Titel der Dissertation:

Knowledge Graphs for Analyzing and Searching Legal Data

Dissertation zur Erlangung des akademischen Grades

eines Doktors

der Sozial- und Wirtschaftswissenschaften an der Wirtschaftsuniversität Wien

eingereicht bei

1. Beurteilerin/1. Beurteiler: **Univ.-Prof. Dr. Axel Polleres**

2. Beurteilerin/2. Beurteiler: **Univ.-Prof. Dr. Jan Mendling**

3. Beurteilerin/3. Beurteiler: **Dr. Serena Villata**

von **Mag.iur. Erwin Filtz, MSc (WU)**

Fachgebiet: **Informationswirtschaft**

Wien, im **April 2021**

Vienna University of Economics and Business

Department of Information Systems and Operations Management

Institute for Data, Process and Knowledge Management

Welthandelsplatz 1, 1020 Wien

Knowledge Graphs for Analyzing and Searching Legal Data

Dissertation

by

Mag.iur. Erwin Filtz, MSc (WU)

Advisors

Dr. Sabrina Kirrane

Univ.-Prof. Dr. Axel Polleres

April 2021

The work presented herein has been supported by two funded projects: *Linked Legal Data for Citizens* funded by the Jubiläumsfonds der Stadt Wien für die WU Wien and *Semantische Suche in Rechtsinformationen* (Semantic Search in Legal Information Systems) funded by Austrian Ministry for Digital and Economic Affairs.

Acknowledgments

I would like to thank my supervisor Axel Polleres for giving me the chance to pursue a PhD, his support, motivation and lots of ideas over the past years. A big thank you to Sabrina Kirrane for her continuous support, encouragement and guidance during this PhD journey. I am very thankful to my examiners Jan Mendling and Serena Villata for joining my committee. I also want to thank my co-authors, Axel, Cristiana, Gerhard, María, Sabrina and Víctor for their valuable help and fruitful discussions. Furthermore, I would like to thank my colleagues, in particular Amr, Javier, Jürgen, Sebastian, Svitlana and Vadim for their motivation and advice. Finally, I would like to thank my family for everything.

Kurzfassung

In Zeiten zunehmender Globalisierung und Zusammenarbeit auf verschiedenen Ebenen wird es immer wichtiger Rechtsinformationen im Internet für die Öffentlichkeit zur Verfügung zu stellen, um allen interessierten Personen die Möglichkeit zu geben diese Informationen abzurufen. Die Rechtsinformationssysteme haben sich im Laufe der Zeit unterschiedlich entwickelt, wodurch sich der Suchprozess, speziell über verschiedene solcher Systeme (etwa über Ländergrenzen oder verschiedene Rechtsbereiche) hinweg, schwierig gestalten kann. Dies sehen wir insbesondere, wenn es um die Verknüpfung von nationalen und internationalen Rechtsdaten geht. Die Europäische Union hat diese Probleme erkannt und Vorschläge zu deren Beseitigung unter Einbeziehung semantischen Technologien gemacht, die es ermöglichen sollen, Rechtsinformationen standardisiert und maschinenlesbar mit eindeutigen Kennungen (unique identifiers) zu repräsentieren und zu annotieren. Die Verwendung von semantischen Technologien ermöglicht es uns Rechtsdaten als „Wissensgraph“ und somit auch deren Verknüpfungen darzustellen und abzufragen. In dieser Dissertation werden die Möglichkeiten der Erstellung und Verwendung eines Wissensgraphen für den Rechtsbereich basierend auf dem österreichischen Rechtsinformationssystem unter Bezugnahme auf die Vorschläge der Europäischen Union diskutiert. Des Weiteren werden verschiedene Wege gezeigt um einen Wissensgraphen automationsunterstützt aus bestehenden Rechtsdaten und -dokumenten zu generieren, um schlussendlich die sich daraus ergebenden erweiterten Suchmöglichkeiten, inklusive der Suche von Rechtsdaten aus anderen Ländern, aufzuzeigen, die mit dem bisherigen Rechtsinformationssystem nur umständlich oder gar nicht möglich sind.

Abstract

In times of increased globalization and cooperation, it is more and more important that governments provide access to legal information via the internet such that all interested people have the possibility to access this information. Over time, legal information systems have developed differently in different countries with regards to the available data, formats and accessibility. This leads to a more complicated legal information search process, especially when legal information from different countries and thus also different legal information systems are involved. In particular, the interlinking of legal information from different countries across borders is missing. In order to overcome these problems, the European Union made proposals to foster easier access to and interlinking of legal information. The goal of these proposals is to provide legal information in a standardized and machine-readable way using unique identifiers and annotations. Semantic technologies allow us the represent legal information as a "Knowledge Graph", which links legal data and enables structured querying. In this dissertation, we investigate the possibilities of creating and querying a legal knowledge graph for the Austrian legal system. The proposed legal knowledge graph is created from the data contained in the Austrian legal information system and modeled based on the EU proposals. Furthermore, we also analyze the available linked legal data from other countries and how this data can be integrated. Different approaches to populate the proposed legal knowledge graph in an ideally automated fashion are demonstrated and compared. Finally, we demonstrate how the proposed legal knowledge graph populated with legal data from different countries enables enhanced legal information search possibilities in order to answer search queries, which are not possible at the moment.

Contents

CHAPTER 1

Introduction

Being able to access legal information is a very important aspect in our everyday life as “the law” is everywhere, for example when we buy something at the supermarket or participating in traffic. A study [World Justice Project, 2019] conducted in 101 countries with 1,000 participants per country shows that around half of the participants across the globe faced legal challenges between the year 2015 and 2017. More detailed numbers for Austria show that only two-thirds of the participants knew where to find legal information. These numbers show that access to legal information needs to be improved and made easier for the remaining one-third of people.

What do we actually mean by “legal information”? Legal information can appear in different aspects, for instance as *law* imposing obligations or prohibitions. More generally, we can define the *law* as a framework of rules that govern our everyday life. Legal information can also be contained in court decisions, which are also used to interpret and refine the law. Typically, legal information is contained in documents, which is why we also call them *legal documents*. Such documents can be, for instance, laws and court decisions but also contracts between individuals, which include specific information about the affected authorities or references to other legal documents. We call these particular sequences of words *legal entities*. Furthermore, legal documents can also include *temporal expressions*, which can be combined with legal entities to indicate *legal events* and describe what happened when.

In contrast to former times, when changes to laws have been published only in printed on official bulletin boards in order to adhere to the legal publication requirements, we can now use *legal information systems*. Legal information systems are used to support the search for and finding of the required information for solving a legal problem [van Opijnen and Santos, 2017]. Such a legal information system is, for instance, the Austrian *Rechtsinformationssystem des Bundes*[1] *(RIS)* provided by the Federal Ministry of Digital and Economic Affairs (BMDW)[2], which is available on the web and can be accessed free of charge. The RIS provides a keyword-based search interface, which allows users to search in different kinds of documents, for instance laws or court decisions. Additional filters can be used to restrict the search, for example to a particular publication date of documents. The search results are then presented as long result lists requiring users to go through all the individual documents and check them whether they contain the required information. Furthermore, the documents are only partly interlinked, for

[1] `https://www.ris.bka.gv.at/`, last accessed 2020-12-28
[2] `https://www.bmdw.gv.at/`, last accessed 2020-12-28

instance law references in a court decision are not linked to the actual law document. This requires users to start an additional search for the law in the RIS law section for each law reference. Hence, missing links in the documents reduce navigability and complicate the search process by making it a unnecessarily tedious and time-consuming. Additionally, the search possibilities are often limited by the available metadata, which means that information contained in the actual documents, for instance legal entities, is not available for the search process. This gets even worse when legal sources from the European Union or foreign countries are required to solve a legal problem. In such cases, foreign legal information systems need to be consulted, which might be organized in a completely different way.

Consequently, the problem with the missing links can be solved by adding the links between the documents. Furthermore, information extraction approaches can be used to extract additional information contained in the legal documents, for instance legal entities, to supplement the existing metadata and make it available for the search process. For this purpose, it is possible to link legal data using the *Resource Description Framework (RDF)* [W3C Working Group, 2014], a machine-readable data format, to enable structured queries and easier navigation through interlinked legal documents. In 2011, the EU started to put effort into initiatives towards solving these problems by proposing standards, which should help to interlink legal information across the EU member states based on RDF. The *European Law Identifier (ELI)*[3] for legislative documents and the *European Case Law Identifier (ECLI)*[4] for judiciary documents have been proposed by the Council of the European Union. Both ELI and ECLI assign unique identifiers to and describe a minimum set of metadata for legal documents. The implementation of the proposed standards is not mandatory for the EU member states, which might be the reason for the slow uptake. In the past years since ELI and ECLI have been proposed, some EU member states have at least assigned identifiers to their legal documents, while other member states have not shown any interest in participating in these initiatives.

Austria is one of the EU member states where the ELI and ECLI identifiers (and only the identifiers) have been assigned to RIS documents in a first step. This means, we can take the current state as a starting point to overcome the disadvantages of the search process outlined above. Additionally, we can also build on the already taken efforts and participate in ELI and ECLI. Furthermore, ELI and ECLI also provide the necessary flexibility to accommodate specific national requirements by extending the ELI and ECLI ontologies with classes and properties specific to the Austrian legal system. Hence, a legal knowledge graph being capable of representing related information, for instance links to other legal documents or documents classified into the same class based on a classification schema, enables enhanced search capabilities. Moreover, information extracted from the legal documents can be used to link entities to external knowledge bases such as Geonames or DBpedia, which also enhances legal information search. Additionally, it supports cross-jurisdictional search requests by integrating legal data from other countries and the European Union. We contribute towards the goals of ELI and ECLI aiming to provide easier access to and interlinking of legal information across Europe, which can only be successful if the various member states participate and use the same system. From a practical point of view that would allow us to enable more

[3] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012XG1026(01)`, last accessed 2020-12-28

[4] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52011XG0429(01)`, last accessed 2020-12-28

complex search queries, which either require a complicated search process or cannot be answered at all with the current system, like the following example questions (Q), which will be explained in more detail in Chapter 3:

Q 1 *Which documents are referenced in a specific court decision?*

Q 2 *Over which districts does a court have competent jurisdiction?*

Q 3 *What are the national transpositions of a specific EU directive?*

Q 4 *Which legal documents regulate a specific legal area, searched with keywords in a foreign language?*

Q 5 *Which events are mentioned in a court decision and could be used for a quick overview of the case?*

Indeed, legal search processes conducted by legal experts involve answering such questions and combinations thereof. Any support for answering and processing them partially automatically would promise to make such search tasks much more effective for legal professionals.

Previous research regarding the processing of legal information to support various tasks has been carried in different scientific domains. The legal informatics field of *Computational Law* looks at the "*mechanization of legal analysis*" [Genesereth, 2018] by combining the formalization of rules and facts in terms of logical expressions and reasoning over them to derive consequences. In the 1980s, *Artificial Intelligence (AI)* started to be applied to the legal domain to support solving legal problems, for instance in legal reasoning [v. d. L. Gardner, 1983]. Later on, another area of work in the legal domain focused on data formats to represent legal information such as Metalex [Boer et al., 2002] and Akoma-Ntoso [Palmirani and Vitali, 2011], both being XML (eXtensible Markup Language[5]) standards used to describe the structure and content of legal documents. In parallel, work on legal ontologies started with the goal to enable the interchange of legal information, for instance with the Legal Knowledge Interchange Format (LKIF) [Hoekstra et al., 2007] and legal domain specific ontologies, for example ontologies for privacy policies [Oltramari et al., 2018, Palmirani et al., 2018], to describe a subset of legal domains or problems. The emerging area of Natural Language Processing in the legal domain started with a template-based extraction of persons from legal documents [Dozier and Haschart, 2000]. Over time, this work has been expanded in terms of extracting different kinds of entities and the classification of legal documents ranging from using rule-based approaches over machine learning to finally deep learning approaches [Dozier et al., 2010, Cardellino et al., 2017a, Chalkidis et al., 2019, Leitner et al., 2019, Tuggener et al., 2020]. However, the focus of these previous efforts has been on the content of legal documents rather than on the connections between them. Only in the last years we can see small signs of a shift towards linking national legal data. The Greek Diavgeia project[6] aims at increasing accessibility to legal information by forcing the authorities to provide their documents via the web, from which linked legal data can be created [Chalkidis et al., 2017]. Similar work using ELI and ECLI for Finnish legislation and case law published as RDF is the Finlex Data Bank [Oksanen et al., 2019].

[5]`https://www.w3.org/TR/xml/`, last accessed 2021-03-12
[6]`https://diavgeia.gov.gr/en`, last accessed 2020-12-28

Thus there is a need for linked legal information that allows professional as well as non-professional users to search and navigate through legal information by interlinking national and foreign legal documents in a *Legal Knowledge Graph (LKG)*. Representing legal information in a graph structure, based on a common ontology used by all EU member states, helps us to ease the access to legal information and support cross-boarder search.

1.1 Hypothesis & Research Questions

With a clear motivation for the creation of an Austrian legal knowledge graph and building on the efforts towards linked legal data, the work presented herein is guided by the following overall hypothesis:

> *A Legal Knowledge Graph can be used to interlink legal documents from national and international sources, leading to an enhanced legal information search process with extended search possibilities that are not possible at the moment using traditional legal information systems.*

From this hypothesis we can derive the following concrete research questions (RQ):

RQ 1 *What is required in order to construct a legal knowledge graph from an existing legal information system?*

To answer this research question, we want to know the requirements and pre-existing building blocks we can use in order to transform the data contained in a traditional legal information system into a knowledge graph. Furthermore, we need to combine existing data with existing ontologies, which need to be extended to support national requirements. The Austrian legal system is embedded into the European system and interplays with the legal systems of other countries, which is why ELI and ECLI serve as a basis for our legal knowledge graph.

RQ 2 *Which approaches can be followed in order to populate the legal knowledge graph from different data sources in an automated fashion?*

Data contained in the Austrian legal information system is available to be transferred into a legal knowledge graph. Hence, we need to find ways to enable the population from different data sources. It is necessary to analyze the available data from the RIS (metadata and documents) and compare it to the properties of the ontologies we need to populate. We can derive three sub-research questions:

RQ 2.1 *Which approaches are available for the population of the legal knowledge graph from structured data and how effective are they?*

In order to answer this research question, we need to analyze which information is available in a structured format provided by RIS as metadata and investigate approaches to use this information for the population of the legal knowledge graph.

RQ 2.2 *Which approaches are available for the population of the legal knowledge graph from text sources (i.e. legal documents) and how effective are they?*

In order to answer this research question, we need to investigate which ELI and ECLI properties cannot be populated from the RIS metadata, but from information contained in the legal documents. We will analyze and compare different approaches to extract legal entities from the documents. Furthermore, we will also investigate approaches allowing us to categorize legal documents into a given set of classes.

RQ 2.3 *Which approaches are available for the extraction of events from legal documents and how effective are they?*

In order to answer this research question, we need to investigate events contained in legal documents. Furthermore, it is necessary to analyze the individual event components and compare the performance of different extraction approaches for these components.

RQ 3 *In how far is it possible to enhance the legal inquiry and search process by linking legal data?*

In order to find an answer to this research question, we need to analyze the current legal information search process for which we use the sample questions outlined above. We investigate whether we can leverage the added links and enhanced metadata for enhanced search queries in order to answer the sample questions.

1.2 Contributions

The contributions of this thesis can be summarized as follows:

- **Contribution to RQ 1:** We analyze the proposed ELI and ECLI ontologies and their suitability when it comes to the Austrian legal data and extend the ontologies where needed. In particular, we describe the legal knowledge graph creation methodology and extend the ELI and ECLI ontologies with classes and properties in order to represent the data contained in the Austrian legal information system. Furthermore, we introduce a new thesaurus containing specific terms used in the Austrian legal language and information where the ELI and ECLI ontologies stipulate national extensions, for example a document classification scheme or country specific document types.
- **Contribution to RQ 2.1:** For the population of an Austrian legal knowledge graph we propose three different population methods based on the available metadata provided by RIS. In particular, we propose three methods for the population of the legal knowledge graph: (i) Methods that allow a direct transfer of the data requiring only a minimum of preprocessing effort; (ii) Methods based on additional conditions and lookups; and (iii) Methods to interlink RIS data with external knowledge bases.
- **Contribution to RQ 2.2:** We propose population approaches based on NLP tools and techniques by (i) extracting information from the documents; and (ii) using the document content to classify these documents into the classes of a given thesaurus. For both tasks, we use state-of-the-art approaches already successfully applied to documents from other domains, whose performance we compare and evaluate based on datasets containing legal documents. In more detail, we provide a new

corpus with 50 manually annotated Austrian Supreme Court decisions, which is used for the legal entity extraction experiments. The performance of classification approaches is evaluated on gold-standard datasets containing legal documents from the European Union.

- **Contribution to RQ 2.3:** We identify the problems of extracting temporal expressions in court decisions. Furthermore, we propose three temporal dimensions along which temporal expressions contained in court decisions can be classified. We provide a new gold standard corpus with temporal annotations of thirty manually annotated court decisions, ten documents from the European Court of Justice, the European Court of Human Rights and the United States Supreme Court each. We use this corpus to compare and discuss the features and performance of ten state-of-the-art, but not legal domain specific, temporal taggers. We provide an overview of the most common errors and problems of these generic temporal taggers. The extraction of legal events from court decisions helps to get quick overview over a case. We introduce two different types of events and define event components to further fragment the information contained in an event. We provide another manually annotated gold standard corpus with thirty court decisions from the European Court of Human Rights annotated with legal events. This corpus is used to extract and classify events contained in the court decisions. For both tasks we analyze the performance of state-of-the-art event extraction approaches.
- **Contribution to RQ 3:** We provide a comparison of the current situation regarding legal information systems and search possibilities for all EU member states. We analyze the availability of legal data and implementation status of ELI and ECLI, the used data formats and additional information. We also describe the access to and features of legal databases of all EU member states from a more general point of view, the document formats used for the dissemination of legal documents and in which languages legal information is available. We describe non-governmental efforts based on ELI and ECLI towards the provision of linked legal data classifying them based on their features. We demonstrate the benefits of linked legal data by showing examples of queries driven by practical legal search use cases, which are possible with a legal knowledge graph, but have not been possible before including integrated legal data from other countries.

1.3 Thesis Structure

The remainder of the thesis is structured as follows:

Chapter 2 presents background information in relation to Knowledge Graphs, Semantic Web, Linked Data and introduces the legal ontologies and thesauri used throughout the thesis. Furthermore, it contains an introduction into Natural Language Processing (NLP) and language models as well as commonly used NLP tasks, approaches and tools.

Chapter 3 describes the challenges with traditional legal information systems exemplified with the Austrian RIS and the derived requirements for the creation of a legal knowledge graph. This chapter also presents the creation methodology and finally introduces the Legal Knowledge Graph Ontology (LKG) containing the new classes and properties to properly represent the Austrian legal system.

Chapter 4 introduces different knowledge graph population approaches from various data sources using Natural Language Processing tools and techniques. In particular, we describe the extraction of entities from legal documents and the classification of documents into a large number of disjoint classes. Experiments are performed and the results compared and discussed for both tasks.

Chapter 5 focuses on temporal information contained legal documents, in particular court decisions. We describe the challenges of extracting temporal information from court decisions and introduce different temporal dimensions. Furthermore, we compare the performance of ten non-domain specific temporal taggers on detecting temporal information. Moreover, temporal information is also part of events that can be extracted from court decisions and presented in a timeline. We compare different approaches to extract legal events from court decisions and discuss their performance.

Chapter 6 presents and compares the initiatives carried out in other European countries with regards to (linked) legal data. An overview shows which countries participate in the EU driven initiatives or decide to go another way. Furthermore, this chapter also introduces non-governmental initiatives in the area of linked legal data. Finally, we present the benefits of linked legal data and present a roadmap towards a linked legal knowledge graph for stakeholders thinking about providing linked legal data or creating a legal knowledge graph.

Chapter 7 summarizes the findings of this thesis, answers the research questions and discusses future research directions.

1.4 Publications and Impact

The content presented in this thesis has been presented and published in different peer-reviewed international conferences and journals and contains material from (in chronological order):

- Erwin Filtz, Sabrina Kirrane, Axel Polleres, and Gerhard Wohlgenannt. Exploiting Eurovoc's Hierarchical Structure for Classifying Legal Documents. On the Move to Meaningful Internet Systems: OTM 2019 Conferences - Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21-25, 2019, Proceedings, volume 11877 of Lecture Notes in Computer Science, pages 164–181. Springer, 2019. [Filtz et al., 2019]

  In this paper, we compare various approaches that can be used to classify legal documents in a multi-label classification setting using corpora with legal documents published by the European Union. We contrast the results with a well-known dataset from the news domain used in classification tasks. Within the thesis, this work is presented in Section 4.3. An extension of this work shows that the results can be boosted by using transformer models [Shaheen et al., 2020].

  This publication contributes to RQ 2.2.

- María Navas-Loro, Erwin Filtz, Víctor Rodríguez-Doncel, Axel Polleres and Sabrina Kirrane. TempCourt: Evaluation of Temporal Taggers on a new Corpus of Court Decisions. The Knowledge Engineering Review, 34:e24, 2019. doi:10.1017/S0269888919000195. [Navas-Loro et al., 2019]

This work focuses on temporal information contained in court decisions and compares the performance of ten non-domain specific temporal taggers. In order to evaluate the performance of these taggers, we created a manually annotated gold standard corpus of court decisions from three different courts. This work is presented in Section 5.1.

This publication contributes to RQ 2.3.

- Erwin Filtz, María Navas-Loro, Cristiana Santos, Axel Polleres, and Sabrina Kirrane. Events matter: Extraction of events from court decisions. Legal Knowledge and Information Systems - JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020, volume 334 of Frontiers in Artificial Intelligence and Applications, pages 33–42. IOS Press, 2020. [Filtz et al., 2020]

  In this article, we introduce two different types of events commonly found in court decisions and compare different state-of-the-art event extraction approaches. In addition, we also extract three event components to describe an event, which enables us to create a timeline to provide a quick overview of a court decision. The content of this work is presented in Section 5.2.

  This publication contributes to RQ 2.3.

- Erwin Filtz, Sabrina Kirrane, and Axel Polleres. The linked legal data landscape: linking legal data across different countries. Artificial Intelligence and Law, pages 1–55. [Filtz et al., 2021]

  In this paper, we describe the Austrian use case for a legal knowledge graph based on the Austrian legal information system and cover all topics from the modeling to the integration of legal data from other countries. The background information from this paper is covered in Chapter 2. The challenges and requirements as well as description of the modeling section are presented in Chapter 3. The description of population approaches is covered in Chapter 4. Finally, the integration of legal data is discussed in Chapter 6.

  This publication contributes to RQ 1, 2.1, 2.2 and 3.

The following additional works have been published by the author and are partially related to the work presented herein, while not having contributed directly to the content presented in the present thesis:

- Erwin Filtz. Building and processing a knowledge-graph for legal data. The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part II, volume 10250 of Lecture Notes in Computer Science, pages 184–194, 2017. [Filtz, 2017]

- Erwin Filtz, Sabrina Kirrane, and Axel Polleres. Interlinking legal data. Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems co-located with the 14th International Conference on Semantic Systems (SEMANTiCS 2018), Vienna, Austria, September 10-13, 2018., volume 2198 of CEUR Workshop Proceedings. CEUR-WS.org, 2018. [Filtz et al., 2018]

- Martin Beno, Erwin Filtz, Sabrina Kirrane, and Axel Polleres. Doc2rdfa: Semantic annotation for web documents. Proceedings of the Posters and Demo Track of the 15th International Conference on Semantic Systems co-located with 15th

International Conference on Semantic Systems (SEMANTiCS 2019), Karlsruhe, Germany, September 9th - to - 12th, 2019., volume 2451 of CEUR Workshop Proceedings. CEUR-WS.org, 2019. [Filtz et al., 2019]

- Zein Shaheen, Gerhard Wohlgenannt and Erwin Filtz. Large Scale Legal Text Classification Using Transformer Models. SEMAPRO 2020 The Fourteenth International Conference on Advances in Semantic Processing, Nice, France, October 25-29, 2020, pages 7-17, IARIA 2020. [Shaheen et al., 2020]

CHAPTER 2

Background

This chapter provides the necessary background information about the technologies and standards used in this thesis. Section 2.1 introduces the term *Knowledge Graph*, which has gained popularity in the last decade. We compare different attempts to define the term and put the definitions into context with respect to the legal domain. Section 2.2 introduces basics of *Semantic Web* including the commonly used data models and serializations as well as the query language used to retrieve the data. Furthermore, we provide a summary of *Linked Data* and its principles. Section 2.3 gives an overview of *ontology engineering* and introduces the *legal ontologies* on which our work is based. In particular, we introduce the ontologies proposed by the European Union for legislative and judiciary documents. Moreover, we include an introduction of the EuroVoc thesaurus maintained by the Publications Office of the European Union as well as the Common Data Model ontology used for the dissemination of documents by the EU. Finally, Section 2.4 provides an overview of *Natural Language Processing (NLP)* tasks and techniques used in this thesis and introduces the concept of language models.

2.1 Knowledge Graphs

Knowledge Graphs [Hogan et al., 2020] are a trending topic, which is attracting increased interest in various domains. In order to organize and link information in an easy manner, such knowledge graphs typically contain both factual and schematic (or, resp., ontological) information, in a flexible and extensible graph structure.

2.1.1 What is a Knowledge Graph?

In 2012, the term *Knowledge Graph* gained popularity when Google used it to denote searching for things instead of strings [Amit Singhal, 2012]. While there are many different proposals for a definition of what a knowledge graph is, there is no formal definition available [Ehrlinger and Wöß, 2016]. In the following, we list different attempts to define the term:

- "*[...] a 'graph'—that understands real-world entities and their relationships to one another: things, not strings.*" [Amit Singhal, 2012]

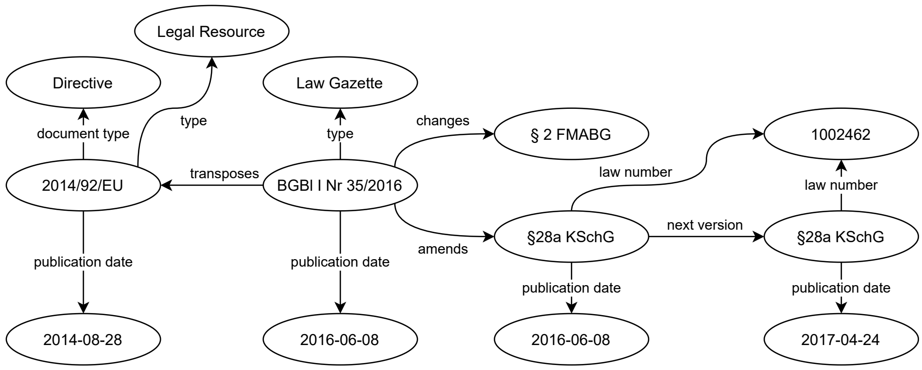

Figure 2.1: Knowledge Graph Example

- "*A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.*" [Ehrlinger and Wöß, 2016]
- "*From a broader perspective, any graph-based representation of some knowledge could be considered a knowledge graph (this would include any kind of RDF dataset, as well as description logic ontologies.)*" [Paulheim, 2017]
- "*[...] we use the term knowledge graph for any RDF graph.*" [Färber et al., 2018]
- "*A graph of data with the intent to compose knowledge.*" [Bonatti et al., 2018]
- "*A graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities.*" [Hogan et al., 2020]

Summarizing the aforementioned definitions of the term "knowledge graph", we can see that they are all centered around a graph and representing data or knowledge as a graph-structure, interlinking entities and describing their relations. The focus is mostly on the graph-structure and the contained knowledge in a graph, only [Färber et al., 2018] explicitly state the RDF data format, which is a standard format to represent knowledge graphs that evolved from W3C's Semantic Web activity. RDF turned out to be very useful for publishing and consuming knowledge graphs in an interoperable and transparent manner, which is why we will also rely on this format in the present thesis.

An example of a knowledge graph is shown in Figure 2.1. The knowledge graph consists of a set of *nodes* and a set of *edges* connecting the nodes. Nodes in a knowledge graph are used to represent *entities* about which information should be stored. Such entities can be tangible objects, for instance a printed document, and intangible objects, for instance the virtual representation of a document shown on a computer screen. In our knowledge graph, the edges connecting the nodes are directed, which is why this graph is called a *directed edge-labelled graph* ([Hogan et al., 2020]). For instance, an edge "publication date" has a start node "2014/92/EU" and an end node "2014-08-28". This tells us that the entity "2014/92/EU" has been published on "2014-08-28". Knowledge represented in a graph also allows for an easy extension of the knowledge simply by adding additional nodes to the graph. More general, knowledge graphs can be easily created by starting with a single node and extending it when required. The knowledge can be acquired from various sources, such as from humans and user communities, text documents or other sources, which already provide information in a structured format [Hogan et al., 2020].

The data source used for populating the knowledge graph also determines the exact population process in terms of preprocessing and data curation, extraction of information from unstructured sources or updating mechanisms. For example, it makes a difference whether the information to populate a knowledge graph is complete and available in a structured format or needs to be extracted from raw text documents.

2.1.2 Knowledge Graph Examples

We distinguish between open and enterprise knowledge graphs as well as domain-independent and domain-specific knowledge graphs [Hogan et al., 2020]. Open knowledge graphs are available on the Web and publicly accessible, on the contrary, enterprise knowledge graphs typically contain private enterprise data and are therefore not publicly available. Domain-independent knowledge graphs may contain data about many different topics and domains, whereas domain-specific knowledge graphs focus on collecting knowledge on a particular domain. Well-known examples for open knowledge graphs are DBpedia [Lehmann et al., 2015], which is created from Wikipedia[1] and Wikidata [Vrandecic and Krötzsch, 2014], which is populated by its users. Domain-specific knowledge graphs can be found for various domains, for instance geography [Stadler et al., 2012] and medicine [Rotmensch et al., 2017]. More examples of open and enterprise knowledge graphs can be found in works such as [Hogan et al., 2020, Heist et al., 2020, Guo et al., 2020].

2.1.3 Knowledge Graphs in the Legal Domain

In recent years knowledge graphs have also been adopted in the legal domain. In the following, we present some examples:

- **Lynx** The Lynx knowledge graph has been developed as part of the Lynx project[2] and focuses on smart compliance services [Montiel-Ponsoda et al., 2017]. The knowledge graph targets selected legal areas, for instance data protection and compliance for small and medium sized enterprises. This knowledge graph includes legal documents from the European Union and Spanish courts.
- **GDPR & PCI DSS** This knowledge graphs aims at integrating the different data protection rules set by the European General Data Protection Regulation (GDPR) and the Payment Card Industry Data Security Standard (PCI DSS). This knowledge graph should enable users to find contradicting regulations and to help companies to automate the checking of data protection breaches [Elluri et al., 2018].
- **Wolters Kluwer** provides knowledge for multiple domains including the legal domain[3]. Their knowledge graph covers a large dataset of German court cases, which are transformed from XML format into a knowledge graph using a custom data structure with the goal to enhance the search process for users [Junior et al., 2020].

[1] `https://www.wikipedia.org/`, last accessed 2021-02-21
[2] `https://lynx-project.eu/`, last accessed 2020-12-28
[3] `https://www.wolterskluwer.com/`, last accessed 2021-02-21

Listing 2.1: Namespaces used in examples throughout the paper

```
PREFIX lkg: <http://data.wu.ac.at/legal/lkg#>
PREFIX av: <http://data.wu.ac.at/legal/austrovoc#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http//www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#date>
PREFIX cdm: <http://publications.europa.eu/ontology/cdm#>
PREFIX frbroo: <http://iflastandards.info/ns/fr/brbr/frbroo/>
PREFIX eli: <http://data.europa.eu/eli/ontology#>
PREFIX ev: <http://eurovoc.europa.eu/>
PREFIX gn: <http://sws.geonames.org/>
PREFIX ris: <https://www.ris.bka.gv.at/>
```

Many open standards and technologies to create, represent, interchange and process Knowledge Graphs originated from the Semantic Web and Data activities within the World Wide Web Consortium (W3C)[4].

2.2 Semantic Web and Linked Data

The Semantic Web has been introduced by Tim Berners-Lee et al. with the goal to bring meaning and structure into existing websites and make information machine-readable [Berners-Lee et al., 2001]. With the Semantic Web it is possible not just to link websites but also to give the links a meaning, which can be processed by a machine.

2.2.1 Resource Description Framework

The *Resource Description Framework (RDF)*[5] is the data model underpinning the Semantic Web and enables machine-readability of the data. The RDF data model follows a graph-structure with nodes describing the subjects and objects of a statement (i.e. resources) and edges describing how they are related. A *resource* can be any tangible or intangible thing, for instance a person, a car or a legal document that exists electronically only, about which information should be stored and is identified by an *Unique Resource Identifier (URI)*[6]. As URIs can be very long with similar prefixes, *namespaces* are used to abbreviate long URIs and help to make RDF more readable by shortening the URIs, for instance `http//www.w3.org/2000/01/rdf-schema#label` can be replaced with `rdfs:label`. Listing 2.1 shows the prefixes and namespaces used throughout this thesis. Besides resources, a *literal* is used to describe properties of a resource. Untyped literals are interpreted as a plain string, for instance "Konsumentenschutz" and can have an additional language tag such as "Konsumentenschutz"@de where "de" is used to indicate that German is the language used in this literal. Typed literals also indicate the datatype that provides information on how this literal is interpreted, for instance as a date (`xsd:date`) or as an integer (`xsd:integer`). An RDF statement, commonly referred to as a *triple*, consists of a *subject*, *predicate* and *object*. The subject of a

[4]`https://www.w3.org/2001/sw/`, last accessed 2020-12-28
[5]`https://www.w3.org/TR/rdf11-concepts/`, last accessed 2020-12-28
[6]`https://tools.ietf.org/html/rfc1630`, last accessed 2020-12-28

Listing 2.2: RDF snippet for EU Directive 2014/92/EU

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX rdfs: <http//www.w3.org/2000/01/rdf-schema#>
3  PREFIX eli: <http://data.europa.eu/eli/ontology#>
4  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#date>
5  PREFIX frbroo: <http://iflastandards.info/ns/fr/frbr/frbroo/>
6  PREFIX eures: <http://publications.europa.eu/resource/authority/resource-type/>
7
8  <http://data.europa.eu/eli/dir/2014/92/oj>
9      rdf:type
10         eli:LegalResource ;
11     eli:type_document
12         eures:DIR ;
13     eli:date_publication
14         "2014-08-28"^^xsd:date .
15 <http://data.europa.eu/eli/ontology#LegalResource>
16     rdfs:subclassOf
17         frbroo:F1_Work .
```

triple must be a URI, the object can also be a literal. A collection of triples describing both the schema and instance data is called *ontology*. There are various formats available for the serialization of RDF data, for instance RDF/XML[7], N-Triples[8], RDF in Attributes (RDFa)[9] for embedding RDF in HTML (Hypertext Markup Language)[10] and the JavaScript Object Notation[11] (JSON) for Linked Data (JSON-LD[12]). Another format is the Terse RDF Triple Language (Turtle)[13], which we use in the examples presented throughout this theses, as it is easy to read. Turtle also supports the usage of shortcuts. A semicolon can be used when all predicate and object pairs belong to the same subject, a comma when all objects belong to the same subject and predicate pair.

Listing 2.2 shows an RDF snippet about the EU Directive 2014/92/EU and represents the information about the directive presented in Figure 2.1. The first six lines contain the namespaces used in this example in order to shorten the URIs and increase readability. This allows us to write `eli:type_document` instead of `http://data.europa.eu/eli/ontology#type_document` by defining `PREFIX eli: <http://data.europa.eu/eli/ontology#>` once, which allows us to use the prefix `eli:` in the whole RDF document without restating the full URI. The lines 19 (`http://data.europa.eu/eli/dir/2014/92/oj`) and 26 (`http://data.europa.eu/eli/ontology#LegalResource`) show two subjects within this RDF snippet, the lines 20 (`rdf:type`[14]), 22 (`eli:type_document`), 24 (`eli:date_publication`) and 27 (`rdfs:subclassOf`) the predicates and lines 21 (`eli:LegalResource`), 23 (`<http://publications.europa.eu/resource/authority/resource-type/DIR>`), 25 (`"2014-08-28"^^xsd:date`) and 28 (`frbroo:F1_Work`) the objects. All objects are resources with the exception of the object in line 13, which is a typed literal of datatype `xsd:date`. Furthermore, this example illustrates the class hierarchy as the EU directive (line 19) is of type (`rdf:type`) `eli:LegalResource`, which itself is a `rdfs:subclass` of `frbroo:F1_Work` (lines 14 to 17).

[7] `https://www.w3.org/TR/rdf-syntax-grammar/`, last accessed 2020-12-28
[8] `https://www.w3.org/TR/n-triples/`, last accessed 2020-12-28
[9] `https://www.w3.org/TR/rdfa-primer/`, last accessed 2020-12-28
[10] `https://html.spec.whatwg.org/multipage/`, last accessed 2020-12-28
[11] `https://tools.ietf.org/html/rfc8259`, last accessed 2020-12-28
[12] `https://www.w3.org/TR/json-ld11/`, last accessed 2020-12-28
[13] `https://www.w3.org/TR/turtle/`, last accessed 2021-03-21
[14] The predicate `rdf:type` is commonly abbreviated with `a`.

Listing 2.3: Example SPARQL Query

```
# Which directives have been published in 2014?
PREFIX eli: <http://data.europa.eu/eli/ontology#>
PREFIX eu: <http://publications.europa.eu/resource/authority/resource-type/>
SELECT (?s as ?Directive)
WHERE {
  ?s eli:type_document eu:DIR .
  ?s eli:date_publication ?d .
  FILTER (year(?d) = 2014)
}
```

RDF Schema (RDFS)[15] and the *Web Ontology Language (OWL)*[16] are used to describe classes of and properties (relations) between resources. The core features of RDFS are summarized in the ρdf subset [Muñoz et al., 2009], which contains properties to define simple taxonomies in terms of class (`rdfs:subClassOf`) and property (`rdfs:subPropertyOf`) hierarchies. Likewise, domain (`rdfs:domain`) and range (`rdfs:range`) restrictions can be used to infer the class membership of subjects (domain) or objects (range) of particular properties as shown in Listing 2.2 line numbers 13 and 14 that the ELI class `eli:LegalResource` is a subclass of `frbroo:F1_Work`. OWL caters for the definition of more complex ontological axioms on classes and properties, which can be used for more complex reasoning.

2.2.2 SPARQL Protocol and RDF Query Language

The *SPARQL Protocol and RDF Query Language (SPARQL)*[17] is used to retrieve RDF data. SPARQL queries search for matches of user defined triples. The set of triples is called *graph pattern*. A `SELECT` query allows users to define a graph pattern, which must match the data and the variables to be returned. Basic graph patterns must match all results in order to be returned, whereas in an `OPTIONAL` query we can also define *optional patterns* that need not occur in all results and return an empty binding (result) if not matched. With *alternative patterns* using `UNION` it is possible to define multiple graph patterns of which at least one must be fulfilled. The number of results can be reduced using a `FILTER` clause, which allows users to restrict results to literals that contain a particular string, or to apply comparison operators such as equals, greater than and so on. Listing 2.3 shows an example SPARQL query that could be used to retrieve all EU directives that have been published in 2014. The `WHERE` clause selects all documents of type `eu:DIR` and the publication date. The result would contain all directives but a `FILTER` reduces the result list to results where the publication year is 2014 only. *Solution modifiers* are used to manipulate query results such as sorting the results in a particular way. With the keyword `ORDER BY` it is possible to sort the results on a particular variable. When no additional information is given the results are sorted in an ascending way by default. The sorting direction can also be set explicitly by adding `ASC(?variable)` or `DESC(?variable)` to set the order of sorting to ascending or descending.

[15]`https://www.w3.org/TR/rdf-schema/`, last accessed 2020-12-28
[16]`https://www.w3.org/TR/owl2-overview/`, last accessed 2020-12-28
[17]`https://www.w3.org/TR/sparql11-overview/`, last accessed 2020-12-28

2.2.3 Linked Data

In order to make machine-readable data more accessible on the Web, Tim Berners-Lee [Berners-Lee, 2006] proposed a set of *Linked Data Principles* for publishing data on the Web, which fundamentally rely on RDF:

"1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI provide useful information using the standards RDF and SPARQL.

4. Include links to other URIs, so that they can discover more things." [Berners-Lee, 2006]

The *things* mentioned in the first principle refer to resources. Identifying resources with HTTP URIs allows consumers to retrieve additional information about these resources on the Web. Information about the resources stored in RDF allows them to be retrieved using SPARQL. The fourth rule stipulates that resources should be linked with other resources and shall allow users or agents to browse through different resources by following links.

The advantages of following the linked data principles are obvious. Using unique identifiers to identify, the standard HTTP protocol to retrieve and RDF to describe the data enables and turns linked data into a "Web of Data" [Bizer et al., 2009]. Data is not only accessible via standard browsers but also interlinked such that the links can easily be followed and data from even different sources can be easily explored [Bizer, 2009]. Furthermore, the data is also self-describing as all the information is included and so far unknown information can be looked up by dereferencing the URIs, which reveals new information, for instance about an entity or the relation between entities. In this thesis, we will follow these principles as closely as possible and in Chapter 6 we investigate the current state of linked data in other EU member states.

2.3 Ontologies

In this section, we provide an overview about ontologies. In the context of this theses, ontologies may be viewed as RDF vocabularies to denote the schemata used to define allowed edge labels and node types in our knowledge graph. From a more general point of view, we can also define an ontology as "*a description of knowledge about a domain of interest*" [Hitzler et al., 2010]. In addition, several more formal definitions of the term "ontology" have been proposed (cf. [Gruber, 1993, Borst, 1997, Studer et al., 1998]) centered around the "conceptualization" (a "model") describing objects and their relations of the real-world in an abstract way [Genesereth and Nilsson, 1988]. An ontology is therefore helpful to represent the knowledge of a specific domain – in our case the legal domain – and to serve as a means for the information exchange between humans and machines [Guarino et al., 2009]. In this section, we provide an overview of ontology engineering approaches to create and extend ontologies. Furthermore, we provide a description of the legal ontologies used throughout this thesis as well as the main ontology used by the EU for the dissemination of documents, the *Common Data Model (CDM)*.

2.3.1 Ontology Engineering

There are three commonly used approaches to create an ontology. In a *bottom-up* approach, the ontology creation process starts with the description of objects in terms of their individual units that cannot be split anymore and their relations to each other from which more abstract concepts ("abstraction") are derived [van der Vet and Mars, 1998]. Ontologies can also be created the other way around in a *top-down* fashion starting with very abstract (general) concepts from which more fine-grained concepts are derived ("specialization"). The third approach is called *middle-out*. This approach starts "in the middle" to create an ontology with generic concepts on the one hand and detailed concepts on the other hand. It aims at finding the missing concepts in the middle serving as a specialization for the more generic concepts and as an abstraction for the detailed concepts at the same time. Each ontology creation approach has its own advantages and disadvantages [Uschold and Gruninger, 1996]. A bottom-up approach might result in a very high level of detail with a huge number of different concepts making it hard to find similarities between them. A top-down approach is at risk of starting with too generic concepts that might properly reflect the required level of detail from the more fine-grained concepts through the specialization process. The middle-out approach is a combination of top-down and bottom-up approaches that should reduce the required effort by limiting the level of detail as the concepts need to be aligned with the more abstract concepts [Uschold and Gruninger, 1996, Ghosh et al., 2016].

There is also a body of work dealing with ontology creation methodologies. Such methodologies describe standard procedures and sequences of tasks carried out during the ontology engineering process to create an ontology. Over time, a large number of different ontology creation methodologies have been proposed, all with their different goals, advantages and disadvantages (cf. [Jones et al., 1998, Pinto and Martins, 2004, Cristani and Cuel, 2005, Iqbal et al., 2013, Simperl and Luczak-Rösch, 2014]). Some of these ontology creation methodologies have also been applied to the legal domain, for instance METHONTOLOGY [Fernandez et al., 1997, Corcho et al., 2003], TERMINAE [Biebow et al., 1999] or Ontology Development 101 [Noy et al., 2001].

Ontology design patterns describe common patterns in the ontology creation process [Gangemi, 2005, Gangemi and Presutti, 2009]. Content Ontology Design Patterns provide building blocks to ensure reusability [Presutti and Gangemi, 2008] and can be used to create (legal) domain-specific ontologies [Gangemi, 2007]. The application of content ontology design patterns has already been demonstrated for the legal domain, for instance for the modeling of licensing [Rodríguez-Doncel et al., 2013] and consumer complaints [Santos et al., 2016].

As ontologies are used to describe domain knowledge there might not be the need to create a new ontology for a specific domain, but instead reuse [Noy et al., 2001] or extend [Uschold and Gruninger, 1996] an existing ontology. This practice reduces the effort to create a new ontology but might lead to an increased effort in adapting the existing ontology to the specific requirements [Bontas et al., 2005]. In the case of legal knowledge graphs, we can build on the already existing ELI and ECLI ontologies, which will be introduced next.

2.3.2 Legal Ontologies

Legal ontologies are ontologies specifically designed for the legal domain. As the legal domain is very broad in terms of different legal areas and jurisdictions, a variety of different legal ontologies has been proposed to cover specific domains or problems (cf. overviews presented in [Leone et al., 2019, de Oliveira Rodrigues et al., 2019]). The work presented herein builds on the ELI and ECLI ontologies.

Article 67 of the Treaty of the Functioning of the European Union[18] (TFEU) stipulates that an area of freedom, security and justice should be constituted while respecting the legal systems of the EU member states. The Council of the European Union identified a number of problems related to legislative [Council of the European Union, 2012, Council of the European Union, 2017] as well as judiciary documents [Council of the European Union, 2011], which can be summarized as follows:

1. Legal information cannot be acquired from EU sources only;
2. Legal information search in various databases is a complex and not user friendly task;
3. Legal information exchange between EU member states is hampered by different legal and technical systems; and
4. Legal documents use a variety of national identifiers and are not necessarily compatible.

In order to overcome the aforementioned problems, the Council also proposed actions to fit the requirements of Art 67 TFEU and therefore concluded:

1. The national systems should not be replaced by a centralized EU platform or system and member states should be able to continue their system;
2. Common identifiers for legislative and judiciary documents enabling interlinking of these documents should be introduced;
3. A minimum set of metadata should be published with the legal information; and
4. Ontologies for legislative and judiciary are proposed.

The outcome of these proposals are the ELI and ECLI ontologies, which we will use as a basis for the legal knowledge graph in this thesis. In addition to the ELI and ECLI ontologies, the EU also uses the standardized *EuroVoc* thesaurus, which is not restricted to legislation alone. The EuroVoc thesaurus contains normative terminology and is available as an RDF vocabulary. Furthermore, the thesaurus is used to classify legal documents published by the EU and its authorities.

European Law Identifier

The European Law Identifier (ELI) [Council of the European Union, 2012] serves as a common system to identify legislative documents and its metadata. It was first proposed in 2011 and followed by additional Council conclusions in 2017 [Council of the European Union, 2017] acknowledging the efforts of the participating countries,

[18] `https://eur-lex.europa.eu/eli/treaty/tfeu_2008/art_67/oj`, last accessed 2020-12-28

Table 2.1: Mandatory properties of the ELI ontology

Property	Description
eli:realizes	Describes that a legal expression materializes a legal resource.
eli:embodies	Describes that a format represents a legal expression.
eli:type_document	Indicates the type of a legal resource.
eli:language	The language in which a legal expression is written.
eli:title	The title of a legal expression.
eli:format	Resource format expressed as URI (e.g. HTML).

introducing an ELI task force and clarifying the three pillars of the ELI system. The three pillars [Francart et al., 2018] the ELI is built on are:

1. to foster the assignment of unique identifiers for laws;
2. to use a common ontology that provides a metadata standard; and
3. to provide said metadata in a machine-readable form.

The EU is required to publish legal acts in various languages and therefore needs the ability to represent different language versions of the same legal act. The ELI ontology distinguishes between three classes of resources and has six mandatory properties. As shown in Table 2.1, a `eli:LegalResource` is a distinct intellectual creation such as a legal act, which is realized by a `eli:LegalExpression` and embodied in a specific `eli:Format`. Hence, a `eli:LegalExpression` has a `eli:title` and `eli:realizes` the base version in a particular language (`eli:language`) of a `eli:LegalResource`, which is of a specific `eli:type_document`, for instance a directive. The `eli:LegalExpression` is published in a `eli:Format`, which is the actual physical representation, whereas physical includes paper as well as electronic formats such as HTML or PDF (Portable Document Format). Despite the goals laid out for ELI, it also contains properties (`eli:type_document`, `eli:passed_by`, `eli:is_about` and `eli:version`) for which the EU member states are encouraged to create their own lists or schemes. This makes sense as these properties are used for information that is very likely different for every country. For example, the type of documents used in a legal system are specific for each EU member state such that it does not make sense to provide a predefined list of document types.

The ELI (both in terms of identifier syntax and in terms of the usage of metadata properties) is modeled in different ways from country to country depending on the respective legal system. Notably, the Council conclusions define all of the syntactic components of the ELI being optional, such that national requirements can be fulfilled and not all components need to be implemented in each national legal system. Additional information for the member states as well as reference files for the ELI ontology are provided in HTML[19], XLSX[20] and OWL[21] format. The ELI follows the principles set forth in the Functional Requirements for Bibliographic Records [International Federation

[19] `http://publications.europa.eu/resource/distribution/eli_documentation/html/doc_user_manual/eli_ontology.html`, last accessed 2020-07-15

[20] `http://publications.europa.eu/resource/distribution/eli/xlsx/owl/eli_ontology.xlsx`, last accessed 2020-07-15

[21] `http://publications.europa.eu/resource/distribution/eli/owl/owl/eli.owl`, last accessed 2020-07-15

Table 2.2: Mandatory properties of the ECLI ontology

Property	Description
`dcterms:identifier`	The URL where the resource can be retrieved.
`dcterms:isVersionOf`	Indicates that a resource is a version of another resource.
`dcterms:creator`	Full name of deciding court.
`dcterms:coverage`	Indicates the country in which the court or tribunal has its seat.
`dcterms:date`	The date when a decision has been rendered.
`dcterms:language`	The language in which this particular is written.
`dcterms:publisher`	The organization that is responsible for the publication of the document.
`dcterms:accessRights`	Defines who can access the resource, *public* or *private*.
`dcterms:type`	Defines the type of the rendered decision.

of Library Associations and Institutions, 2009] (FRBR) ontology but uses the object-oriented version of FRBR [International Federation of Library Associations and Institutions, 2016] for the ELI ontology (prefix `frbroo:`), for instance `eli:LegalResource` is a `rdfs:subClassOf frbroo:F1_Work` and `eli:LegalExpression` is a `rdfs:subClassOf frbroo:F22_Self-Contained_Expression` [Publications Office of the European Union, 2020b]. The ELI syntax is very flexible and can be adjusted to national requirements by adding and removing individual components. The syntax of the ELI identifier is defined as the base URI followed by *eli* with the rest of the components being optional and separated by slashes, for instance the ELI for a EU directive such as `http://data.europa.eu/eli/dir/2014/92/oj` looks different from an Austrian legal provision, like `https://www.ris.bka.gv.at/eli/bgbl/1979/140/P28a/NOR40180997`.

Regarding the modeling of ELI as proposed by the EU, [Francesconi et al., 2015] highlight the disadvantages of coupling resources with the corresponding FRBR classes. The authors state that such a coupling leads to complex queries that are needed in order to retrieve metadata for all FRBR levels (resource, expression, etc...). Although their proposed alternative modeling reduces complexity, it does so at the cost of interoperability. Considering, that linking is necessary to support the legal inquiry process across different jurisdictions, the proposed optimization needs to be built into the ELI and ECLI standards and adopted by the participating EU member states.

European Case Law Identifier

The European Case Law Identifier (ECLI) [Council of the European Union, 2011] has been created to introduce an identifier for *case law*, and to define a minimum set of metadata for judiciary documents (e.g. court decisions). The ECLI does not define any specific classes and uses the properties of the Dublin Core Metadata Initiative (DCMI)[22] ontology with the prefix `dcterms`. In contrast to the ELI, there is no separate formal ontology specification provided by the EU, but rather only a recommendation of nine mandatory (listed in Table 2.2) and eight optional properties, which should be used to describe metadata relating to the documents. Moreover, the ECLI conclusion makes particular suggestions for the use of the `dcterms` vocabulary, for instance that the object of `dcterms:coverage` should be used for the country (or more closely defined location)

[22] `https://dublincore.org/`, last accessed 28-12-2020

where the court is seated. Unfortunately, these suggestions are given without explicit ontological commitments or formal axioms, e.g. in terms of explicit range restrictions. Although not explicitly mentioned, the ECLI also contains properties for which the creation of country-specific lists seems appropriate. For example, especially for the properties with a corresponding intention such as `dcterms:type`, which is used to describe the document type and `dcterms:subject` used to describe the legal area.

The syntax of the ECLI identifier is more restricted compared to the ELI as it consists of five components separated by a double colon, for instance `ECLI:AT:OGH0002:2018:0100OB00060.17X.0220.000` is the ECLI identifier used for a decision of the Austrian Supreme Court. The order of the components is fixed and starts with the abbreviation `ECLI` and is followed by a country code, for instance `AT` for Austria, or `EU` for the courts of the European Union. The third component is the court code of the deciding court (`OGH002`), which is individually assigned by each participating country and the year of the decision (`2016`). The last component (`0100OB00060.17X.0220.000`) is an unique ordinal number of the decision, which translates to the official case number *10Ob60/17x* assigned to this case in the Austrian legal information system.

EuroVoc

Since 1982 the Publications Office of the European Union has been publishing and regularly updating the *EuroVoc*[23] thesaurus. EuroVoc is a multi-domain and multilingual thesaurus initially created as an indexing tool for the processing of documents from the EU authorities [Publications Office of the European Union, 2021]. The EuroVoc thesaurus is based on the Simple Knowledge Organization System (SKOS)[24], a well-known standard to represent information using RDF. The latest version of the EuroVoc thesaurus has been published on 18th December 2020 (V4.12) and is available for download as Resource Description Framework (RDF) or Extensible Markup Language (XML), as well as accessible via a SPARQL endpoint[25]. EuroVoc contains more than 6,000 terms (also called concepts, classes and descriptors) in the languages of the EU member states and therefore supports multilingual search. EuroVoc can also be used by national authorities but the hierarchy and available terms are deliberately generic in nature and might not be sufficient to meet the requirements of the envisaged national use case out of the box. The thesaurus is organized into 21 domains (for instance "law" and "trade"), 127 microthesauri and more than 600 top terms. With an exception of the geography domain, there is no polyhierarchy. This means that each term belongs to one superior term only. The terms are manually assigned to a class and in cases where a term would fit multiple classes, assigned to the class that seems to be the most natural fit for users. This might lead to confusion as, for instance, "commercial law" belongs to the domain "trade", "company law" to the domain "business and competition" and "civil law" to the domain "law". However, from a legal perspective all the mentioned terms are referring to a legal area for which the "law" domain seems to be the best fit.

The hierarchy and naming of terms in EuroVoc is illustrated in Listing 2.4. The individual terms are of type `skos:Concept` and a collection of concepts, a microthesaurus, is of type `skos:ConceptScheme`, for instance `ev:100195`. Each term has one preferred term (`skos:prefLabel`) and optional multiple non-preferred terms (`skos:altLabel`),

[23] `https://op.europa.eu/s/n3kP`, last accessed 02-01-2020
[24] `https://www.w3.org/2004/02/skos/`, last accessed 28-12-2020
[25] `http://publications.europa.eu/webapi/rdf/sparql`, last accessed 28-12-2020

Listing 2.4: Example EuroVoc

```
# Microthesaurus
<http://eurovoc.europa.eu/100195>
    a skos:ConceptScheme;
    skos:prefLabel "2026 Verbrauch"@de, "2026 consumption"@en .
# Topterm
<http://eurovoc.europa.eu/138>
    a skos:Concept;
    skos:inScheme ev:100195 ;
    skos:topConceptOf ev:100195 ;
    skos:narrower ev:2836 ;
    skos:prefLabel "Verbraucher"@de, "consumer"@en .
# Term
<http://eurovoc.europa.eu/2836>
    a skos:Concept;
    skos:inScheme ev:100195 ;
    skos:broader ev:138 ;
    skos:prefLabel "Verbraucherschutz"@de, "consumer protection"@en ;
    skos:altLabel "Maßnahmen zugunsten der Verbraucher"@de, "consumer rights"@en .
```

hence providing support for synonyms and colloquially used terms. The example shows that the EuroVoc term labeled *"consumer protection"@en* (`ev:2836`) has preferred and non-preferred labels, and a broader concept `ev:138` labeled *"consumer"@en* (`skos:broader`). The predicate `skos:topConceptOf` of concept `ev:138` indicates that it is a top term of the microthesaurus labeled *"2026 consumption"@en* (`ev:100195`). We can also see that the hierarchy within a microthesaurus is organized using the predicates `skos:narrower` and `skos:broader` and each concept is also directly linked to its microthesaurus with the predicate `skos:inScheme`.

Common Data Model

The Publications Office of the European Union (OP) uses the Common Data Model (CDM)[26] for their published resources, which is based on FRBR [Francesconi et al., 2015, Publications Office of the European Union, 2020a]. The resources that can be accessed via the EUR-Lex SPARQL endpoint are represented using the CDM ontology rather than the ELI and ECLI ontology. An RDF dump of the EUR-Lex data using ELI, up until 2018, is available on the EU Open Data Portal[27]. The usage of the CDM ontology results in using a different identifier for the documents in the EUR-Lex database CELLAR, the repository of the EU Publications Office, instead of the ELI identifier. A mapping between CELLAR and ELI identifiers is however provided using the predicate `owl:sameAs`.

2.4 Natural Language Processing

Legal documents are typically composed of mainly text, which humans can understand but cannot be directly processed by machines [Allahyari et al., 2017]. We therefore need methods to enable machines to work with text. The research area dealing with this matter

[26] `https://op.europa.eu/en/web/eu-vocabularies/model/-/resource/dataset/cdm`, last accessed 2021-01-15

[27] `http://data.europa.eu/88u/dataset/eli-european-legislation-identifier-eurlex`, last accessed 2021-01-15

is called *Natural Language Processing (NLP)*. More precisely, NLP can be defined as "*Natural language processing employs computational techniques for the purpose of learning, understanding, and producing human language content.*" [Hirschberg and Manning, 2015]. In order to use a machine for the automatic processing of natural language, we will briefly introduce the modeling of a language and the different NLP tasks, approaches and frameworks.

2.4.1 Language Models

A language is a "*system of communication*" [Collins Dictionary, 2021], which typically occurs in oral or written form as sequences of words. While using a language is a natural way of communication between humans, machines require a different representation of a language in the form of a model, which is defined as follows: "*A language model is a function that puts a probability measure over strings drawn from some vocabulary.*" [Manning et al., 2008]. A vocabulary is the collection of all words from a corpus, which can be a single document or a collection of documents. The vocabulary is created by tokenizing the sequences of words into individual parts, so called *tokens*, which are grouped sequences of characters that belong together [Manning et al., 2008]. For instance, the word sequence "European Court of Justice" is tokenized into the individual tokens "European", "Court", "of" and "Justice".

The most simple language models are statistical models based on a sequence of words, which are called *n-grams* [Manning et al., 2008], where *n* denotes the number of words considered for probability calculation. For example, a *unigram (n=1)* consists of only a single word (e.g. "European"), a *bigram (n=2)* consists of two words (e.g. "European Court"), a *trigram (n=3)* consists of three words (e.g. "European Court of") and so forth. Such n-grams can be used to calculate the probability of words given the preceeding word(s), for instance the word "European" is likely to be followed by the word "Union" or "Court", the sequence "European Court" to be followed by "of Justice" or "of Human Rights". The quality of such a language model depends on the size of the vocabulary. However, the drawback is the size of the model when all co-occurrences in a large corpus need to be stored and processed.

A different way is the representation of words in dense vectors, which are called *word embeddings*. State-of-the-art language models are based on neural networks and are able to capture the semantics of a text enabling a calculation of semantic similarities between words. Examples for such language models are for instance *word2vec* [Mikolov et al., 2013a, Mikolov et al., 2013b], which can be used in two flavors, either predicting a word given a context or predicting the context given a word. However, static word embeddings like word2vec have the disadvantage that only one vector is created for each word in a corpus. This means that these vectors are not context-dependent. Contextual word embeddings [Peters et al., 2018] are able to capture the context in the language model generation process and are therefore able to provide context-dependent vectors. Prominent contextual embeddings are trained on transformer-based architectures, for instance BERT [Devlin et al., 2019] and derivatives like DistilBERT [Sanh et al., 2019].

In order to properly train a language model, a large corpus of text is required. Luckily, a number of pretrained language models are provided for the different types of language models with an emphasis on the English language and generic texts. Many language models are pretrained on news articles or the content of knowledge bases such as

Wikipedia. Such a pretrained model might be sufficient for generic tasks, but the legal language is specific in terms of formulations and semantics of words. On the one hand, pretrained models trained on large corpora are available, but trained on generic corpora. On the other hand, legal corpora are not large enough to properly train a domain-specific language model. In order to overcome this problem, a transfer learning method such as ULMFiT [Howard and Ruder, 2018] can be applied. The basic idea is to build on a generic language model and fine-tune it with a domain-specific corpus.

2.4.2 Tasks

There are a number of different tasks in NLP research [Collobert et al., 2011], we focus on the tasks used in this thesis, namely *Named Entity Recognition (NER)* and *text classification*.

Named Entity Recognition (NER) is the process of extracting named entities from a text and classifying them into categories [Grishman and Sundheim, 1996]. Named entities are understood as expressions for objects sharing the same proper names [Nadeau and Sekine, 2007]. Such named entities can be generic, for instance person, place and organization, but also domain-specific, for instance, judge, judicial district and court. Furthermore, named entities are also sequences that are commonly not considered being entities, such as temporal expressions (e.g. a date) and can be of arbitrary length, hence covering a *span* of tokens, for instance the sequence “22 July 2005” consists of three tokens, which forms one named entity “date”.

Text classification is the process of assigning one or more categories out of a set of categories to a given text [Sebastiani, 2002]. Text classification belongs also to the research area of text mining [Allahyari et al., 2017]. Text classification is sometimes also called *document classification*. We note that “text” (and “document”) do not indicate a specific format of the text nor its length. In fact, text indicates a sample that is used in the particular task and can be a single sentence, paragraph or a document such as a legal document consisting of multiple pages.

2.4.3 Approach

We distinguish between three types of NLP approaches and all of them have their advantages, disadvantages and have been employed in NLP tasks on legal documents. All of these approaches can be used for the aforementioned tasks and in combination with the language models.

Rule-based approaches use predefined rules, which are usually created by humans tailored to a specific task and corpus. This approach can be used for named entity recognition and text classification tasks. For example, the rule to detect a named entity of type “legal rule” in a court decision can be defined as “the string ‘RS’ is followed by a number”. The concrete representation of a rule depends on the used engine to enforce the rules. Rules are easy to understand, to update and to fix in case of errors, but might get complex easily and tedious to create [Chiticariu et al., 2013].

Machine learning-based approaches use the input data to automatically search for patterns and relations based on the NLP task as well as the chosen machine learning algorithm. We use a supervised learning approach to solve NLP tasks, which means that the algorithm takes a labelled set of training data to build a model for the NLP task. For example, in the NER task a named entity of a specific type is annotated as a member of this type. Machine learning-based systems are not as rigid as rule-based systems (a rule fires or not), but need a large set of training data and are not so simple to understand [Chiticariu et al., 2013].

Deep learning-based approaches use a neural network, which consists of interconnected processing units and can be arranged in different ways. Deep learning approaches are more powerful than machine learning approaches as they are able to learn from raw data and are less dependent on carefully preprocessed training data [LeCun et al., 2015]. Furthermore, *Recurrent Neural Networks (RNN)* are often applied in NLP tasks, because they are suitable in detecting sequences as text is a sequence of symbols (words and delimiters, numbers, etc.).

Over time, all of these approaches have been used to solve entity extraction and classification tasks in the legal domain and often contain a performance comparison of different approaches to each other (e.g. [Dozier et al., 2010, Steinberger et al., 2012, Boella et al., 2015, Dragoni et al., 2017, Angelidis et al., 2018, Leitner et al., 2019, Chalkidis et al., 2019]).

2.4.4 Tools and Frameworks

In order to execute an NLP task or to create and use a language model, we can rely on existing NLP tools and frameworks. When there is no pretrained corpus available, which can be used, it is necessary to create a corpus with own data. For example, in order to extract legal entities from legal documents, the documents need to be annotated first. A tool supporting the annotation process for a NER task is, for instance, the General Architecture for Text Engineering (GATE) [Cunningham et al., 1999] with a graphical user interface allowing users to directly mark and tag a sequence of words with a class. Furthermore, GATE also contains an engine to create rules and supports the evaluation of rule-based approaches.

A popular framework for machine learning is scikit-learn[28], which supports many different machine learning algorithms commonly applied to NLP tasks, for instance Support Vector Machines (SVM) [Noble, 2006] and Random Forests (RF) [Breiman, 2001]. For Conditional Random Fields (CRF) [Lafferty et al., 2001] the sklearn-crfsuite[29] provides a handy tool.

The state-of-the-art approach to NLP tasks is using neural networks that can be either implemented manually or using deep learning frameworks, for instance fast.ai[30] and Flair[31]. The advantage of using these frameworks is that they are already fully implemented and support a range of NLP tasks out of the box. The concrete supported task might vary from framework to framework. Furthermore, these frameworks usually also support the integration of pretrained language models and some of them also the creation of new language models. As their big advantage, they can be applied without going deep into neural networks and concentrate on the actual NLP task.

[28] `https://scikit-learn.org/stable/`, last accessed 2021-03-12
[29] `https://sklearn-crfsuite.readthedocs.io/en/latest/`, last accessed 2021-03-12
[30] `https://www.fast.ai/`, last accessed 2021-03-12
[31] `https://github.com/flairNLP/flair`, last accessed 2021-03-12

CHAPTER 3

Legal Knowledge Graph Modeling

In this chapter, we describe the creation process of a legal knowledge graph for the Austrian legal system. In Section 3.1, we start with an introduction of the current Austrian legal information system. We describe the data that can be found there as well as the challenges users face when searching for legal information. This is important for the knowledge graph creation process described in Section 3.2, which is not only based on the ELI and ECLI ontologies (cf. Chapter 2), but also on the available data in RIS. In order to properly represent Austrian legal data using ELI and ECLI, we propose the *Legal Knowledge Graph Ontology (LKG)* and extend the ELI and ECLI ontologies with the required classes and properties in Section 3.3. In Section 3.4, we introduce the Austrian vocabulary *AustroVoc*, which contains Austrian-specific terms and concepts. Finally, the related work is presented in Section 3.5 and Section 3.6 summarizes the chapter including a view on possible future research directions.

3.1 Austrian Legal Information System

The *Rechtsinformationssystem des Bundes (RIS)*[1] is the legal information system of the Republic of Austria. RIS serves as a single point of information from which legal documents issued by various authorities can be searched and accessed. In addition to the web interface, RIS also provides access to its data via a REST API[2], which enables users to access RIS data in JSON[3]. Through the web interface different backend databases – subdivided into different parts of the legislation – such as “Bundesrecht” (federal law), “Landesrecht” (state law of the nine Austrian states) or “Judikatur” (judiciary) and many more – can be accessed. Documents in RIS can be retrieved in different formats like HTML, XML, RTF (Rich Text Format)[4] and PDF. Although the RIS web interface gives the impression that it is a single database containing all legal information, it is in fact a collection of independent databases, which are not currently connected nor interlinked underneath.

Figure 3.1 shows the RIS search interface for judiciary documents, in particular for documents of the supreme court and subordinated courts. The screenshot illustrates

[1] `https://www.ris.bka.gv.at/`, last accessed 2021-01-15
[2] `https://data.bka.gv.at/ris/api/v2.5/`, last accessed 2021-01-15
[3] `https://tools.ietf.org/html/rfc8259`, last accessed 2021-01-15
[4] `http://msdn2.microsoft.com/en-us/library/aa140277(office.10).aspx`, last accessed 2020-12-28

Figure 3.1: Austrian Legal Information System (RIS) Interface

that RIS looks a single database but the menu on the top of the page ("Bundesrecht", "Landesrecht", "Gemeinderecht", "Judikatur", etc.) shows the different databases that can be accessed and searched. Within these categories, further subcategories can be selected. For example, federal law gazettes and legal provisions can be accessed via "Bundesrecht" and the court decisions of different courts (Supreme Court, Constitutional Court, etc.) can be accessed via "Judikatur". As shown, the search interface provides the possibility to enter multiple search parameters, which can be used on their own or in combination. Keywords ("Suchworte") can be used for a very broad search and users can also select whether the search should only be conducted in the legal rules ("Rechtssätze (RS)"), in the decision texts ("Entscheidungstext (TE)") or both. Additionally, the search can be restricted to the decision date ("Entscheidungsdatum") of the judgments, new ("Neu im RIS seit") and updated documents by date ("Änderungen seit"). Furthermore, it is possible to search for decisions of a specific court ("Gericht") and a specific legal provision ("Norm"). All of these search possibilities return a (possibly long) list of search results. The search option based on a case number ("Geschäftszahl") is a very specific search and returns the documents found for the entered case number. The search interfaces for law gazettes and legal provisions follow the same principles, thus allowing users to search for keywords and additional filters to restrict the search space. Now, we can revisit the example questions to explain the drawbacks of the current Austrian legal information system.

3.1.1 Example Questions

Q 1 *Which documents are referenced in a specific court decision?*

Court decisions are based on the law and are, besides other documents such as legal rules, referenced in court decisions. Although it is possible to search court decisions for a given legal provision or to retrieve all documents with a specific legal provision via keyword search, it is not possible to retrieve all referenced documents for a specific court decision, optionally with the document text. Instead, users have to read the court decision and look up every referenced legal provision separately in the respective database.

Interlinking the referenced documents and allowing users to retrieve the court decision with (the content of) the referenced documents would reduce the time spent for and increase the efficiency of the legal information search process.

Q 2 *Over which districts does a court have competent jurisdiction?*

The competent jurisdiction of a court is also dependent on spatial aspects, for instance where a specific property is located. Currently, it is only possible to search for decisions of a specific court. In order to get the information about the court having competent jurisdiction over a specific spatial entity (e.g. a village), users have to consult the website of other authorities as this information is not contained in RIS. Furthermore, it also not possible to search for cases that occurred in a specific spatial area. A workaround is to enter the spatial area as a keyword, which is only successful when the area occurs literally in the document.

Integrating spatial and court information would enable enhanced search queries such as tracking appeal stages or analyzing court decisions on a spatial level. Furthermore, we consider courts and their competent jurisdiction as an essential part of legal information and therefore it should be part of a legal information system.

Q 3 *What are the national transpositions of a specific EU directive?*

An EU directive is a legal act that needs to be transposed into national law. The RIS contains this information as part of the metadata in § 0 of a law, where all changes to the entire law (which means all changes to all legal provisions of this law) are listed. A service[5] provided as part of the European EUR-Lex database does not include the actual transposed texts or links to the national documents. As a result, users have to consult the national legal information systems to find the corresponding national transpositions. In the worst case, the national legal information systems might not be available for the public or do not contain this information at all.

Including the information about national transpositions with their actual texts in the legal knowledge graph could enable a cross-jurisdictional search of EU legal acts and opens new possibilities for comparative law.

Q 4 *Which legal documents regulate a specific legal area searched with keywords in a foreign language?*

Legal information is typically provided in the official language(s) of the issuing country. Sometimes, the "most important" laws (e.g. the constitution and the civil code) are also provided in additional languages. This requires users to find the appropriate term in the foreign language in order to conduct a search process.

Integrating a thesaurus such as EuroVoc, which contains terms in multiple languages could enable the search across multiple languages. Hence, users could enter a keyword in one language (e.g. Italian) and retrieve documents in another language (e.g. German).

[5] `https://eur-lex.europa.eu/collection/n-law/mne.html`, last accessed 2021-02-15

Q 5 *Which events are mentioned in a court decision and could be used for a quick overview of the case?*

Court decisions are potentially very long and although they typically follow a common structure, at least per court or jurisdiction, reading these documents is a very time-consuming process. At the moment, these decisions are presented as a continuous text describing what happened in the particular case, for instance that the applicant filed a complaint or the applicant married someone. The same applies to contracts or large collections of court or law firm internal files where it could be helpful to know what happened when. Users have to go through the entire documents and search or remember such events.

Extracting events from court decisions and presenting them in a timeline results in a quick overview about the events including additional information such as temporal aspects, the acting subject and the actual description of what happened.

3.1.2 Challenges

Primary challenges, based on the current Austrian legal information system and the sample queries in order to facilitate the answering of such complex questions in a more automated manner, include the following:

Information Overflow. Search results are presented in a long list. For example, a search for the keyword "Auto" (car) returns almost 1,500 documents covering different legal areas ranging from issues with the sales contract, car accidents and even about the lawfulness of GPS surveillance of employee cars. Our objective is to enhance the search process by providing additional search possibilities, for instance by classifying legal documents into categories allowing users to reduce the search space.

Unstructured and Missing Information. Information about legal documents can be contained in both structured metadata but also within unstructured text, for instance law references in court decisions are not contained in metadata. Further, some connections between documents are only implicitly available in the text and while these can be detected by a human reader, a machine would struggle with the same task. In addition, the mandatory and optional properties within the ELI and ECLI ontologies can only be partially constructed from the document metadata alone.

Data Silos. The Council of the European Union identified the need to disseminate and exchange legal information[6] across the EU member states. Unfortunately, the exchange of legal information is hindered by different legal systems and incompatibilities of national legal information systems. At the moment, these legal information systems are still separate silos. Our objective is that two steps will help to reduce the problem of data silos. In a first step, linking legal data nationally across so far disconnected backend databases reduces the data silos on a national level. This is followed by a second step to interlink legal data across Europe leading to a reduction of data silos on a European level. It is worth noting that automatic extraction from and linkage of existing databases

[6]2011/C127/01, 2012/C325/02: Identification of needs

should avoid the need to maintain the same information at multiple places, while also allowing the data to be easily integrated with other sources.

Redundant Data Storage. Considering that legal documents contain references to each other, the legal information search process typically involves searching across different databases. At the moment, additional information that should be made available for search but is not part of the particular database, is stored in an additional database column just to enable search. Still, this leads to redundant data storage and does not add any beneficial additional information except enabling search. Furthermore, this situation results in anomalies, which must be considered on insert, update and delete operations. Linking data across databases helps to avoid these anomalies as it does not require to store the same information redundantly at multiple places and therefore provides more flexibility.

3.1.3 Requirements

From the challenges outlined above, we derive three core requirements for the creation of a legal knowledge graph.

Extraction. It must be possible to extract information that is missing in the metadata from the document text. For this purpose, it is required that the documents, from which additional information needs to be extracted, are accessible and available in an ideally open and structured format. Furthermore, information extraction techniques need to be applied, which are tailored to the legal domain.

Integration. We need to integrate legal data from various national and international data sources into a single knowledge base. For such a successful integration of legal data across multiple sources, it is required to agree on using the same data model. Moreover, the data needs to be provided by all authorities and to be publicly accessible on the web, without any additional restrictions.

Normalization. It is necessary to represent legal information in a consistent way, both on a national level and international level. This enables the integration of legal data from different sources and can be achieved by using unique and structured identifiers for referencing legal documents instead of plain text references to avoid redundancies and inconsistencies.

3.2 Legal Knowledge Graph Creation Methodology

Legal ontologies, such as ELI and ECLI, serve as an input for the legal knowledge graph creation process as they are the basis for the modeling of the knowledge graph. In the first step, we model the ontology to represent the Austrian legal system based on ELI and ECLI and create a national thesaurus *AustroVoc* in order to encode Austrian specific terms, not covered in existing terminologies such as EuroVoc. Since ELI and ECLI are only describing a minimum set of metadata in order to be applicable to all EU member states, we need to create additional classes and properties for our legal knowledge graph to reflect Austrian specific requirements. In our particular case, we can build on the already existing ELI and ECLI ontologies. However, on the one hand the existing ontologies are in parts not fine-grained enough. On the other hand, legal documents and their metadata, available in the Austrian legal information system, provide us with additional required information on the missing parts. Therefore, we extend these ontologies in a middle-out fashion (cf. Section 2.3). This approach has also been described to be effective in the legal domain in a similar setting with existing legal ontologies that are extended based on underlying legal documents [Ghosh et al., 2016].

In the bottom-up phase, we analyze the available metadata and the data extracted from the Austrian legal documents, which could be used to find the additional classes and properties to be added. For this process, we keep in mind that our primary goal is interlinking of the documents, rather than describing the actual content of the documents. For instance, we already know that the Austrian legal information system contains different kinds of documents in their databases. It is therefore appropriate to create additional classes for these different kinds of documents such as the law gazettes, legal provisions and court decisions. The same procedure also applies to the properties derived from the documents and their associated metadata. For example, specific document types also have special properties like dates when a bill has passed a council.

After the bottom-up phase, we investigate the existing ELI and ECLI ontologies and review their classes and properties in the top-down phase. The analysis reveals that the ELI ontology only contains a single class (`eli:LegalResource`) for all legislative documents but no more fine-grained classes, while the ECLI ontology does not provide a class for judiciary documents at all.

Finally, we matched the classes and properties derived in the bottom-up phase with the appropriate classes and properties in the top-down phase, hence we refined and extended classes, properties, as well as taxonomic terminologies/thesauri, where needed.

Based on the resulting combined ontological schema, the resulting model has been populated with data from RIS and linked to external knowledge bases. In a final step, we integrate external legal data from the European Union, the European thesaurus *EuroVoc* containing terms from different domains in the official languages of the EU member states and also legal data from selected other countries.

3.3 Legal Knowledge Graph Model

Since both ELI and ECLI are targeting a variety of different legal systems within the EU member states, they only provide two classes of legal documents, which we extended in order to represent specific legal document types used in Austria's national legal publication process, such as law gazettes and legal provisions. In our examples herein, we exemplify our legal knowledge graph with a focus on federal law as well as jurisdiction by the justice branch, which includes decisions of the supreme court and lower courts. Figure 3.2 depicts our legal knowledge graph model with the specific classes we added colored gray. Nodes denote classes and edges represent properties connecting their respective domain and range classes.

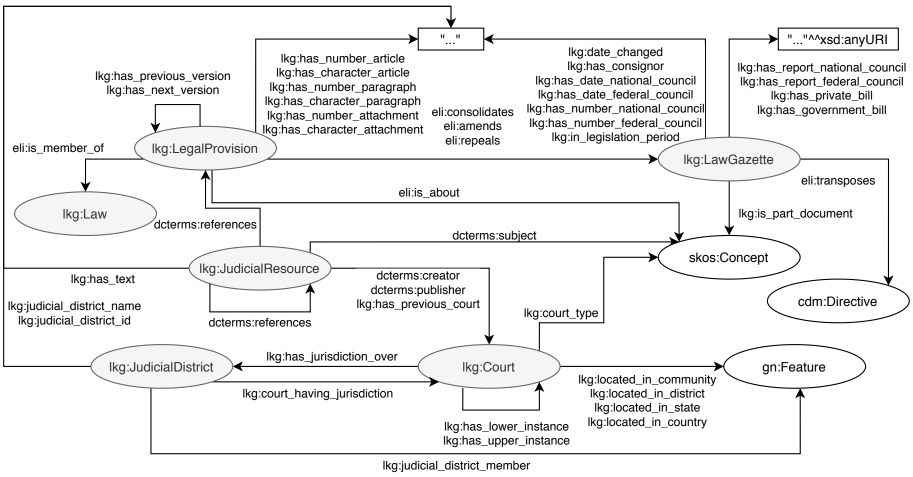

Figure 3.2: Legal Knowledge Graph Model

Listing 3.1: Example Law Gazette (new properties)

```
<https://www.ris.bka.gv.at/eli/bgbl/I/2016/35/20160608>
       a lkg:LawGazette ;
       lkg:date_changed
              "2016-06-08"^^xsd:date ;
       lkg:has_consignor
              "BMASK (Bundesministerium für Arbeit, Soziales und Konsumentenschutz)" ;
       lkg:has_date_federal_council
              "2016-05-12"^^xsd:date ;
       lkg:has_date_national_council
              "2016-04-28"^^xsd:date ;
       lkg:has_government_bill
              "http://www.parlament.gv.at/PG/DE/XXV/I/I_01059/pmh.shtml" ;
       lkg:has_number_federal_council
              853 ;
       lkg:has_number_national_council
              126 ;
       lkg:has_report_federal_council
              "http://www.parlament.gv.at/PG/DE/BR/I-BR/I-BR_09579/pmh.shtml" ;
       lkg:has_report_national_council
              "http://www.parlament.gv.at/PG/DE/XXV/I/I_01095/pmh.shtml" ;
       lkg:in_legislation_period
              "XXV" ;
       lkg:is_part_document
              av:leg_bg .
```

3.3.1 Law Gazette

A law gazette is used to publish new laws or any changes to existing laws, which happen in editorial instructions (e.g. "in § X change amount Y to Z"). As exemplified in Listing 3.1 for law gazette *BGBl I Nr. 35/2016*, we represent a law gazette with class `lkg:LawGazette` (subclass of `eli:LegalResource`). We introduce new properties to provide background information about the legislative process, which is a useful source that is often used to solve legal interpretation problems. These properties cover dates when law changes have been discussed in the national and federal councils (`lkg:has_date_national_council`, `lkg:has_report_national_council`) and links

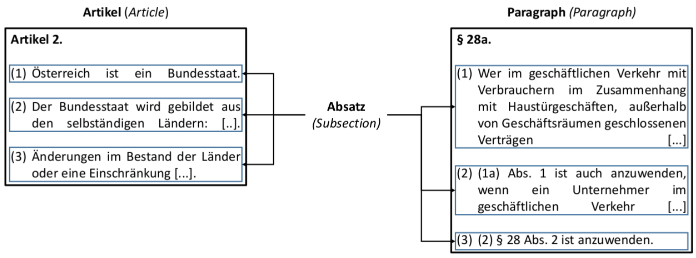

Figure 3.3: Legal provision naming convention

Listing 3.2: Legal Provision §28a Consumer Protection Law (new properties)

```
<https://www.ris.bka.gv.at/eli/BGBl/1979/140/P28a/NOR40180997>
        rdf:type
                lkg:LegalProvision ;
        lkg:has_number_paragraph
                28 ;
        lkg:has_character_paragraph
                "a" ;
        lkg:has_next_version
                ris:eli/BGBl/1979/140/P28a/NOR40192489 ;
        lkg:has_previous_version
                ris:eli/BGBl/1979/140/P28a/NOR40173437 .
```

to the reports about the parliamentary discussion[7], which are available on the web (`lkg:has_report_national_council`, `lkg:has_report_federal_council`). These reports are useful in case there is a loophole in the law and the will of the parliament needs to be discovered. Bills initiate the legislative process and are linked using the properties `lkg:has_private_bill` and `lkg:has_government_bill`. The authority bringing in a bill is indicated with the property `lkg:has_consignor`. We use `lkg:is_part_document` to determine the type of the law gazette such as "constitutional law" or "order". The legislation period in which a law gazette has been published is included for legal analysis and is indicated with the predicate `lkg:in_legislation_period`.

3.3.2 Legal Provision and Law

A `lkg:LegalProvision` (subclass of `eli:LegalResource`) is a resource containing the actual norm. In Austria, each legal provision is an individual document with a *NOR* number used as a unique technical identifier, for instance "NOR40180997" (see Listing 3.2) and a label used in legal practice, for instance "§ 28a KSchG" (Paragraph 28a of the Consumer Protection Law). Figure 3.3 shows the legal provisions "Artikel 2 B-VG" (Art. 2 of the Constitution) and *§ 28a KSchG*. A legal provision can be labeled "Artikel" (article) or "Paragraph" (paragraph) and is always seen in its entirety for modeling, irrespective of whether there is only one "Absatz"[8] (subsection) or multiple subsections.

Listing 3.2 depicts an RDF snippet for legal provision *§ 28a KSchG* with the new properties we introduced in our extended `lkg:` ontology. Besides the "Artikel" and "Paragraph" there is also a "Anlage" (attachment) usually used for transitional provisions,

[7]Publicly available at the Austrian parliament's website: `https://www.parlament.gv.at/`

[8]The English translation of "Absatz" is "paragraph", but we call the "Absatz" subsection to avoid confusion, as the word "Paragraph" in Austrian/German legal language rather refers to law articles.

Listing 3.3: Judicial Resource (new properties)

```
<https://www.ris.bka.gv.at/Dokumente/Justiz/JJT_20180220_OGH0002_0100OB00060_17X0000_000/
      JJT_20180220_OGH0002_0100OB00060_17X0000_000.html>
       rdf:type
              lkg:JudicialResource ;
       dcterms:creator
              <http://data.wu.ac.at/legal/lkg/court#court_1> ;
       lkg:has_previous_court
              <http://data.wu.ac.at/legal/lkg/court#court_2> ;
       lkg:has_text
              "OGH 20.02.2018 10Ob60/17x Der Oberste Gerichtshof hat [...]" .
```

which combines both "Artikel" and "Paragraph", for instance "Artikel 1 § 1". We introduce new properties to model numbers as well as characters in the labels of legal provisions, for instance `lkg:has_number_paragraph` and `lkg:has_character_paragraph`. Analogously, for legal provisions named by article or attachment we use the properties `lkg:has_number_article`, `lkg:has_character_article` and `lkg:has_number_attachment`, `lkg:has_character_attachment` respectively. Two temporally subsequent legal provisions are linked with `lkg:has_next_version` and `lkg:has_previous_version`. We create the class `lkg:Law` because legal provisions can be a part of a *law book*, which is a collection of legal provisions containing regulations about the same topic. The membership between a `lkg:LegalProvision` and `lkg:Law` is indicated with the ELI property `eli:is_member_of`.

Legal provisions are the basis for court decisions and it is therefore important to link a judgment with the correct version of a legal provision. The linking between judgments and legal provisions is achieved by following a *date-based linking approach*, which links a judgment to the legal provision that is in force at the decision date because this will be the correct version most of the time. In cases, where the court has to apply a specific version of a legal provision, this is indicated in the court decision. Furthermore, a specific version of a legal provision is always the sum of the initial version with all its amendments over time and is called consolidated version.

3.3.3 Judicial Resource

The class `lkg:JudicialResource` (subclass of `frbroo:F1_Work`) is used for judiciary documents, which are modeled based on the ECLI suggestions. Austria assigns an ECLI identifier to judiciary documents as metadata, which is, in contrast to the ELI, not used as a web identifier. Furthermore, the Austrian judiciary documents cannot be searched in the ECLI search engine[9] of the European e-Justice Portal. We add the text of a court decision with the property `lkg:has_text`. The EU Publications Office (OP) provides *Named Authority Lists (NAL)*, which are vocabularies used to standardize the inter-institutional legal data exchange. Some of these NAL can be used by all countries, for instance the NALs for languages or countries, while other NAL are very EU-specific, for example court-types, which contain EU courts only and therefore cannot be used for national courts. We use these NALs for the ECLI properties that indicate in which country the deciding court is seated (`dcterms:`

[9] `https://e-justice.europa.eu/content_ecli_search_engine-430-en.do`, last accessed 2021-01-15

Listing 3.4: Example of court and judicial district for *Bezirksgericht Leopoldstadt*

```
<http://data.wu.ac.at/legal/lkg/court#court_8>
      rdf:type
         lkg:Court ;
      rdfs:label
         "Bezirksgericht Leopoldstadt" ;
      lkg:court_type
         av:bg ;
      lkg:has_jurisdiction_over
         <http://data.wu.ac.at/legal/lkg/judicialdistrict#judicialdistrict_900M> ;
      lkg:has_upper_instance
         <http://data.wu.ac.at/legal/lkg/court#court_3> ;
      lkg:has_upper_instance
         <http://data.wu.ac.at/legal/lkg/court#court_4> .
```

coverage), the language of the decision (dcterms:language) and the access rights (dcterms:accessRights). Properties populated with Austrian specific values, such as dcterms:type, dcterms:publisher, lkg:previousCourt, are linked with concepts contained in the *AustroVoc* thesaurus we created for this purpose. Listing 3.3 shows a snippet of the Austrian Supreme Court decision *10Ob60/17x*. The new predicate lkg:has_previous_court is used to reference to the higher regional court that took a decision in this case before it came to the supreme court, and the new predicate lkg:text is used to reference the actual text of the court decision.

3.3.4 Court and Judicial District

A judgment in the judiciary branch is rendered by a lkg:Court of a specific type indicated with lkg:court_type as shown in Listing 3.4. Furthermore, courts are organized in a hierarchical manner and have a higher instance indicated with the predicate lkg:has_upper_instance and a lower instance indicated with the predicate lkg:has_lower_instance. The location of a court is important as some legal matters define the competent court by the location of the affected object, for instance a property. A court is located in a community (lkg:located_in_community), district (lkg:located_in_district), state (lkg:located_in_state) and country (lkg:located_in_country). A district court also lkg:has_jurisdiction_over a geo-spatial entity of lkg:JudicialDistrict[10], which are different from political districts. Similarly, the property lkg:court_having_jurisdiction indicates the court having spatial competent jurisdiction, thus it relates the judicial district to the court. The competent jurisdiction is assigned to the lowest level of authorities, hence district courts. Since we know that a district court has competent jurisdiction over a particular area and that court has an upper instance, we can also infer that a higher court has competent jurisdiction over all areas of all lower courts assigned to the higher court. In order to represent spatial information about judicial districts and courts, we use the publicly available database *Geonames*[11], which provides identifiers and spatial information for locations in multiple languages as well as a small ontology (prefix gn:) to describe these properties.

[10] https://www.statistik.at/web_de/klassifikationen/regionale_gliederungen/gerichtsbezirke/index.html, last accessed 2021-01-15

[11] https://www.geonames.org/, last accessed 2021-01-15

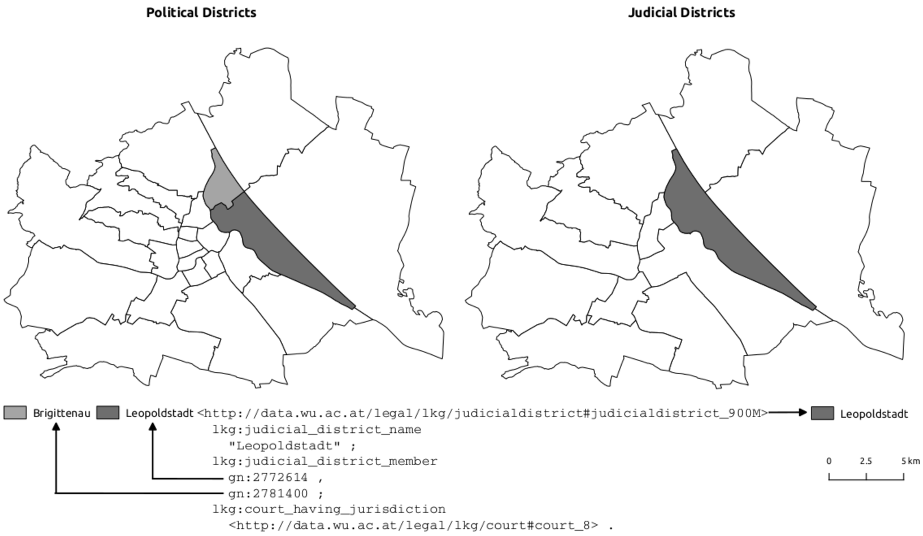

Figure 3.4: Illustration of political and judicial districts for the Austrian capital Vienna.

Figure 3.4 illustrates the difference between political and judicial districts for the capital of Austria, Vienna, which is divided into 23 political districts but only 12 judicial districts. The two political districts *Leopoldstadt* (`gn:2772614`) and *Brigittenau* (`gn:2781400`) are the members (`lkg:judicial_district_member`) of the single judicial district named (`lkg:judicial_district_name`) *Leopoldstadt*. The link between the judicial district and the court is established with the predicate `lkg:court_having_jurisdiction`, which is the inverse to `lkg:has_jurisdiction_over` as illustrated for the district court Leopoldstadt in Listing 3.4.

3.4 AustroVoc

We propose a SKOS-based thesaurus *AustroVoc* containing Austrian specific terminology. ELI and ECLI encourage member states to create their own schema for the properties indicating a document type (`eli:type_document` and `dcterms:type`) and a document classification to describe the content or legal area of a document (`eli:is_about` and `dcterms:subject`). We create three different schemes for *Gericht-typ* (court type), *Bundesrechtindex* (law index) and *Resource-typ* (resource-type).

3.4.1 Court Type

The court types provided in the NAL[12] of the EU Publications Office cannot be used 'as is' since they only contain EU courts. Thus we create an additional *court-type* scheme, which contains the different types of Austrian courts. We distinguish between public tribunals, for instance the constitutional court (`av:vfgh`), and ordinary courts, for instance the supreme court (`av:ogh`), which are responsible for different legal areas and are organized in a hierarchical way. Adding this information enables a search for

[12] `https://op.europa.eu/s/oFnE`, last accessed 2021-01-15

Listing 3.5: Law index example (shortened)

```
# Legal provision linked with law index
<https://www.ris.bka.gv.at/eli/BGBl/1979/140/P28a/NOR40180997>
   eli:is_about av:bri2006 .
# Law index
av:bri2006
   a skos:Concept;
   skos:broader av:bri20;
   skos:prefLabel "Konsumentenschutz"@de ;
   rdfs:seeAlso ev:2836 .
# EuroVoc term
<http://eurovoc.europa.eu/2836>
   a skos:Concept;
   skos:inScheme ev:100195 ;
   skos:broader ev:138 ;
   skos:prefLabel "Verbraucherschutz"@de, "consumer protection"@en ;
   skos:altLabel "Maßnahmen zugunsten der Verbraucher"@de, "consumer rights"@en .
```

judgments rendered by courts of a particular type and superior or subordinate courts and legal analysis.

3.4.2 Law Index

The *law index* is an index for Austrian federal law[13] provided by RIS, which organizes the law in a hierarchical manner. As shown in Listing 3.5, every legal provision is assigned to an entry in this index with the property `eli:is_about`, which allows users to search for legal provisions belonging to a specific legal area, for instance §28a KSchG is linked to the law index `av:bri2006`. We also use the law index to indicate the legal area of judgments dependent on the legal provisions they are based on using `dcterms:subject`. Finally, where possible (for details, see Section 4.4 below) we link the national law index items with corresponding items to the European thesaurus *EuroVoc* using the property `rdfs:seeAlso` to enable a multi-lingual search across jurisdictions. For instance, the AustroVoc law index `av:bri2006` (*"Konsumentenschutz"@de*) is linked to the EuroVoc concept `ev:2836` (*"Verbraucherschutz"@de*).

3.4.3 Resource Type

As with the court-types mentioned above, the resource-types contained in the NAL[14] are EU specific and incomplete as they do not contain specific resources used and required in Austria. We again created our own schema for such specific *resource-types* in RIS. These mainly include different document types, for instance judiciary documents can be subdivided into “Entscheidungstext” (decision text) or a “Rechtssatz” (legal rule), which is a case summary from which general legal rules can be inferred. The properties used to indicate the document types are already available in ELI (`eli:type_document`) and ECLI (`dcterms:type`). These properties used to indicate the document types are not to be confused with the property `rdf:type`, that is used to indicate to which class a document belongs to, for instance judiciary documents are of

[13] `https://www.ris.bka.gv.at/UI/Bund/Bundesnormen/IndexBundesrecht.aspx?TabbedMenuSelection=BundesrechtTab`, last accessed 2021-01-15

[14] `https://op.europa.eu/s/oFnF`, last accessed 2021-01-15

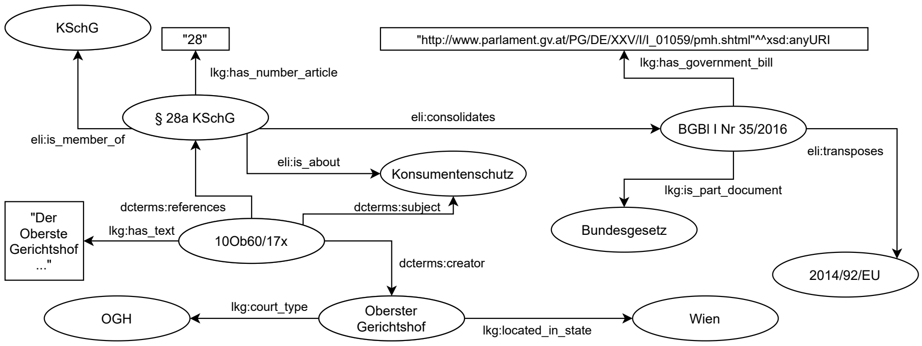

Figure 3.5: Legal Knowledge Graph Model Example

type `lkg:JudicialResource` and legislative documents are of type `lkg:LawGazette` or `lkg:LegalProvision`.

Figure 3.5 shows a sample knowledge graph resulting from our extraction/population methods, which we describe in the following chapter. The sample knowledge graph shows a court decision with case number *10 Ob 60/17x* (a `lkg:JudicialResource`), which has been decided by (`dcterms:creator`) the Austrian Supreme Court *Oberster Gerichtshof* (`lkg:Court`), is of type `lkg:ogh` and located in *Wien* (`lkg:located_in_state`). The court decision references (`dcterms:references`) § 28a KSchG (a `lkg:LegalProvision`). Both, the court decision and the legal provision, have been classified with the topic "Konsumentenschutz" (`skos:Concept`) from the AustroVoc thesaurus. Furthermore, the legal provision belongs to the law (`lkg:Law`) *KSchG* and consolidates (`eli:consolidates`) the changes as announced in the `lkg:LawGazette` *BGBl I Nr 35/2016*. The law gazette is based (`lkg:has_government_bill`) on a government bill available on the website of the Austrian parliament to transpose the European Union Directive *2014/92/EU* (a `cdm:Directive`).

3.5 Related Work

The goal of this section is to present related work that has been carried out by other researchers in terms of legal information exchange and other semantic technology based initiatives in the legal domain including work beyond legal knowledge graphs.

Several works also use ELI and ECLI to interlink legal documents, for instance Nomothesia [Chalkidis et al., 2017] (Greece) and Semantic Finlex [Oksanen et al., 2019] (Finland), which describe their approach to the interlinking of legal information for a particular jurisdiction. Similar to our work, the Semantic Finlex project also extends the ELI and ECLI ontologies, while the Nomothesia project includes the ELI and ECLI ontologies. The EUCases [Boella et al., 2015] project creates linked legal data from XML sources and the Lynx [Montiel-Ponsoda et al., 2017] project focuses on compliance. We used these works as an inspiration for the creation of our knowledge graph and provide are more detailed analysis in Section 6.2.

Several formats have been proposed enabling or simplifying the exchange of legal information in a structured and standardized manner. [Boer et al., 2002] described the

XML standard MetaLex, which can be used to encode the structure and the content of legal documents. Another open and extensible XML standard for the exchange of legislative and judiciary documents is Akoma Ntoso[15], which provides schemes for the structure and metadata of legal documents. Other standards for the XML-based exchange of legal information are, for instance, LegalDocML TC[16], which is based on Akoma-Ntoso aiming at the creation of a standard for a worldwide exchange of legal information using a standardized set of metadata. LegalRuleML [Palmirani et al., 2011, Athan et al., 2013] focuses on the expression of rules and constraints in the legal domain in XML format. These proposed standards are mainly used for the markup of legal documents and to structure their metadata, as well as content, in a standardized way, which should enable an easier exchange of legal information. A legal document and knowledge management system building on such XML standards to represent legal information is Eunomos [Boella et al., 2016], which can be used to support the legal information search process. XML files can also be used as a data source for the creation of a legal knowledge graph as shown by [Junior et al., 2019], which use XML to RDF mapping languages. In contrast to these works, we use RDF and the ELI and ECLI ontologies for our legal knowledge graph.

Complementary work mostly focusing on ontologies specific for particular legal domains has been carried out in the past. The Legal Knowledge Interchange Format (LKIF) proposed by [Hoekstra et al., 2007] is an ontology, which supports the interchange of legal information between different legal systems modeling the semantics contained in the text of legal documents [Boer et al., 2008]. A summary of existing legal ontologies has been published by [Breuker et al., 2009]. The authors compare 23 ontologies and categorize them by application (information retrieval, general language for expressing legal knowledge,...), type (knowledge representation) or character (general vs domain-specific). A recent extensive study conducted by [de Oliveira Rodrigues et al., 2019] analyzes legal ontologies found in various digital libraries based on multiple dimensions, such as formalization, legal theories, semantic problems and ontology engineering problems in a systematic manner. [Leone et al., 2019] conducted a review of legal ontologies and classify legal ontologies according general, modeling and semantic information. [de Oliveira Rodrigues et al., 2019] and [Leone et al., 2019] show in their studies, that a large number of legal ontologies have been proposed over time and are available for reuse. Unfortunately, the majority of these ontologies is not based on the ELI and ECLI ontologies. For instance, [Ajani et al., 2016] proposed the European Legal Taxonomy Syllabus (ELTS) as a lightweight ontology, that should help to relate national and European legal terminology to represent the differences in the national legal systems of the EU member states. A legal knowledge management system based on ELTS that can be used to semi-automatically classify and interlink documents has been proposed by [Boella et al., 2019]. Ontologies for particular legal areas are for instance the Open Digital Rights Language (ODRL)[17], which is used to model regulatory constraints [De Vos et al., 2019] and to encode the GDPR [Agarwal et al., 2018], Linked Data Rights (LDR)[18] and the Media Contract Ontology (MCO) [Rodríguez-Doncel et al., 2016] used to model policies, and LOTED2 [Distinto et al., 2016] and PPROC [Muñoz-Soro et al., 2016] for the procurement domain. Ontologies related to data protection are,

[15] http://www.akomantoso.org/, last accessed 2021-01-15

[16] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legaldocml, last accessed 2021-01-15

[17] https://www.w3.org/community/odrl/, last accessed 2021-01-15

[18] http://vocab.linkeddata.es/ontologies/purl.oclc.orgNETldrns.html, last accessed 2021-01-15

for instance, GDPRtEXT [Pandit et al., 2018], which is an extension of the ELI ontology to model the GDPR, the SPECIAL ontologies for GDPR compliance checking [Bonatti et al., 2020], PrivOnto [Oltramari et al., 2018], PrOnto [Palmirani et al., 2018] to model privacy policies and a similarly named ontology to represent product information called PRONTO [Vegetti et al., 2011].

In our case, we use the ELI and ECLI ontologies to model the Austrian legal knowledge graph, because Austria wants to contribute towards a common European framework to represent and interlink legal information. Furthermore, the goal of our legal knowledge graph is to model the Austrian legal system as a whole and not at only specific aspects of a particular legal area.

3.6 Summary and Future Directions

In this chapter, we introduced the Austrian legal information system and showed how it is used to search legal information. We then described the limitations of the current legal information system based on the example questions. Furthermore, we introduced the challenges and derived requirements from the current Austrian legal information system, which served as an input for the legal knowledge graph creation.

The legal knowledge graph is based on the ELI and ECLI ontologies, which were proposed by the EU to foster easier access to and cross-border interlinking of legal information. We described the knowledge graph creation process using a middle-out approach, bringing together these ontologies and the available data in the Austrian legal information system. When we compared the proposed ELI and ECLI ontologies, we saw that there are Austrian documents that cannot be cannot be mapped directly/completely to the existing ELI/ECLI model and that is why we extended the ELI and ECLI ontologies with Austrian-specific classes and properties. The extension of the ELI and ECLI ontologies resulted in the Legal Knowledge Graph (LKG) ontology, which adds six classes and more than 30 properties in order to properly model the Austrian legal system in a legal knowledge graph. Moreover, the ontologies provide enough flexibility to enable national extensions. Even more, for some properties the ELI ontology explicitly states that member states should create their own lists. Hence, we introduced the AustroVoc thesaurus, a vocabulary containing Austrian specific terminology.

To summarize, we created a KG model by combining external ontologies following best practices from ontology engineering for the case of the Austrian legal system, with the ultimate goal of interlinking this model with legal data from other countries.

In future work it would be worthwhile investigating how the ELI and ECLI ontologies could be improved. [Francesconi et al., 2015] made a proposal for the improvement of the ELI modeling to reduce the query complexity. Furthermore, it would be interesting to analyze whether other ontologies, specific for a particular legal area, for instance GDPRtEXT [Pandit et al., 2018] to model the GDPR, could be used to extend the legal knowledge graph.

CHAPTER 4

Legal Knowledge Graph Population

In this chapter, we populate the legal knowledge graph, which we modeled in the previous chapter, with actual data from various sources by following different approaches. In Section 4.1, we describe the population from structured data and distinguish between direct and indirect population of the knowledge graph. Moreover, we also describe the population from external sources providing data in a structured format, for instance external knowledge bases like Geonames. In Section 4.2, we focus on the population from unstructured data, hence data extracted from the legal documents. For this purpose, we create a new corpus, which is composed of 50 Austrian Supreme Court decisions from RIS manually annotated with legal entities. This corpus is used to compare the performance of rule-based and deep learning-based information extraction approaches. In Section 4.3, we demonstrate another way of populating the legal knowledge graph by document classification. We use legal documents to evaluate different approaches to classify them into EuroVoc categories by utilizing the hierarchical structure of the EuroVoc thesaurus. We propose a method for the alignment of heterogeneous schemes, in particular to align the concepts of the AustroVoc thesaurus and the EuroVoc thesaurus, in Section 4.4. Finally, the related work is presented in Section 4.5 and Section 4.6 summarizes the chapter including a view on possible future research directions.

4.1 Population from Structured Data

Structured data is organized in a certain way, allowing it to be directly queried and processed [Baars and Kemper, 2008], for instance data contained in relational database. For the creation of our legal knowledge graph, we were provided with a dump of the relational RIS database, which contains the metadata as well as the text of the legal documents we can use for the population from structured data. However, the used database schema does not satisfy the ELI or ECLI metadata requirements upfront. In addition, each RIS application is currently stored in a separate relational database.

4.1.1 Direct Population

A direct mapping (in analogy with the terminology used in R2ML [W3C Recommendation, 2012])[1] of the legal knowledge by mapping attributes to URLs is possible where the required metadata is available. This is typically applicable to properties that have a literal as an object and preprocessing of the data is limited to a minimum. An example would be transforming a date from datetime to date format, for instance for the properties `dcterms:date`, `dcterms:issued`, `eli:first_date_entry_in_force`, `eli:date_no_longer_in_force`, `eli:date_document` and `eli:date_publication` in ISO 8601[2] format (YYYY-MM-DD). Other properties that have a literal as their object, such as `eli:title`, `eli:title_short` and `eli:title_alternative`, are transformed without modification.

4.1.2 Indirect Population

This approach is used when there is data available in a structured format that cannot be directly fed into the legal knowledge graph. For example, in case of resource types represented as simple strings in the database, which need to be mapped to/replaced with the AustroVoc vocabulary terms based on mappings between the input and the output data. Furthermore, in cases where linking requires additional lookups or conditionals. In more detail, RIS document types are indicated as strings or integers in the database, but we created a concept scheme `av:resource-types` as suggested by the ELI and ECLI ontologies in AustroVoc. For instance, a legal provision of type “BG” (Bundesgesetz, federal law) is replaced with the AustroVoc concept `av:leg_bg`, where the resource can be linked to its type using the properties `eli:type_document` for legislative documents and `dcterms:type` for judiciary documents. We proceed similarly when it comes to mapping the law index of legal provisions using the property `eli:is_about`. The law index item is also replaced with the corresponding `av:bundesrechtindex`. We use the legal provisions mentioned in the text to assign judiciary documents a class. We look up the law index for each of the found legal provisions and assign the law index to the judiciary document in order to populate the `dcterms:subject` property for each judiciary document. Furthermore, references extracted from the document text are strings, which need to be replaced with the actual URI of the referenced documents and linked using the `dcterms:references` and `eli:cited_by_case_law` properties.

4.1.3 Population by Interlinking External Sources

Although the RIS database contains relevant legal information, for instance legal provisions and court decisions, it does not provide additional structured background information. What is more, such information could also be interesting in terms of enhancing the legal search process by adding respective search attributes as well as enabling advanced analysis of the legal system. Such background information might include spatio-temporal information about geographic entities or events mentioned in court decisions, for instance the deciding courts or case relevant dates. Similar techniques for enhancing search by interlinking information from spatio-temporal knowledge graphs

[1] However, as opposed to the strict definition in the R2RML standard, note that we speak herein also about direct mapping, when minor, straightforward syntactic literal transformations are applied.

[2] `https://www.iso.org/iso-8601-date-and-time-format.html`, last accessed 2021-02-22

Listing 4.1: Example *Bezirksgericht Leopoldstadt* (extension of Listing 3.4)

```
<https://data.wu.ac.at/legal/lkg/court#court_8>
    rdf:type
        lkg:Court ;
    rdfs:label
        "Bezirksgericht Leopoldstadt" ;
    lkg:court_type
        av:bg ;
    lkg:located_in_community
        <http://sws.geonames.org/2772614/> ;
    lkg:located_in_country
        <http://sws.geonames.org/2782113/> ;
    lkg:located_in_district
        <http://sws.geonames.org/2761333/> ;
    lkg:located_in_state
        <http://sws.geonames.org/2761367/> ;
    rdfs:seeAlso
        <https://www.openstreetmap.org/relation/1651546> .
```

have already proven successful for Open Data search [Neumaier and Polleres, 2019]. As for geo-references, we enhance the court information with external data from Nominatim[3], the search engine of OpenStreetMap (OSM)[4], and Geonames[5] from which we use an RDF dump that we import in our legal knowledge graph. In order to get information about the Austrian courts, we compile a list of court names and query Nominatim for address information, for instance for "Bezirksgericht Leopoldstadt"[6]. The result contains an entry "display_name" and provides address information such as street, community, district, state and country. We extract this information and use Geonames in order to populate the properties `lkg:located_in_community`, `lkg:located_in_district`, `lkg:located_in_state` and `lkg:located_in_country` as shown in Listing 4.1. In addition, we also include the OSM court information page using `rdfs:seeAlso`, which allows users of the legal information system to retrieve location and contact information for the respective authorities.

4.2 Population by Information Extraction

Not all properties of the ELI and ECLI ontologies can be populated from the metadata. However, this missing information can be extracted from the document text using Natural Language Processing (NLP) tools and techniques. The process of extracting entities from a text and classifying them into a set of classes (e.g. person, organization, etc...) is called Named Entity Recognition (NER) [Grishman and Sundheim, 1996]. In our case, we extract legal entities, such as courts, legal provisions and law gazettes. For instance, court decisions contain references to other documents that are not available in the metadata, such as legal provisions and legal rules mentioned in the court decision text. We note though, that rather than structured hyperlinks, the references used in legal practice are oriented on the use by

[3]`https://nominatim.openstreetmap.org/`, last accessed 2021-01-22
[4]`https://www.openstreetmap.org/`, last accessed 2021-01-22
[5]`https://www.geonames.org/`, last accessed 2021-01-22
[6]`https://nominatim.openstreetmap.org/search/Bezirksgericht Leopoldstadt?polygon_geojson=1&format=json&countrycode=AT&type=administrative`, last accessed 2021-01-22

humans and therefore use simple textual labels such as *§ 28a KSchG* rather than URIs like `https://www.ris.bka.gv.at/eli/BGBl/1979/140/P28a/NOR40180997` to reference a legal provision. In order to transform such unstructured references to machine-readable links in our knowledge graph, we extract such textual entities to find corresponding ELI or ECLI identifiers of referenced documents, linking both documents with the properties `dcterms:references` (`lkg:JudicialResource` –> `lkg:LegalProvision`) and vice versa `eli:cited_by_case_law` (`lkg:LegalProvision` –> `lkg:JudicialResource`). Multiple approaches are available to extract information from document text, which could help us to link the documents with each other. We herein specifically compare a *rule-based approach* used in combination with *gazetteers* with more advanced approaches such as *conditional random fields* and *deep learning*. A comparative assessment of these orthogonal approaches helps to increase confidence in the extraction results in the legal domain.

4.2.1 Dataset(s)

For a performance comparison between the different approaches, we need an annotated training corpus of legal documents. To the best of our knowledge, there is no gold standard Austrian legal corpus available, thus we manually annotate 50 randomly selected decision texts from the Justice branch. The documents are quite varied in length with an average of 11,669 tokens with ± 7,741.88 tokens standard deviation (SD), and 260.12 (± 262.71 SD) sentences. For the population of our knowledge graph, we extract the following legal entities and show examples (the relevant legal entities are highlighted in boldface) taken from the Austrian Supreme Court decision *10 Ob 60/17x* [Austrian Supreme Court, 2016]:

Case reference is a reference to another decision text, which is used to refer to decisions taken or arguments brought up in previous cases. In the corpus, a document contains on average 33 (± 23 SD) case references.

> "*Auch Gesprächsnotizen, die vorgedruckte und standardmäßige verwendete Formulierungen enthalten, unterliegen der verbraucherschutzrechtlichen Geltungs- und Inhaltskontrolle (**1 Ob 46/10m**); dies trifft auch auf in Websites und deren Subpages enthaltene vorformulierte Allgemeine Vertragsbedingungen zu (**2 Ob 59/12h**).*" [Austrian Supreme Court, 2016]

Contributor contains the names of the judges involved in a decision. The number of judges involved in a decision amounts to 5 (± 2 SD), which is caused by the different compositions of the senates.

> "*Der Oberste Gerichtshof hat als Revisionsgericht durch den Senatspräsidenten **Dr. Neumayr** als Vorsitzenden, die Hofrätinnen **Dr. Fichtenau** und **Dr. Grohmann** sowie [...]*" [Austrian Supreme Court, 2016]

Court is mentioned in the decision text to indicate the court taking the decision, but there are also courts in the appeal stages. Courts are mentioned 15 (± 6 SD) times in a document.

> "*Der **Oberste Gerichtshof** hat als Revisionsgericht durch den Senatspräsidenten [...] gegen das Urteil des **Oberlandesgerichts Wien** als Berufungsgericht [...]*" [Austrian Supreme Court, 2016]

Listing 4.2: Example snippet JAPE rule for the extraction of legal rules

```
Input: Token
Rule: rs
(
  {Token.string == "RS"}
  {Token.kind == "number"}
):rs
-->
:rs.LegalRule = {legalrule = :rs@string}
```

Legal rule is a summarizing statement of a ruling from which general rules are inferred and are often cited in decision texts to back up the decision. Legal rules are cited 23 (± 22 SD) times on average in the documents of the corpus.

> "*[...] zugunsten des obsiegenden Klägers (RIS-Justiz **RS0079624** [T14]). Ein berechtigtes Interesse des obsiegenden Beklagten an der Urteilsveröffentlichung ist dann gegeben, wenn der Rechtsstreit eine gewisse Publizität erlangt hat (RIS-Justiz **RS0079511**), etwa wenn [...]*" [Austrian Supreme Court, 2016]

Legal provision is mentioned in the decision text and forms the legal basis on which the decision is grounded. Court decisions must be based on the law, it is therefore not surprising that 87 (± 72 SD) legal provisions are cited on average.

> "*[...] einen Verstoß gegen **§ 6 Abs 3 KSchG** und **§ 879 Abs 3 ABGB** geltend. Die Klausel lasse eine beträchtliche Entgelterhöhung im Wege einer Zustimmungsfiktion iSd **§ 6 Abs 1 Z 2 KSchG** aufgrund [...]*" [Austrian Supreme Court, 2016]

Law gazette is cited in cases where the court wants to refer to a specific version of the law. A law gazette is usually cited together with a legal provision to indicate the specific version of the legal provision the court is referring to. Given the purpose of citing a law gazette in a court decision, the number of citations is comparatively low with an average of 4 (± 6 SD) per court decision.

> "*[...] in Kraft getretene Zahlungsdienstegesetz (ZaDiG, **BGBl I 2009/66**), von Relevanz [...]*" [Austrian Supreme Court, 2016]

Literature is used to cite legal literature used to back up the decision. We also extract these references as they are relatively high with 50 (± 36 SD) citations on average and thus constitute a very important source. However, the literature is mostly (at least in Austria) only available against a paid subscription from various legal publishers. Furthermore, the citation style depends on the type of the cited literature (e.g. commentary, book,...) and is sometimes abbreviated when the citation is repeated in the same court decision.

> "*[...] vgl. **Mayrhofer/Tangl in Fenyves/Kerschner/Vonkilch, Klang3 § 6 Abs 1 Z 2 KSchG** Rz 1 [...]*" [Austrian Supreme Court, 2016]

4.2.2 Approach

Rule-based approach. Given that legal documents follow a relatively regular structure and citation style, we apply a rule-based approach for the information extraction and use the Java Annotation Pattern Engine (JAPE) [Cunningham et al., 1999], which is part of the General Architecture for Text Engineering (GATE)[7]. An example of how we can exploit the standardized citation style in legal documents is shown in Listing 4.2, which illustrates a (shortened) JAPE rule used to extract references to legal rules in a court decision. A JAPE rule has a left hand side where the rule is defined and a right hand side that defines what to do with the extracted information. Both sides are separated with a "-->". After a tokenizer (splitting the text into its individual parts) has been applied, the JAPE rule takes a *Token* as an input and looks for the defined pattern in the rule section. In this example, a legal rule must start with a token consisting of a string "RS" directly followed by a token of kind "number". The returned result is the complete legal rule string, for instance *RS0042781*. We can look up the legal rule string in the database in order to replace the literal text with its actual URI, thus generating a link between the two documents. Rules can easily be supported by gazetteers, which are lookup lists that are very suitable for static, recurring entities, hence entities that do not change frequently. We use gazetteers to assist with the detection of contributors (a list with most common names and academic degrees), courts, legal provision (a list with all law abbreviations) and literature (a list with the most common legal journals used in Austria). Note that we included a score for a strict and a lenient evaluation for the rule-based approach. The strict evaluation of rules only counts occurrences as correct when the annotation of the rule matches the gold standard annotation exactly. Lenient results also count occurrences as correct when both annotations overlap with the rule (adding or omitting some words).

Conditional Random Fields. An alternative, common approach to label textual sequence data using probabilistic models are Conditional Random Fields (CRF) [Lafferty et al., 2001]. We use the implementation of the sklearn-crfsuite[8]. The features of a token, for instance position and casing, are used to calculate the probabilities of tokens following each other. In the legal domain, CRF have already been used in the context of entity extraction tasks where it has shown good results (e.g. [Dozier et al., 2010, Cardellino et al., 2017a, Leitner et al., 2019]).

Deep learning approach. For experiments involving embeddings and deep learning we use the Flair framework[9], which provides all the necessary functionality required for our evaluation and in addition also supports importing *pretrained German language models*, which we were hoping to boost the accuracy for our German legal document corpus. We compare the following language models: (i) *Flair*, which uses contextualized character level embeddings [Akbik et al., 2018] trained on a mixed corpus of web and Wikipedia documents; (ii) Language models using a transformer based architecture [Vaswani et al., 2017] provided by HuggingFace[10] [Wolf et al., 2019] known as *Bidirectional Encoder Representations from Transformers (BERT)* [Devlin et al., 2019] trained on German Wikipedia, German open legal data and news articles; and (iii) *DistilBERT* [Sanh et al., 2019] a faster and smaller version of BERT also trained on Wikipedia articles and web documents. DistilBERT uses a teacher-student setting to distill the knowledge from the teacher (the BERT model) to the student (DistilBERT model).

[7] `https://gate.ac.uk/`, last accessed 2021-01-31
[8] `https://sklearn-crfsuite.readthedocs.io/en/latest/`, last accessed 2021-01-31
[9] `https://github.com/flairNLP/flair`, last accessed 2021-01-31
[10] `https://huggingface.co/`, last accessed 2021-01-31

Table 4.1: Evaluation results of legal entity extraction

P=Precision, R=Recall, F=F-score. Best results highlighted in boldface.

			Case reference	Contributor	Court	Legal provision	Law gazette	Legal rule	Literature
Rules	Rules strict	P	0.9782	0.7631	0.9892	0.8742	0.9150	1	0.6814
		R	0.9817	0.9406	0.9659	0.9074	0.9683	1	0.7865
		F	0.9799	0.8426	0.9774	0.8905	0.9409	**1**	0.7302
	Rules lenient	P	0.9806	0.7631	0.9919	0.8923	0.9200	1	0.8095
		R	0.9842	0.9406	0.9685	0.9262	0.9735	1	0.9343
		F	**0.9824**	0.8426	**0.9801**	0.9090	0.9460	**1**	0.8674
CRF	CRF	P	0.9868	0.9161	0.9852	0.9452	0.9638	0.9994	0.9145
		R	0.9710	0.9557	0.9416	0.9483	0.9364	1	0.8611
		F	0.9787	0.9328	0.9616	0.9459	**0.9473**	0.9997	**0.8866**
Deep Learning	Flair	P	0.9783	0.9187	0.9455	0.9324	0.9263	1	0.8596
		R	0.9800	0.9780	0.9486	0.9526	0.9245	1	0.8671
		F	0.9791	0.9435	0.9456	0.9414	0.9215	**1**	0.8629
	BERT	P	0.9687	0.9481	0.9557	0.9447	0.9546	0.9971	0.8497
		R	0.9738	0.9710	0.9762	0.9536	0.9336	1	0.8409
		F	0.9712	**0.9583**	0.9654	0.9489	0.9396	0.9986	0.8448
	DistilBert	P	0.9759	0.9316	0.9407	0.9446	0.9392	0.9979	0.8663
		R	0.9786	0.9878	0.9784	0.9600	0.9529	1	0.8604
		F	0.9772	0.9551	0.9586	**0.9521**	0.9437	0.9989	0.8626

4.2.3 Evaluation and Discussion

For our experiments, we did not apply any preprocessing to the documents and apply a 5-fold cross-validation approach using a train/test/validation split of 80%/10%/10%. All models have been trained with default settings, in particular the deep learning models with a maximum of 150 epochs, starting learning rate of 0.1, patience 3 and an anneal factor of 0.5. The training stops automatically when the learning rate becomes too small.

Table 4.1 shows the results for the different legal entities, whereby approaches with the best F-scores are highlighted in boldface. Looking at the evaluation results, we can see at first glance that there is no single clear best approach outperforming all other approaches on all legal entities. Furthermore, it can also be noted that the results of all extraction methods are comparable across all methods for the individual legal entities. In particular, the numbers show that rules perform well, when the entities under investigation are highly structured and always follow the same pattern, for instance case reference (e.g. “14Os108/20v)” and legal rule (e.g. “RS0042781”), which are very easy to recognize. Moreover, we use gazetteers to support rules for the extraction of the contributors. The rule looks for a degree (from a gazetteer) followed by a last name (from a gazetteer) within the head of the document. When adding more variations and more complexity to the legal entities, the performance of the rule-based and gazetteer supported approach deteriorates and machine learning-based approaches perform better. The numbers of the legal provision, law gazette and literature show this effect. The citations of legal provisions can be simpler (e.g. “§ 41 ZPO”) and more complex (e.g. “§§ 41, 43 Abs

2 erster Fall und § 50 ZPO"), which adds a lot of complexity to the rules and, as a result, makes the result much harder to create. The citations of the law gazettes changed over time by adding additional information (e.g. from "BGBl. 1969/207" to "BGBl. I Nr. 134/2015"). The most complex entity to extract is the literature, because there are various types of literature (e.g. commentaries, books, articles,..) and citation styles used. References to literature might also include names, which might be mistaken for contributors or legal provisions in the reference. The higher complexity of literature references is also reflected in the evaluation results. While the best F-scores for the other legal entities are somewhere in the 94% range, the best F-score for literature is achieved by CRF with only 88%. The numbers also show that the gap between the rules and automatic approaches is bigger the more complex the rules (with gazetteer support) need to be. However, the gap between the individual approaches is very small. The F-scores of the three deep learning approaches (Flair, BERT, DistilBERT) are within 2% across all legal entities, thus we cannot nominate a clear winner in this segment. Also the difference across all approaches and legal entities falls within a range of 4%.

Although the evaluation results show that the extraction approaches perform mostly equally well, we should also take the effort into account, that is required to set up such a system for the extraction of legal entities. Rules can be easily and quickly created with only a few sample documents that cover the possible variations in which legal entities can appear. In addition, rules are easy to interpret and explain. The outcome of a rule is clear from the beginning, as a rule either matches a sequence of tokens or not. Gazetteers are suitable for entities that do not change frequently, for instance courts or names. Moreover, gazetteers have a maintenance requirement and might need to be updated on a regular basis, otherwise rules using these gazetteers will start to fail over time. By contrast, approaches using (deep) machine learning promise to be more flexible and are also able to cover variations in patterns where a rule would fail. However, these approaches are less explainable and predictable, hence working with probabilities of the results and selecting the right algorithm for the right task is necessary.

In addition, we remark that it requires considerable effort to annotate documents required for training machine and deep learning approaches as well as computational power and resources to perform both model fine-tuning and training. In our case, the experiments with our corpus of only 50 documents used the full capacity of our machine with 16GB of memory and requires a powerful GPU (we use a GTX 1080 Ti with 16GB memory) to perform the computations in a timely manner.

Summarizing the results shown by the experiments, there is no clear best approach to extract legal entities from text. Thus the approach should be chosen based on the requirements, the available data from the legal information system acting as a data source and human resources. We conclude that rules, in combination with gazetteers, are a viable alternative and can keep up with state-of-the-art NLP techniques using complex neural networks for the relatively well-structured texts in our domain, offering maintainability and explainability of extraction results.

4.3 Population by Classification

The process of assigning a category out of a set of categories to a document is called *text classification* [Sebastiani, 2002]. We distinguish between different types of

Table 4.2: Multi-label datasets

Dataset	Domain	# Doc	# Labels	Avg. # tokens	Std. Dev.	Skewness	Kurtosis
JRC-Acquis V3	Legal	17,519	3,563	3.065,90	8,931.94	8.61	112.82
EUR-Lex 4K	Legal	19,513	3,969	3,021.38	8,606.06	7.74	88.98
Reuters-21578	News	21,578	120	151,05	152.16	7.05	54.37

classification, namely binary classification, multi-class classification and multi-label classification [Tsoumakas and Katakis, 2007]. Binary classification is the simplest of these three classification types as it assigns two mutually exclusive categories to a text. Binary classification is often used to classify, for instance, reviews (positive, negative) or a text satisfying conditions (true, false). One class out of a set of disjoint classes is assigned to a document in multi-class classification tasks, while multi-label classification tasks involve the assignment of one or more partially overlapping classes out of a set of classes to a document, which is considered to be the hardest task of all document classification tasks. These tasks are typically approached with machine learning and more recently also with deep learning systems. The selection of applicable algorithms is task dependent and ranges from decision trees, probabilistic and rule-based classifiers [Hotho et al., 2005, Aggarwal and Zhai, 2012, Allahyari et al., 2017] to recurrent and convolutional neural networks [Howard and Ruder, 2018, Jacovi et al., 2018].

The ELI (`eli:is_about`) and ECLI (`dcterms:subject`) ontologies provide properties to link legal documents with classes in a classification schema, for instance the EuroVoc or a national thesaurus such as AustroVoc. Furthermore, legal documents contain a lot of semantic information, which can be used for the classification task [Altinel and Ganiz, 2018]. The main idea of using the semantics of the documents in the classification task is that documents talking about same topics do also have a similar semantics. The challenging part of classifying legal documents is that they are written in a very domain-specific language, including the usage of many abbreviations and the large number of different classes (more than 6,000) available in the EuroVoc thesaurus. That is why it could help to exploit the hierarchical structure of the EuroVoc thesaurus resulting in a reduced number of potential classes and to increase the classification results.

In order to analyze the possibilities of using classification algorithms for the population of a legal knowledge graph, we use two corpora with legal documents from the European Union, as there is no gold standard dataset for Austrian legal documents available. These documents are assigned to multiple classes, hence a multi-label classification problem, which is evaluated on statistical, machine learning and deep learning-based approaches as well as combinations thereof.

4.3.1 Dataset(s)

For our experiments, we use two legal corpora and a corpus of news articles for comparison. The structural features of the datasets are presented in Table 4.2. The metrics of two legal datasets (JRC-Acquis V3 and EUR-Lex 4K) compared to the popular Reuters-21578 dataset from the news domain, shows that the Reuters-21578 dataset is comparable on the number of documents in the corpus. However, it includes only 120 classes to classify the documents, which is less than 5% of the possible EuroVoc

Table 4.3: Overview of dataset features

Dataset	Dataset Version	# Doc	# Labels	Label Cardinality	Avg. # Doc / Label
JRC-Acquis V3	full	17,519	3,563	5.41	26.62
JRC-Acquis V3	topterms	17,519	489	4.59	164.21
JRC-Acquis V3	microthesauri	17,519	126	4.60	634.88
EUR-Lex 4K	full	19,513	3,969	5.39	26.15
EUR-Lex 4K	topterms	19,513	512	4.65	177.02
EUR-Lex 4K	microthesauri	19,513	126	4.82	741.59
Reuters-21578	-	21,578	120	1.26	202.57

labels in legal datasets. In addition, the length of news documents is much shorter than the documents from the legal domain. The skewness describes the symmetry of the label distribution. A skewness value in the range -0.5 to 0.5 describes a symmetrical distribution and a high positive or negative skewness value indicates highly asymmetrical, hence highly skewed data. Comparing the skewness values for all three datasets, we can clearly see that label usage in all three datasets is highly skewed. The kurtosis of a dataset refers to the outliers in the distribution, with a value of 0 showing that the distribution follows the standard distribution. All three datasets have a positive kurtosis indicating larger tails, and a power-law distribution of labels usage.

The *EU Acquis Communautaire* is the collection of the legal documents and obligations within the European Union containing regulations, directives, decisions, treaties and many more. Version 3 of the JRC-Acquis corpus contains documents in various languages from institutions of the European Union dating from 1958 to 2006. There are around 20,000 documents available per language. The English version, which we use, contains 20,682 documents in XML format. The documents have been manually classified into the different EuroVoc classes and include the identifiers of the respective EuroVoc classes [Steinberger et al., 2006]. The JRC-Acquis corpus, which is the property of the European Commission, is available free of charge for commercial and non-commercial use under the provisions laid out in the Commission Decision of the 12$^{\text{th}}$ of December 2011[11]. Our second dataset, the EUR-Lex 4K dataset [Loza Mencía and Fürnkranz, 2010] also consists of documents taken from the EUR-Lex database and is provided by the Technical University of Darmstadt. The most important dataset properties of the test datasets we created from these corpora are summarized in Table 4.3. We created two additional dataset versions from the original (*full*) datasets, which contain the document and class assignments as they are provided. The *topterms* and *microthesauri* versions are based on the original EuroVoc class assignments, but exploit the hierarchy to reduce the number of different classes. Note that, although as mentioned above, there are 20,682 documents in the original JRC dataset, only 17,519 documents are actually annotated with EuroVoc classes. We pruned non-annotated documents from the dataset and kept only those documents, which actually have EuroVoc classification labels. Furthermore, note that despite the fact that there are more than 6,000 EuroVoc classes available, only 3,563 are actually used by the documents in the full JRC-Acquis dataset. For the creation of the topterms version of the dataset, we extracted all top terms from the EuroVoc thesaurus and replaced all EuroVoc leaf classes in the full JRC dataset with the top

[11] `http://data.europa.eu/eli/dec/2011/833/oj`, last accessed 2021-01-31

term classes (489) they belong to. Similarly, the microthesauri version of the dataset is generated by replacing annotations with the unique microthesaurus they belong to (126). The same approach reduces the number of classes of the EUR-Lex 4k dataset from 3,969 classes to 512 classes for the topterms and again 126 classes for the microthesauri version of the dataset.

The class reduction is based on the hierarchy of the classes in the EuroVoc thesaurus and works as follows: For each EuroVoc class for a given document the top term (microthesaurus) is looked up and replaced with the found top term (microthesaurus). Since multiple EuroVoc classes for a document can belong to the same microthesaurus, we only take each result once, hence a set. For instance, class 575 is a narrower term of class 573, hence we replace 575 with 573. This way, we reduce the overall number of classes available for classification by 86% to 489 labels in total. For the microthesauri version of the dataset, we apply the same procedure and are therefore able to reduce the number of classes to 126. Notice that the EuroVoc thesaurus has 127 microthesauri of which we use only 126. The 127$^{\text{th}}$ microthesaurus is a general microthesaurus to which every EuroVoc class belongs. Hence, this missing microthesaurus does not contribute to the classification problem and has therefore been removed.

The label cardinality describes the average number of EuroVoc classes assigned to each document. Documents from the original dataset have 5.41 class labels on average. The decrease of the label cardinality for the topterms (4.59) and microthesauri (4.6 classes per document) versions of the dataset is caused by going up the hierarchy in the EuroVoc thesaurus and reducing the number of classes. Moreover, some documents are annotated with multiple EuroVoc classes sharing the same top term or microthesaurus. The decrease in available EuroVoc classes also affects the number of documents per class. While, in the original full dataset there are on average only 27 documents available per class, we have 164 documents per class in the topterms and 635 documents per class in the microthesauri version of the dataset. Since the number of documents remains the same for all three versions of the dataset, the average number of tokens per document of 3,066 as well as the standard deviation of $\pm$ 8,932 also remain the same.

We use version 1.0 of the Reuters-21578 dataset, available for research purposes, to compare the approaches. The Reuters-21578 dataset contains documents that appeared in the Reuters Newswire, which have been manually annotated with 120 classes. The label cardinality is also much lower compared to the two legal datasets, but the learning process can make use of around 200 documents per label.

4.3.2 Approach

Preprocessing. The first step is to do the preprocessing of the raw text files, not only to reduce the size of the documents, but also to reduce the runtime of all subsequent processing steps. We opted to separate this preprocessing step from the actual classification process and runtime measurement. Preprocessing includes lowercasing as well as removing stopwords from the text using the standard English NLTK[12] stopword list. We also remove punctuation and special characters from the text and replace all words with their lemma using the spaCy[13] lemmatizer to reduce the morphological variations of each word to their lemma. In addition to these standard preprocessing steps, we also

[12] `https://www.nltk.org/`, last accessed 2021-01-31
[13] `https://spacy.io/`, last accessed 2021-01-31

include specific preprocessing steps tailored to the legal documents, which include the removal of references to other legal documents (e.g. "[..] amended by Directive 83 / 191 / EEC [...]") and the removal of all brackets and their contents for the same reason. Also the structure of legal documents can be used in preprocessing in order to remove all headings contained in the documents (e.g. "Article" or "Appendix").

Term Frequency - Inversed Document Frequency. The most basic approach used for classification is based on counting the numbers of term occurrences in documents. *Term Frequency (TF)* indicates the number of occurrences of each term in a document. Under the assumption that more important terms occur more often, we could say that the higher the frequency, the higher the importance (relevancy) of a term. However, there might be terms that occur many times, but are not unique to a particular document in a corpus. For instance, the term "regulation" might occur very often in legal documents from the European Union but rarely in tweets. In order to account for the descriptiveness of a term in relation to the entire corpus, term frequency is typically contrasted by *Inverse Document Frequency (IDF)* [Jones, 2004] to measure the descriptive power of a term in a corpus. The assumption is that a term is less descriptive and specific if it appears in a high number of documents. Terms that appear in only a fraction of documents are useful to distinguish those documents from others, and consequently are useful for classification. Finally, the TF-IDF score is the product of the TF and IDF scores. For our corpus, this means that many of the generic domain-specific terms, such as "regulation", "directive", "commission", "EC", "EEC", etc., are considered to have low discriminative power and the remaining terms are weighted higher.

Word2Vec. In order to apply neural language modeling to large-scale text corpora in a run time-efficient manner, in recent years new methods based on simplified neural network architectures have been proposed. The first, and most well-known approach, in this area is Word2Vec [Mikolov et al., 2013a]. Word2Vec trains a model on text in an unsupervised way, and as a result generates low-dimensional, dense, floating-point vector representations for each word in the corpus. There is the possibility to download pretrained models, which are trained on different corpora (e.g. from github[14]), or to train one's own corpus-specific model. Furthermore, Word2Vec includes two different algorithms for model training, the *Continuous Bag of Words (CBOW)* model and the *skip-gram* model. The former is primarily used to predict a word from a given context, while the latter aims at predicting the context given a word.

First, we tested large-scale pretrained language models trained with general-purpose text corpora such as GoogleNews and the CommonCrawl, but as expected both performed badly on the legal dataset. For example, the CommonCrawl model reached an F-score of 0.38 and the GoogleNews model an F-score of 0.31. Therefore, we opted to train our own model based on the JRC-Acquis corpus. Despite the fact that for using Word2Vec the corpus size typically has a large impact on model quality, we achieve better results by training a model on our 17,519 documents than reusing the large pretrained models: at the very least, this seems to confirm our base assumption that generic language models do not work well on the domain-specific language used in legal documents. As for the training parameters, we use the standard settings with a vector size of 300 and a minimum count of 1 due to the homogeneous corpus and in order to capture very specific

[14] `https://github.com/3Top/word2vec-api`, last accessed 2021-01-31

words in legal documents. We use the CBOW model for the classification task because it outperforms skip-gram by more than 15% in terms of the F-score (0.4 for skip-gram vs 0.55 for CBOW). We employ a simple method to create the document vectors by summing up the vectors of all words contained in a document and computing an average vector. Our assumption is that these average vectors of documents specific to a given document topic (represented by their EuroVoc classifications) are similar.

Doc2Vec. While Word2Vec creates global word representations, Doc2Vec creates a vector for an entire document. Doc2Vec uses word vectors and extends the vectors by adding paragraph vectors, which allows for the predictions of words in the context of a paragraph [Le and Mikolov, 2014]. Similar to Word2Vec, Doc2Vec also allows users to train two different kinds of models: *Distributed Bag of Words (DBOW)* and *Distributed Memory (DM)*. For our training, we use a vector size of 300 and minimum count of 1.

TF-IDF weighting embeddings. In order to filter the domain corpora, and to exclude generic legal terms without discriminative power in the legal domain, we use the weighting approach as suggested in [Lilleberg et al., 2015] to remove common words from the embeddings. We achieve this by combining the statistical TF-IDF approach mentioned above, with the word embeddings of Word2Vec and Doc2Vec. In the first step, we calculate the TF-IDF scores for all words in the corpus. Since the number of words varies from document to document and the TF-IDF scores are also different, we do not set a hard limit for the TF-IDF scores, instead we calculate the TF-IDF scores for all words in a document and rank them according to these scores. Afterwards, we set a threshold for the TF-IDF scores and remove all words with a score below the set threshold. The threshold is set as the top x percent of words, in particular experiments showed that the top 10% of the words are most descriptive and a setting of e.g. 25% of the top words decreases the results. We also cannot set the number of words to be considered to a fixed value (e.g. 10 words per documents) as we do not know the TF-IDF score distribution. The training parameters for Word2Vec and Doc2Vec are the same as in the individual approaches.

For all approaches mentioned above, we use Random Forest (RF) and a Support Vector Machine (SVM). We apply GridSearch to find the best training parameters for both algorithms. We mainly use the standard parameters, but set the *class_weight = balanced* to compensate for the skewed label distribution and *C = 100* for the SVM. All machine learning tasks are performed using Python 3 and the Scikit-learn library[15].

fast.ai. As a representative of currently popular (deep) neural network training approaches, we also compare the above-mentioned approaches to the powerful fast.ai[16] framework: fast.ai is a library for training fast and accurate neural nets. It is based on deep learning research and tries to incorporate current best practices. The fast.ai framework provides support for different task types, such as computer vision, NLP, tabular data and recommender systems. As for input corpora, we experiment both with the preprocessed dataset and the original JRC dataset. In both cases, fast.ai also applies its own preprocessing on top, which includes lowercasing, marking the start and end of sentences, etc. Additionally, fast.ai applies an iterative model training process, which

[15]`https://scikit-learn.org/stable/`, last accessed 2021-01-31
[16]`https://github.com/fastai/fastai`, last accessed 2021-01-31

Table 4.4: Evaluation results for JRC corpus

			Full			Topterms			Microthesauri	
Approach	**Algorithm**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Baseline	-	0.44	0.52	0.47	-	-	-	-	-	-
TF-IDF	RF	*0.88*	0.24	0.37	*0.90*	0.30	0.45	*0.89*	0.39	0.55
Word2Vec	SVC	0.52	*0.59*	0.55	0.43	*0.80*	0.56	0.50	*0.85*	0.63
Doc2Vec	SVC	0.74	0.40	0.52	0.65	0.61	0.63	0.71	0.69	0.70
TF-IDF + Word2Vec	SVC	0.62	0.47	0.53	0.54	0.69	0.61	0.59	0.77	0.67
TF-IDF + Doc2Vec	SVC	0.62	0.45	0.52	0.46	0.71	0.56	0.53	0.78	0.63
fast.ai	LSTM	0.61	0.55	**0.58**	0.70	0.63	**0.67**	0.75	0.73	**0.74**

Table 4.5: Evaluation results for KED corpus

			Full			Topterms			Microthesauri	
Approach	**Algorithm**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Baseline	-	0.40	0.46	0.42	-	-	-	-	-	-
TF-IDF	RF	*0.84*	0.12	0.21	*0.86*	0.20	0.33	*0.88*	0.34	0.49
Word2Vec	SVC	0.29	*0.63*	0.40	0.34	*0.77*	0.47	0.44	*0.83*	0.57
Doc2Vec	SVC	0.53	0.41	0.46	0.60	0.52	0.56	0.69	0.63	0.66
TF-IDF + Word2Vec	SVC	0.15	0.26	0.19	0.20	0.38	0.26	0.29	0.50	0.36
TF-IDF + Doc2Vec	SVC	0.16	0.25	0.19	0.22	0.36	0.27	0.31	0.47	0.38
fast.ai	LSTM	0.54	0.49	**0.52**	0.64	0.59	**0.61**	0.73	0.69	**0.71**

includes two basic steps: (i) fine-tuning a pretrained language model with the domain corpus, and (ii) learning the classifier. The process as well as additional techniques such as slanted triangular learning rates are explained in [Howard and Ruder, 2018]. In training the models, we follow mostly the recommended architecture given in the fast.ai examples[17], which in the first basic step includes the fine-tuning of the provided `AWD_LSTM` RNN language model with the JRC corpus. When training the multi-label classifier, techniques such as gradual unfreezing of the network, weight decay (set to 0.1) and *momentum* are used. Further, we apply the default loss function for multi-label text classification, BCEWithLogitsLoss.

4.3.3 Evaluation and Discussion

In this section, we present the experiment results. The experiments using embeddings were carried out on a 2.1 GHz machine with 24 cores and a memory of 246 GB. In order to run the fast.ai experiments we used a i7-8700 CPU with 3.76 GHz, 16 GB of memory and a GeForce GTX 1080 Ti graphics card. The code for the embedding experiments is available on Google[18] and the Jupyter notebooks with all fast.ai related experiments on Github[19].

[17]`https://nbviewer.jupyter.org/github/fastai/course-v3/blob/master/nbs/dl1/lesson3-imdb.ipynb`, last accessed 2021-01-31
[18]`https://drive.google.com/open?id=1Pl4H1pFNuFvcGQwHjkhcUJ9SMHrYjdQl`, last accessed 2021-01-31
[19]`https://github.com/gwohlgen/JRC_fastai`, last accessed 2021-01-31

Table 4.6: Evaluation results for Reuters-21578 corpus

			Full	
Approach	**Algorithm**	**P**	**R**	**F**
TF-IDF	RF	*0.97*	0.63	0.76
Word2Vec	SVC	0.50	*0.94*	0.66
Doc2Vec	SVC	0.82	0.84	0.83
TF-IDF + Word2Vec	SVC	0.05	0.38	0.09
TF-IDF + Doc2Vec	SVC	0.14	0.27	0.18
fast.ai (no preprocessing)	LSTM	0.90	0.87	0.88
fast.ai (with preprocessing)	LSTM	0.92	0.88	**0.90**

We evaluate our approaches on the three multi-label datasets JRC-Acquis, EUR-Lex 4K and Reuters-21578. The results for each dataset are presented in a separate table, Table 4.4 for the JRC-Acquis dataset, Table 4.5 for the results of the EUR-Lex 4K dataset and finally Table 4.6 contains the results of the Reuters-21578 dataset. Each result table contains a column indicating the chosen approach for the classification task. Furthermore, for each dataset version (full, topterms and microthesauri) we present the evaluation metrics *Precision*, *Recall* and *F-score*. A - means that there is no result available. The best result for each dataset version is highlighted in boldface, while the best precision and the best recall for each dataset version are highlighted in italic. All results have been achieved using the preprocessed documents and a test set size of 20%.

We used the JRC EuroVoc Indexer JEX tool to calculate the baseline results. The JEX tool can be downloaded with a pretrained English model to calculate the metrics for the JRC-Acquis and EUR-Lex 4K full legal datasets, which is uses a profile-based ranking algorithm for text classification [Steinberger et al., 2012].

Although we tested Random Forest (RF) and Support Vector Machine (SVM) as learning algorithms, the results show that RF performs better only on TF-IDF, while for all other machine learning approaches SVM is the superior learning algorithm. Furthermore, the results clearly show that RF has the highest precision but also the lowest recall on all three versions of the dataset. The increase of the F-score with the TF-IDF approach also shows that a decrease of candidate classes by 87% leads to an increase of 10% of the F-score.

Looking at the result metrics, we can say that using TF-IDF in combination with a Random Forest leads to a very high precision, independent of the number of candidate classes. In contrast, the recall is very low and only shows marginal improvement in the case of reduced classes.

The Word2Vec and Doc2Vec approaches and the combinations of both with TF-IDF show the best results using a SVM. However, there is no clear answer to which approach performs best. Having a look at the results for the full dataset, the F-score ranges from 0.52 to 0.57 and therefore perform better than the baseline with the exception of TF-IDF with an F-score of only 0.37. Also the values for precision and recall are evenly distributed. Furthermore, the relation of precision and recall changes with the decreasing number of candidate classes. While, the Word2Vec and Doc2Vec precision remains almost steady across all dataset versions ($\pm$0.09), the Word2Vec recall increases strongly from 0.59 to 0.85 (+0.26) for Word2Vec and from 0.40 to 0.69 (+0.29) for

Doc2Vec on the JRC dataset. The increase of precision and recall on the EUR-Lex 4K is a little bit lower compared to the JRC dataset, but still shows a good increase over the different dataset versions.

The TF-IDF weighting approaches do not show an increase on the overall performance compared to the individual Word2Vec/ Doc2Vec approach for both legal datasets. Only on the JRC dataset the TF-IDF + Word2Vec approach performs better than Word2Vec only, but solely on the dataset versions with the reduced number of classes. The performance of the combined TF-IDF + Doc2Vec is always lower compared to Doc2Vec. The metrics of the TF-IDF weighting approaches applied to the EUR-Lex 4K dataset are much lower compared to the individual approaches

Our approach using a neural network, with language model transfer learning and the deep LSTM architecture of fast.ai, delivers the best F-scores on all three versions of the dataset although it never has the best precision or recall values. The precision and recall change depending on the threshold value for label selection, thus we used a threshold which provides a good F1 result. The results also demonstrate that the multi-label document classification with such a high number of classes and a strongly biased class distribution is very complex and very hard to handle even for deep neural networks, which have proven to be very successful in recent year on a variety of NLP tasks. On the full dataset fast.ai performs only 3% better than the non-neural network approach using Word2Vec. The advantage of fast.ai on topterms and microthesauri datasets on the JRC dataset is 4% in both cases. The metrics for the EUR-Lex 4K dataset are lower in general, but fast.ai performs better by 6% on the full and 5% on the topterms and microthesauri dataset versions.

The Reuters-21578 results show the impact of the low number of classes in combination with the lower label cardinality. The best approach using embeddings is Doc2Vec with an F-score of 0.83, while the highest precision is achieved by TF-IDF (0.97) and the highest recall by Word2Vec (0.94). Also fast.ai outperforms all other approaches with an F-score of 0.9.

Overall, the prediction performance significantly increases with the reduction of the number of candidate classes by taking advantage of the hierarchy of terms, and that a neural network outperforms classic approaches. However, the differences in the results are small and therefore a final answer as to which approach performs best cannot be given. Particularly, predicting rare labels instead of resorting to the coarser, upper level prediction, is, as expected hardly possible, due to the lack of training data for rare labels. We hope, in the future to address this issue by investigating new methods to combine coarse-label and fine-label predictions and exploit other semantic connections to also enable predictions of these rare terms.

4.4 Alignment of heterogeneous schemes

Last but not least, our AustroVoc thesaurus contains a law index, which is very suited to be linked with related terms in EuroVoc, thereby, directly enabling a multi-lingual search (given that EuroVoc is available in multiple languages). As the main obstacle herein, legal language is diverse even within German speaking countries, plus EuroVoc contains "German" German whereas Austria often uses specific "Austrian" German terms. For example, in Austria we use the term *Konsumentenschutz* but EuroVoc contains the term *Verbraucherschutz* for “customer protection”.

4.4.1 Approach

We want to link the concepts of the Austrian law index (contained in AustroVoc) with EuroVoc concepts. For this purpose, we apply two approaches: One approach is based on a direct lookup and the second approach uses external knowledge bases to find a match. As some of the law index items are compound of multiple terms, we split them into separate terms but map the compound law index items to all found EuroVoc classes for each split term. In case a match is found, we can link the AustroVoc term with the corresponding EuroVoc term using the property `rdfs:seeAlso`. For instance, we find a match from “Konsumentenschutz” to “Verbraucherschutz” and add the triple `av:bri2006 rdfs:seeAlso ev:2836` to AustroVoc as shown in Listing 3.5. In cases were no match is found, we exploit the hierarchy of the law index and map the term for which no match is found with the EuroVoc class of the next higher law index item with a found EuroVoc class.

Direct lookup. The simplest way to find a match of a law index term in the EuroVoc thesaurus is to perform a direct lookup of a law index item in EuroVoc. The EuroVoc thesaurus contains for each concept a preferred term (`skos:prefLabel`) and – if available – also non-preferred terms (`skos:altLabel`). An example for a match found with this approach is the term “Strafrecht”, which is contained in the Austrian law index and in the EuroVoc thesaurus (`ev:573`) where it is also called “Strafrecht”. For the direct lookup of a term we accept both, preferred and non-preferred terms as matches for a law index term because they belong to the same EuroVoc concept.

External knowledge bases. The basic idea of including external knowledge bases is to use them as a dictionary, which might provide additional language versions of a term. This can help to find a match in case a law index item can neither be found in the preferred nor in the non-preferred items. We use three different external knowledge bases, two of which contain terms in multiple languages and one in German and English. The two knowledge bases containing terms in multiple languages are *DBpedia*[20] and *Wikidata*[21]. The *Standard Thesaurus Wirtschaft (STW)*[22] contains terms in German and English and is used as the third dictionary. In order to find a match between the law index and the corresponding EuroVoc term, we query the external knowledge base for this term and retrieve all language versions of this term that are available in this knowledge base. We then use this list of terms for the EuroVoc lookup. Furthermore, we also combine Wikidata and DBpedia by querying one of these knowledge bases and use the results as an input to query the other knowledge base. The combined results are then used for the EuroVoc lookup.

Table 4.7: Evaluation results of law index and EuroVoc alignment

	Matches found	Preferred term	Correct	Correct %
Direct lookup	112	56	93	0.8304
Wikidata	134	47	94	0.7015
DBpedia	133	52	98	0.7368
STW	78	36	57	0.7308
Wikidata + DBpedia	147	42	81	0.5510
DBpedia + Wikidata	145	44	82	0.5655

4.4.2 Evaluation and Discussion

We perform a manual evaluation of the results shown in Table 4.7. A result found in EuroVoc is counted as correct if the law index and EuroVoc items match. In cases without a literal match, we count the result as correct when it matches semantically from a legal or common sense point of view. In total, we can map 169 law index items directly to a EuroVoc class. It is not surprising that the direct lookup of law index items in EuroVoc delivers the best results in terms of overall correct matches, while the combination of Wikidata and DBpedia (and vice versa) find the most matches, but only around 55% of them are correct. The results also show that the STW dictionary found 78 matches only, which is likely caused by the fact that it only contains English terms and no other language. Furthermore, STW does provide any additional match that is not found either by the direct lookup or using the other knowledge bases.

As for the overall findings, we can say that the direct lookup works best especially for very common terms not specific for a particular domain or jurisdiction, for instance criminal law or labour law. Compound terms indicating different scopes, for instance “generic public services law” vs “specific public services law” are much harder to find while very narrow and non-common terms, for example “steam boiler law” (Dampfkesselrecht) cannot be found at all. Furthermore, using external knowledge bases can lead to false mappings based on wrong translations provided by them and therefore require a manual curation of the found mappings.

4.5 Related Work

For the extraction of legal entities from legal documents we can relate to previous work in the research area of named entity recognition. [Dozier and Haschart, 2000] extracted person names, such as the names from the involved judges, from American judiciary documents using a template-based extraction system. Later on, the range of extracted entities relevant for legal documents has been expanded to courts and jurisdictions [Dozier et al., 2010], using rule-based and statistical models reaching F-scores of up to the 90% range. [Leitner et al., 2019] focus on extracting fine-grained legal entities from a German corpus of legal documents (e.g. the entity of category “person” is further

[20] https://wiki.dbpedia.org/, last accessed 2021-01-31
[21] https://www.wikidata.org/, last accessed 2021-01-31
[22] http://zbw.eu/stw/version/latest/about, last accessed 2021-01-31

split into the categories “judge” and “lawyer”) and compare CRF and deep learning approaches. The authors use a pretrained German Word2Vec language model and show that using neural networks can outperform classic approaches, such as CRF, by up to 10 % depending on the entity class. A similar work using Greek legislative documents is also using deep learning approaches with a custom language model for the extraction of named entities [Angelidis et al., 2018]. In more detail, the authors of this work extract, next to persons and geospatial entities, references to legislation or documents of public organizations and achieve F-scores in the high 80% range.

From a broader perspective, interest in named entity recognition gained traction in the 1990 years, starting with the extraction of entities for which specific designators are available, for instance “person” or “organization” [Nadeau and Sekine, 2007]. These early works mainly relied on rules and lists, which need to be maintained before approaches without such lists were proposed by introducing statistical models to recognize named entities in texts [Mikheev et al., 1999]. Of course, over time machine learning algorithms have been applied in the NER task [Mansouri et al., 2008] before the shift to the application of neural networks [Lample et al., 2016] started. Like us, they also use pretrained word embeddings, which are fine-tuned for the training. Over time and with enhanced methods to extract more and more different named entities, the F-score increased and state-of-the-art approaches reaching results well in the 90% range.

The mentioned related works from the legal domain, but also from other domains, show the shift from rule-based to deep learning approaches. We took these previous works as an inspiration to design our own approach. As rule-based and deep learning approaches showed good results, we also applied them to our scenario of extracting legal entities from court decisions. Of course, these approaches were not directly applicable due to different datasets and extracted entities, but the results are in line to what is reported in previous works.

Previous research work on classifying legal documents in the EU mostly focuses on documents from the European legal database EUR-Lex, either based on the JRC-Acquis corpus, a multilingual aligned parallel corpus with 20+ languages containing documents taken from the European legal database [Steinberger et al., 2006] or the EUR-Lex 4K dataset provided by the Knowledge Engineering Group of the Technical University Darmstadt [Loza Mencía and Fürnkranz, 2010]. The Joint Research Centre (JRC) of the European Commission published the JRC EuroVoc Indexer JEX tool, which treats the classification problem as a profile-based ranking task and reaches – on the former corpus – an F-score between 0.44 and 0.54 depending on the language by ranking the typical features of a class which form the profile [Steinberger et al., 2012].

One of the core findings in their work is that adjusting the stopwords to the domain (which is already a strong hint on the special nature of language of the legal domain) is the most efficient way to boost classification results. Another approach is proposed by [Boella et al., 2015] who transform the multi-label into a single-label problem in order to enable processing by a Support Vector Machine. The authors claim to reach an F-score of 0.75 for the Italian version of the JRC-Acquis corpus, however, the algorithm description [Boella et al., 2012] was not reproducible and the results of an F-score of 0.75 on the classification task cannot be directly deduced from the paper. While details are vague, we suspect that the high F-score is due to the fact that the authors restrict themselves to only the most commonly used labels (above a certain threshold), which makes the classification task significantly easier: one of the main problems in the JRC and EUR-Lex 4K training corpora is that certain labels hardly appear in the training data and in general the label usage is extremely skewed. Other previous work on document classification in the legal domain also shows the common problem of classification tasks with a vast amount of classes and therefore either confirm the bad performance of classification algorithms [Loza Mencía and Fürnkranz, 2010] or approach the problem by reducing the number of classes to boost the results [Alkhatib et al., 2017, Quaresma and Gonçalves, 2010]. An exploratory excursion to an ontology-based training-less classification method by [Alkhatib et al., 2018] shows the same problems of having a skewed class distribution with a micro F-score of 0.29. Taking different routes, some authors exploit semantic methods [Altinel and Ganiz, 2018] or specific sub-domains like sentiment classification [Liu and Chen, 2015]. Many surveys explore the area of text mining in general or describe classification methods in particular [Aggarwal and Zhai, 2012, Allahyari et al., 2017, Hotho et al., 2005]. Our idea in the present work is, inspired by these related works, also attempting to take into account both the semantics and the hierarchical tree structure of the EuroVoc thesaurus and its keywords, in order to boost performance of multi-label document classification.

From a more general point of view, while text classification dates back to the 1960s, it started to gain a lot of interest from the information systems community in the 1990s with the large availability of digital documents and the rise of the machine learning (ML) paradigm [Sebastiani, 2002]. An early overview of multi-label document classification approaches is provided by [Tsoumakas and Katakis, 2007], and the problem transformation strategies, which enable classical methods like SVMs to be applied to the multi-label case, for example using binary classifiers for each class separately. In recent years, a lot of work has focused on *extreme classification*, a term used for multi-label classification in situations where there is a large number of classes, often with a skewed class distribution, and potentially a large number of documents [Zhang et al., 2017]. Some benchmark datasets, and also real-world applications, contain hundreds of thousands of classes, therefore the focus of extreme classification is not only on prediction accuracy but also on computational performance. The datasets discussed herein (based on EUR-Lex and EuroVoc) fall into the category of *small* extreme classification datasets. Some extreme classification methods like SLEEC [Bhatia et al., 2015] reduce the effective number of classes by projecting the output space into a low-dimensional, continuous subspace [Chen and Lin, 2012] – similar to the idea of using word embeddings instead of one-hot encoding. Others use a tree hierarchy as a structural constraint, where trees or forests filter a fraction of classes on each node visited [Prabhu and Varma, 2014]. This leads to logarithmic prediction time. Finally, [Zhang et al., 2017] present a greedy algorithm that combines the low runtime complexity of the primal-dual sparse approach with the simple parallelization of training and the small memory footprint of one-versus-all approaches.

4.6 Summary and Future Directions

In this chapter, we focused on the population of the legal knowledge graph with data from the Austrian legal information system as well as from external sources such as the EuroVoc thesaurus and spatial information from Geonames. We introduced a distinction between different population methods based on the available data and the necessary steps in order to populate the legal knowledge graph. Structured data, for instance available as metadata for the legal documents can be either used for a direct population without additional transformation steps. In addition, we showed that structured data such as the location of courts and their competent spatial jurisdiction can be used to populate the legal knowledge graph. Moreover, additional information also extend the search possibilities leading to a increased information availability for the enduser. The indirect population with structured data requires a few syntactical preprocessing steps such as converting date formats. Finally, we proposed two approaches to align the Austrian law index as part of AustroVoc with the EuroVoc thesaurus.

We showed two possible methods for the population of the legal knowledge graph from unstructured data. The first method we demonstrated is based on the extraction of legal entities, for instance case reference, contributor, court, legal provision, law gazette, legal rule and literature from the document text. Due to the lack of a gold-standard corpus for Austrian legal documents from the RIS, we manually annotated 50 court decisions from the Austrian Supreme Court with the aforementioned legal entities. This new corpus was used to evaluate three different approaches based on rules, conditional random fields and neural networks. The evaluation of the different approaches shows that the performance of a traditional approach using rules can keep up with the state-of-the-art advanced methods using neural networks and language models pretrained on a large corpus of documents without fine-tuning. Even more, the evaluation results show that rules work the better the more structured the legal entities to be extracted are. This is proven with the results for the legal rule entity, while rules perform worst on the less structured literature entities, which occur in various ways. What is more, we cannot determine a “clear winner” regarding the approach to choose for the extraction of legal entities based on the results. The rules (including gazetteers) work best for the extraction of case reference, court and legal rule. CRF delivers the best results for the law gazette and literature entities while neural networks are the best choice for contributor and legal provision. Overall, the results for each approach and over all legal entities we compared are within approximately 15%. This gap might appear to large, we also want to note that the effort required for the approaches is varying. While, rule-based approaches only require a relatively small set to create the rules, ensuring that all variations of how the legal entities can appear are covered, the automated approaches (CRF and neural networks) require a large training set. Such a training set needs to be annotated by humans and computational resources capable of handling a larger number of documents. This puts the 15% gap in perspective as the choice of the “best” approach is a trade-off of required resources and desired performance.

The second method to populate the legal knowledge graph from unstructured data, is the annotation/classification of documents into a given set of classes by the EuroVoc thesaurus. For the evaluation of this method, we used two corpora of legal documents from the European Union (JRC-Acquis and EUR-Lex 4K) and contrasted it with a well-known dataset (Reuters-21578) usually used for such a task. We compared statistical and machine learning approaches as well as neural networks classifying the legal documents into more than 6,000 EuroVoc classes. Furthermore, we demonstrated that exploiting the hierarchical structure of a thesaurus can be used to boost the classification results with a limited information loss by replacing the labels with the topterms or microthesaurus they are associated with. The results show that statistical approaches are outperformed by approaches including the semantics of the documents. The results of the classification task show that the application of neural networks can be suggested as it usually provides the best results. However, the performance difference to the other approaches using semantics is not that large and the statistical approach delivers very good results for precision such that is a question of the goal of the classification task.

The alignment of the Austrian law index items with the EuroVoc thesaurus showed that we can achieve good results using methods mainly based on literal matching. The best results are achieved were a direct mapping of the terms in both thesauri can be found. External knowledge bases providing additional language versions of a term used as a directory can help to align items were no direct (literal) matching is found. For the latter method, a manual curation of the found matches is required as these external knowledge bases might lead to inappropriate mappings.

Future work for the extraction of legal entities and document classification is going into the direction of extending the corpus of annotated Austrian legal documents. While we used a corpus of 50 Supreme Court decisions to demonstrate the extraction of legal entities, it would certainly be interesting to add documents from other (Austrian) courts, for instance the public tribunals. Moreover, we showed that the approaches work for court decisions, but legal professionals also work with other legal documents such as contracts and it would be interesting to investigate if the approaches reach the same performance on these documents as well. Of course, these documents are usually not publicly available and need to be manually annotated. Furthermore, creating an extended language model from Austrian legal documents including laws and court decisions could be beneficial for further research. Such a language model could be used for further experiments with legal entities to investigate whether the already good, an F-score within the 90% range, results could be further improved. The results of the document classification experiments also leave room for some improvement. Extensive experiments with additional approaches using neural networks and language models trained on different corpora have already been investigated by [Shaheen et al., 2020] and show that the results can be improved by adjusting the hyperparameters of the neural network. Another research direction than tweaking the parameters of the used approaches is directed towards the training corpus. The skewed label distribution could be addressed by adding additional documents specific for the underrepresented thesaurus classes or to add external documents giving definitions of the classes. Both approaches would help to sharpen the semantic profile of the labels. For the alignment of the concepts of two thesauri it would certainly be interesting to test approaches including the semantics of these terms, which requires a language model that includes all terms and is able to find similar terms with a good confidence.

Overall, we can say that whether to choose information extraction or classification for the population of the legal knowledge graph is dependent on the properties to be populated. Furthermore, the experiments show that the gap between machine learning and deep learning approaches is not so big whenever texts are highly structured, such as it often is the case for legal documents, and that also rule-based approaches can deliver good results. The final choice of the approach to be used should be taking the required human and computational resources into account and what should be achieved.

CHAPTER 5

Temporal Information in Court Decisions

It is also important to represent time-related information in a structured way to enable further processing. For that purpose, dedicated formats have been published, for instance TIDES TIMEX [Ferro et al., 2005] and TimeML [Pustejovsky et al., 2003b]. Temporal information is also contained in legal documents, especially in judiciary documents (court decisions) from which they can be extracted and used for further processing. Temporal information occurs in absolute, for instance as a date, and relative formats, for instance as time intervals or references to another points in time. In order to process and use them, for example to summarize documents in the form of a list of consecutive events or visualize them in a timeline, we need to extract temporal information and put them into context, hence associating them with surrounding information, for example the actor or subject. In Section 5.1, we start with an analysis of temporal information in legal documents and introduce three temporal dimensions specific for the legal domain. Furthermore, we manually create a new corpus (named TempCourt, freely available online[1]) composed of 30 legal documents from three different courts, namely the European Court of Human Rights, the European Court of Justice and the United States Supreme Court. We use this corpus to compare the performance of ten state-of-the-art temporal taggers applied to legal documents. Moreover, temporal expressions also play an important role for legal events as presented in Section 5.2. More in detail, we provide another corpus consisting of 30 documents of the European Court of Justice annotated with legal events and their components. This second corpus is used to extract legal events from court decisions with the goal to provide a quick overview of what happened when in a case. In order to do so, we evaluate three different approaches based on rules, conditional random fields and deep learning methods. Finally, the related work is presented in Section 5.3 and Section 5.4 summarizes the chapter including a view on possible future research directions.

5.1 Temporal Tagging of Court Decisions

An emerging area of research is the use of text analytics to derive structured data from legal text (e.g. norms, opinions, recommendations or court decisions). One of the most

[1] `https://tempcourt.github.io/TempCourt/`, last accessed 2021-01-31

relevant activities is the automatic extraction and processing of events and temporal expressions with a view to creating timelines. In this context, a *temporal expression* (TE) is a word or sequence of words making reference to a time instant (e.g. "seven o'clock") or a time interval (e.g. "from seven to ten"). Temporal expressions frame events or happenings implicitly or explicitly mentioned in the document. *Temporal relations* bind TEs to events and determine the relative position of events with respect to other events. Such relations are, for example, "before" and "after".

The example below is a text excerpt from a court decision of the European Court of Human Rights describing the facts of the *Aras v. Turkey* case (no. 21824/07, 20 July 2017). The text contains three TEs (in bold below), two of them being in an absolute form (e.g. "11 December 2002") and one in a relative form ("same day").

> "*On **11 December 2002** the applicant's statement was taken by the public prosecutor and, on the **same day**, the judge at Istanbul State Security Court ordered her detention on remand. On **7 December 2002** the applicant was arrested on suspicion of membership of a terrorist organisation.*"[2]

This temporal information is related to three events, namely, (i) the public prosecutor taking the statement; (ii) the judge ordering a detention; and (iii) the applicant being arrested. Each of the events is related to the other entities, either named ("Istanbul State Security Court") or not ("the applicant"). Even though the two absolute dates in the text above appear in the same format, this is not always the case and very often different formats are used even within the same document. Although our exemplary legal case can be used to motivate an investigation into both temporal and event extraction (e.g. [Schilder, 2005, Uzzaman et al., 2010]), in this section we focus specifically on temporal expressions.

Temporal taggers operate on texts like the one above, performing different tasks, namely TE *identification*, *normalization*, and *classification*. *Identification* (also called *detection* or *extraction*) is a task, which involves finding TEs and their start and end position in the text. *Normalization* (or *anchoring*) is a task that interprets TEs to obtain specific instants and intervals represented in a standard format. This task resolves relative TEs (as "the same day") from context information, localizes time formats (i.e. mm/dd/yy vs dd/mm/yy), considers timezones and enables the reformatting of the TEs into a standard format (e.g. ISO 8601[3]). In contrast, a *classification* task is used to determine which kind of TEs have been found. For instance, **on 7 December 2002** is most likely a time point, while **from 7 December 2002 to 12 December 2002** is a time interval. The temporal expressions found by the temporal taggers are usually represented in domain-agnostic formats, such as TimeML[4]. TimeML is the most widely accepted mark-up language for temporal expressions, and its use is justified over domain-specific formats (e.g. Akoma Ntoso[5] in the legal domain) as it permits representing more details and nuances specific to the temporal terms.

Although several temporal taggers have been proposed and investigated in different domains, the suitability of existing methods to extract temporal information from legal texts has been relatively unexplored to date as being only a side effect for other tasks, for

[2]European Court of Human Rights, Aras vs Turkey, case no. 21824/02, 20 July 2017

[3]`https://www.iso.org/iso-8601-date-and-time-format.html`, last accessed 2021-01-31

[4]`https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_specs_1.2.1.html`, last accessed 2021-01-31

[5]`http://www.akomantoso.org/`, last accessed 2021-01-31

instance document classification or reasoning over documents. Additionally, the lack of temporal resources in the domain is a major drawback when it comes to research in this direction. To the best of the authors' knowledge, there is no preexisting temporal annotation gold standard based on legal text corpora. Consequently, there is no previous evaluation of how well standard temporal tagging tools perform in the legal domain.

5.1.1 Particularities of Legal English

Temporal information has an effect on the version of the applicable law and it creates a chronological order of events in a legal case. Sometimes it is important to know whether event *A* or event *B* happened first. In addition, temporal information is also used to assess whether past events may be time-barred.

When it comes to the automatic extraction of temporal information from legal documents, it is important to highlight that legal documents, and in particular court decisions, slightly differ in structure and writing style from documents from other domains. These differences include deeper parse trees, differences in part-of-speech distributions and more words per sentence [Dell'Orletta et al., 2012].

Judgments are usually framed in legal processes following specific procedures. Events and timings mentioned in the judgment constitute context information that should not be lost in the annotation process. An example of this is the concept of *preliminary ruling*, a legal term referring to a phase previous to the decision when the European Court of Justice is asked how a law should be interpreted, being therefore a reference to this period and a hint for temporal localization of other events. Also specific events happening in legal frameworks must be considered when processing legal texts, as done in other domains such as the medical domain [Styler IV et al., 2014].

Structure of Judgments

Table 5.1 illustrates the differences in document structure for judgments made by the European Court of Justice (ECJ) and the United States Supreme Court (USSC), and preliminary assessments of applications submitted to the European Court for Human Rights (ECHR). The court decisions from the European courts follow a similar structure that already hints which categories of TE could be expected in different parts of the texts. In particular, both ECJ and ECHR start with a description of the involved parties (section A) and are then followed by a case summary (B), stating concisely why this case has been brought to the respective court and what happened so far in terms of the legal proceedings. In ECJ decisions, the legal proceedings are followed by the applicable legal framework and then by the case description, whereas the ECHR structure is the other way round. The decisions of the ECJ and ECHR courts conclude with the matching of the law with the facts of the case under the legal basis (E) and the resulting judgment (F). In contrast to ECHR documents, the "Legal framework" section (D) in ECJ documents cites European and local legislation, without any direct references to the case itself, and as such this information was excluded from the final documents in the TempCourt corpus. Although TEs corresponding to other related events such as prior decisions could be extracted from these sections, we focus on case-related temporal information and leave the extraction of events for future work. Apart from beginning with the involved parties (A) in a particular case, the structure of USSC decisions is quite different. The second

Table 5.1: Structure of ECJ, ECHR and USSC decisions.

Section	ECJ	ECHR	USSC
A	Involved parties	Involved parties	Involved parties
B	Case summary	Procedure	Syllabus
C	Legal framework	Circumstances of the case	Main Opinion
D	Circumstances of the case	Legal framework	Concurring and dissenting opinions
E	Court assessment	Court assessment	
F	Judgment	Judgment	

section (B) is called “syllabus” and contains a short summary of the case. It is followed by the main opinion (C), that includes the final decision of the court and explains how the court came to this decision, by referring to the legal foundations. The last part of a decision states, where applicable, the concurring and dissenting opinions of the involved justices (D). An opinion is called “concurring” if a justice follows the main opinion but grounds the decision on a different legal rationale. A dissenting opinion is issued in cases where a justice disagrees with the main opinion and the underlying legal rationale. Following a consistent structure makes legal documents comparable, and fulfills the expectations of readers who are used to find a specific kind of information always at the same place in the same kind of legal document. Furthermore, the consistent structure of legal documents (from the same authority or within a jurisdiction) leads to expectations with respect to the type of temporal information, which could be expected in each section of the document. We expect temporal references describing the facts of the case (“what happened when?”), which could be used for generating timelines for document summarization, to be present in the *case summary* (ECHR), *Circumstances of the case* (ECJ) or *syllabus* (USSC) sections in the judgment of the respective deciding court, but mentions to general temporal events to appear throughout the entire document. The structural properties of legal documents could also be exploited for the automatic creation of timelines as legal documents can be very long. For the analysis of a judgment it is necessary to understand the order of the events as this can affect the legal proceedings. The easier understanding could be supported with a visual representation of the order of events, hence a timeline that shows the important events and provides a visual summary of the case.

Dates are used in virtually every domain. In contrast to posts published in social media, e.g. Facebook or Twitter, where every user might write dates in different formats, documents from official authorities, such as courts, usually use the same format to represent dates in all documents. Differences in date representation that can be noticed are for instance the order of day and month or the used separators. Therefore, differences in date representation are seldom found within a document, but may vary from court to court.

Mistaken or Misleading Temporal Expressions in Legal Documents

References to legal documents often include some sort of temporal information, usually forming a text pattern prone to be confused with a true temporal expression (see examples in Table 5.2). Typical references containing temporal information are references to

Table 5.2: Examples of mistaken and misleading temporal information.

Source	Example	Description
ECHR	no. 7334/13, 127 - 128, ECHR **2016**	Reference to another case
ECHR	Timoshin v. Russia (**dec.**)	Reference to a decision (dec.), often confused with the month of December
ECJ	OJ **2008** L 348 p. 98	Reference to official journal of the EU
ECJ	Directive **2008**/115/EC	Reference to a directive published in 2008
USC	See Va. Code Ann. §53.1-165.1 (**2013**)	Law reference
USC	[...] 772 F. 3d **1328**, **1333** (CA10 **2014**)	Precedent case reference

previous court cases, laws or legal literature, where the temporal information indicates a point in time when the respective reference has been decided or published. However, temporal information contained in references is not considered relevant for a specific case in terms of describing "what happened when?".

For example, the expressions in Table 5.2 convey some temporal information, e.g., four-digit sequences that could be recognized as years, but which only in some cases indeed refer to actual years. Tagging these kind of expressions as TEs may become a major problem and lead to additional errors —for instance nearby dates in the text can be normalized from these wrong references leading to further errors in the entire text. Additionally, references to other legal documents often present their creation date, that must be differentiated from dates in the document timeline of referred case events. An example of this, where the given date refers to the date of a Council Directive of the European Union and thus is irrelevant for the narrative of the text, is the excerpt below:

> "*Council Directive 93/13/EEC of **5 April 1993** on unfair terms in consumer contracts must be interpreted as not precluding (...)*"[6]

For processing these kinds of expressions, we could first detect and hide them from the temporal tagger (e.g., replacing them for an innocuous expression before the processing and restoring them afterwards) or alternatively we could filter them in a post-processing step.

5.1.2 Incompleteness of the TimeML Standard for the Annotation of Legal Documents

During the annotation of the TempCourt corpus, we also detected relevant information that the TimeML standard is not able to represent. The main drawbacks of the TimeML standard applied to legal documents are summarized in the following subsections.

Specific Legal Terminology as Modifiers

Documents in the legal domain are rich in non-colloquial noun phrases representing temporal information. For example, the sentence "the *expiry* of the three-day period" is badly understood by parsers in comparison with "the *end* of the three-day period".

[6] Council Directive 93/13/EEC, `https://eur-lex.europa.eu/eli/dir/1993/13/oj`, last accessed 2021-03-12.

Similarly, when the extension of a duration is uncertain (e.g, the range between two points in time, such as in "period of between seven and thirty days"), there is no way to properly represent the uncertainty. Likewise, when referring to different possibilities frequently found in the legal language such as "was a year or two more of prison time", this information cannot be properly annotated — even if some taggers such as SUTime [Chang et al., 2012] provide alternative values for similar expressions, i.e. "from one to two years", the standard specification does not allow them.

The standard should be able to represent all these particularities of the legal domain. Similarly, a temporal tagger for the legal domain should be able to reason with this level of granularity.

Missing Levels of Granularity

Not only points in time, which are used to determine the applicability of a particular law, but also durations are of high importance, especially in formal laws determining the limitations of time (e.g. to plead the statute of limitations) for actions that must be taken before they preclude. For instance, in the legal domain a different way to count days is often applied. While `DURATION` is sufficient to indicate the absolute lapse of time, TimeML is not capable to indicate a non-absolute duration such as "10 working days".

Temporal taggers could be enhanced with external knowledge to recognize special constraints being applied to durations, for instance, work calendars where working days are identified. Eventually, also the capability to reason at this level of granularity would be desirable.

Exhaustive List of Attributes

The TimeML attribute `functionInDocument` allows for the marking of some temporal expressions as special reference ones, but just as one among: *"creation_time"*, *"expiration_time"*, *"modification_time"*, *"publication_time"*, *"release_time"*, *"reception_time"* or *"none"*. This is not enough for legal documents, where domain expressions such as *"lodgement_time"*, *"argued_time"* or *"decision_time"* would be more useful. Domain-specific extensions to the TimeML standard could be used to solve this particular problem.

Limited Expressivity of the Existing Format

There are temporal expressions whose anchor time is not the DCT (Document Creation Time) nor are they related to any temporal expressions in the text, but in other legal documents cited in the text, such as in:

> "*The dissent also relies heavily on Missouri v. Frye, 566 U. S. 134 (2012), and Lafler v. Cooper, 566 U. S. 156 (2012). (...) Lafler, decided* ***the same day*** *as Frye (...).*"[7]

To cover this issue, a temporal tagger needs to be combined with a co-reference system in order to find the matching events to which a certain temporal expression relates. This

[7] US Supreme Court, Lee vs United States, 23 June 2017

could be addressed by making use of the clear structure of legal documents, which usually use the same citation style in all documents such that temporal expressions appearing next to case references can be annotated as belonging to them.

The official TIMEX3 tags cannot properly represent precise intervals on their own. A time interval such as "between 12.45 and 18.45" can only be represented as a DURATION (of 4 hours) or as two unrelated datetime points. This is a problem in cases where exact intervals are needed to solve legal problems, such as confirming an alibi or evaluating exact timespans.

While some of these limitations could also be found anecdotally in other kinds of texts, they are common in legal documents, and relevant to their temporal dimension. Other non legal issues raised when using the TimeML standard are the correct extent of the tags or how to deal with multiple normalization options. For example, "one decade" can be represented as "`P1DE`" or "`P10Y`", and "a few weeks later" can be seen as a duration with a known *beginPoint* or as a `FUTURE_REF`).

5.1.3 Temporal Dimensions

In legal texts temporal expressions can be attributed to different temporal dimensions. We identify three different temporal dimensions and illustrate them based on the example decision *Sophie Mukarubega v Préfet de police and Préfet de la Seine-Saint-Denis* [European Court of Justice, 2014].

Temporal Dimension of the Legal Process

Each court proceeding is based on some formal rules and new events are added with the gradual advancement of the legal proceeding. This temporal dimension covers events related to the legal process itself, for instance the date a lawsuit has been filed, date of the hearings or the decision date.

> "*By a decision of **21 March 2011**, adopted after hearing the person concerned, the Director General of the Office francais de protection des refugies et apatrides (OFPRA) (Office for the protection of refugees and stateless persons) rejected her application for asylum. (...)*" [European Court of Justice, 2014]

This temporal expression indicates that a certain event has happened, in this case the rejection of asylum.

Temporal Dimension of the Case

This temporal dimension covers factual information about the case, which serves as the basis for a judgment.

> "*Ms Mukarubega, who was born on 12 March 1986 and is of Rwandan nationality, entered France on **10 September 2009** in possession of a passport bearing a visa. (...)*"[8] [European Court of Justice, 2014]

[8]Please note that the same sentence contains two temporal expressions, which are attributed to two different temporal dimensions.

The highlighted date refers to a fact of the case, hence the point in time when the person entered France.

Temporal Dimension of the Legal Context

Temporal information can also affect the legal context and determine the applicable law and the degree of the resulting penalty. This is especially relevant when determining the limitation of liability in time or when checking a legal reference to know the applicable law version. We can illustrate this in the following example of a preliminary ruling request to the European Court of Justice with the dates marked in bold.

> "*(...) This request for a preliminary ruling concerns the interpretation of Article 6 of Directive 2008/115/EC of the European Parliament and of the Council of* ***16 December 2008*** *(...)*" [European Court of Justice, 2014]

> "*Ms Mukarubega, who was born on* ***12 March 1986*** *and is of Rwandan nationality, entered France on* ***10 September 2009*** *in possession of a passport bearing a visa. (...)*"[9] [European Court of Justice, 2014]

> "*By a decision of* ***21 March 2011****, adopted after hearing the person concerned, the Director General of the Office francais de protection des refugies et apatrides (OFPRA) (Office for the protection of refugees and stateless persons) rejected her application for asylum. (...)*" [European Court of Justice, 2014]

The first ("16 December 2008"), third ("10 September 2009") and fourth ("21 March 2011") temporal expression refer each to a point in time that is relevant for the legal context. A preliminary ruling for the interpretation of an article requires the article to exist (first date). In the second paragraph, the birth date is general information about the defendant, which does not affect the *temporal dimension of the case* but might influence the *temporal dimension of the legal context*. This is especially important in criminal cases when the birth date in conjunction with the date of the offence constitutes the application of the criminal law relating to juvenile offenders. The third date, on the other hand, refers to a fact of the case, the day of entrance in the host country, being therefore part of the *temporal dimension of the case*. Finally, the fourth date indicates when a decision on the case in the legal process was reached, so this TE corresponds to the *temporal dimension of the legal process*.

Conflict of Temporal Dimensions

One could wonder whether there is the possibility of an overlap of temporal dimensions such that a single event might be part of the *temporal dimension of the legal process* and of the *temporal dimension of the case*. For instance, in cases that go through the entire hierarchy of courts, decisions are reversed by higher courts and referred back to the previous court. In these cases, the judgments of the previous courts do have an influence on the following proceedings. This means that courts might be bound to former judgments or receive an order to investigate certain parts of a former proceeding in more detail and do more investigation work. However, from our perspective the *temporal*

[9]Please note that the same sentence contains two temporal expressions, which are attributed to two different temporal dimensions.

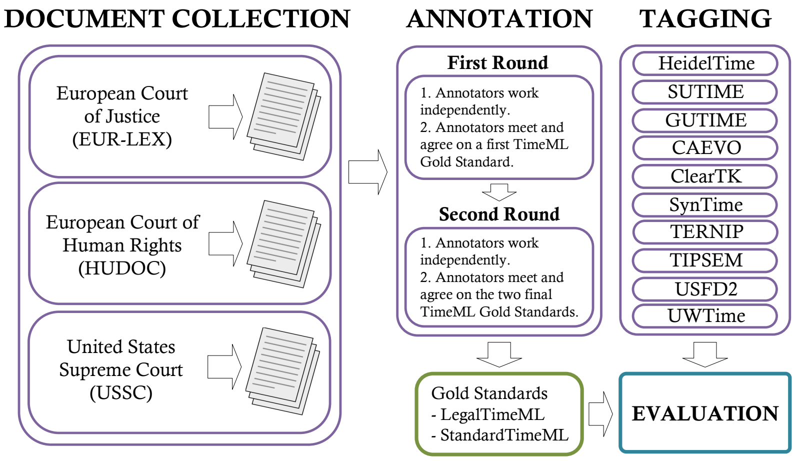

Figure 5.1: Outline of our work, including document collection, annotation and evaluation of taggers.

dimension of the case encompass the events inherent to the case, while revisions and case remands do not change anything in the temporal order of events in the original case, instead such information adds context being relevant for the further proceeding without affecting the *temporal dimension of the case*.

In this section, we outlined the particularities of documents in the legal domain, which encompass the special structure of judgments and legal terminology. Furthermore, we described annotation standards such as TimeML and its incompleteness for annotation tasks in the legal domain and introduced a classification schema of temporal dimensions present in judgments.

5.1.4 Corpus and Annotation

In this section, we aim at evaluating in how far the automatic identification (and normalization) of temporal expressions is feasible using existing taggers, and to test the effectiveness of such tools. In order to enable such an evaluation, we propose two gold standards, one domain focused (LegalTimeML, composed of temporal information important for the facts of the case) and one generic (StandardTimeML, including all temporal information). Both gold standards can be used to compare the results of temporal taggers and to determine which of them is most suited to be used when working with legal documents. The temporal annotation of all documents used in this work is based on the TimeML annotation language[10]. Figure 5.1 illustrates the methodology we followed in order to create and evaluate our proposed gold standards. In the document collection phase we retrieve the documents, and in the annotation phase we create the gold standards in two iterations, which are then used to compare the results retrieved from the temporal taggers in the tagging phase.

[10] `https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_annguide_1.2.1.pdf`, last accessed 2021-01-31

Document Collection

Although different types of documents could have been chosen to create a gold standard in the legal domain, our proposed corpus TempCourt is composed of judgments and preliminary assessments of applications as they contain a large number of temporal expressions.

As many of the taggers do not have full support to other languages, we selected court decisions in English to enable a fair comparison of the results of the temporal taggers. Also, in order to increase the variety of ways in which temporal information is represented in different types of courts, we decided to investigate the judgments of courts acting in different jurisdictions and domains. Specifically, we focus on the court decisions of the European Court of Justice (ECJ), which is the highest court of the European Union, the United States Supreme Court (USSC), and on preliminary assessments of applications submitted to the European Court of Human Rights (ECHR). The documents for the two European courts are available in the respective databases, namely EUR-Lex[11] for the ECJ and HUDOC[12] for the ECHR, while the USSC documents were collected from the website of the United States Supreme Court[13]. The corpus created for this work, named TempCourt, consists of thirty court decisions, composed of an even distribution of ten documents per court in each subcorpus. Documents provided by the European Court of Human Rights are allowed to be reproduced for private use or for the purposes of information and education in connection with the Court's activities when the source is indicated and the reproduction is free of charge[14]. The same policy applies to documents retrieved from EUR-Lex whose documents are allowed to be reused in conjunction with the Commission Decision of 12 December 2011 on the reuse of Commission Documents[15] for commercial and non-commercial purposes given the source is acknowledged[16]. Documents published by US governmental institutions (such as the US Supreme Court) are in the public domain[17].

Legal documents often contain names of persons, especially court decisions. The documents in our corpus contain the names of the involved judges and the names of parties in a non-anonymized way. Names are considered personal data and need to respect the General Data Protection Regulation[18] (GDPR), which in the case of public data involves providing transparency with respect to the processing on request (Article 14 GDPR). Consent for the processing of personal data from the data subject is not required for public data.

For the purpose of temporal annotations, we are mainly interested in the section of the court decisions describing the facts of a case, because we expect to find the most valuable temporal information about the chronology of a case in this section, whereas temporal information in other sections is expected to be relating to laws or previous cases. Therefore, we omitted the "Legal framework" section of the ECJ documents in our corpus.

---

[11] http://eur-lex.europa.eu/, last accessed 2021-01-31

[12] https://hudoc.echr.coe.int/, last accessed 2021-01-31

[13] https://www.supremecourt.gov/, last accessed 2021-01-31

[14] https://echr.coe.int/Pages/home.aspx?p=disclaimer&c=, last accessed 2021-01-31

[15] http://data.europa.eu/eli/dec/2011/833/oj, last accessed 2021-01-31

[16] https://eur-lex.europa.eu/content/legal-notice/legal-notice.html#droits, last accessed 2021-01-31

[17] https://www.copyright.gov/title17/92chap1.html#105, last accessed 2021-01-31

[18] Regulation (EU) 2016/679.

Table 5.3: Corpus statistics

Corpus	# Doc.	# Tokens	Doc. Size (Avg. KB)	Doc. Size (Avg. Tokens)	Sentence length (Avg. Tokens)
ECHR	10	7,252	4	725	13
ECJ	10	53,044	32	5,304	32
USSC	10	50,874	25	5,087	18
Total	30	111,170	20	3,705	21

Table 5.4: Statistics of corpora annotated with TimeML in literature.

Corpus	# Doc.	# Tokens	Doc. Size (Avg. Tokens)
TimeBank[19]	183	61,000	428.7
AQUAINT[20]	73	34,154	467.9
TempEval-3 Eval. (*Platinium*) [UzZaman et al., 2013][21]	20	~6,000	~300
WikiWars [Strötgen and Gertz, 2016]	22	119,468	5,430.4
Time4SMS [Strötgen and Gertz, 2016]	1,000	20,176	20.2
Time4SCI [Strötgen and Gertz, 2016]	50	19,194	383.9

The figures in Table 5.3 illustrate the differences between documents depending on their source. Although we include documents from three different courts, the corpus statistics show that the documents in the ECJ and USSC subcorpora are similar in terms of document size and length. The documents in the ECHR subcorpus are only one fifth in terms of size in comparison with the other two subcorpora. As stated previously, legal texts often make use of very long and complicated sentences to explain legal details, thus we also included the average sentence length in tokens for each corpus. We show that the sentences of the ECHR are roughly one third of length compared to the USSC court decisions, and also tend to be shorter than the ones in the ECJ corpus. These numbers contrast with those relating to corpora from other domains and sources, such as Wikipedia articles (25.1 words per sentence [Kajiwara et al., 2016]), the CONLL 2007 corpus of documents from the Wall Street Journal (24 and 23.4 tokens per sentence in training and test data, respectively [Nivre et al., 2007]) and the basic corpus of everyday documents [Pellow et al., 2014], including all kind of common texts, such as banking or shopping documents (with an average of 17.2 words per sentence). Regarding the amount of documents in each corpus, Table 5.4 provides an overview of the size of referential corpora manually annotated with TimeML. These figures provide evidence that despite the fact that we have less documents per corpus, our corpus is substantially bigger in terms of tokens than most of the previous corpora.

[19] `https://catalog.ldc.upenn.edu/docs/LDC2006T08/timebank.html`, last accessed 2021-01-31. The website mentions 61k non-punct tokens, [Strötgen and Gertz, 2016] report 78,444 tokens.

[20] `http://www.timeml.org/timebank/aquaint-timeml/aquaint_timeml_1.0.tar.gz`

[21] Just approximate figures were provided [UzZaman et al., 2013].

Annotation

For each subcorpus (ECJ, ECHR and USSC), the ten documents were selected at random. In order to compare the results of different temporal annotation tools, all thirty documents have been annotated in multiple steps. In the first part of the annotation process, two different annotators performed the manual annotation of the documents following the TimeML guidelines[22]. Once manual annotation, which was done independently by two persons using General Architecture for Text Engineering (GATE) [Cunningham et al., 2013], was completed, they met to create a gold standard with annotations agreed by both annotators. When doubts arose, the TimeML guidelines were consulted specifically looking for similar cases; if the doubt persisted, also the TIDES TIMEX2 guidelines[23] were examined, as referred to in the TimeML annotation guidelines. However, due to the particularities of the legal domain, some annotation decisions needed further discussion as shown in the following examples:

1. The word `now` is heavily used in legal documents and was only annotated when it was not used as an adverb, hence the meaning is not `currently` or `at the moment`. For instance in the case ECJ C-34/13[24]: "*[...] so the provision is* `now` *worded as follows [...]*".

2. For the annotation of references to the present time, some taggers use the *PRESENT_REF* token as a value, while others normalize to a date, which is usually the creation date. We decided that we should follow the latter approach for the legal domain, since all the documents in the corpus contain this information and humans would also be able to derive it.

3. Legal documents, especially judgments, often contain references to previous court decisions in the legal grounding of a decision. The citation of such preceding cases depends on how decisions of such courts are usually referenced. Typically, a year is contained in the citation and annotated as a temporal reference. Temporal information contained in identifiers used to refer to collections of court decisions (e.g. *2006*I) or included in the document identifier, should not be annotated (e.g. EC:C:*2013*:180).

4. Expressions such as `the date indicated`, appearing for instance in the excerpt "*the application lodged on* ***the date indicated*** *in (...)*" are not considered as temporal references but as co-references, being therefore not annotated in the gold standard, since a temporal tagger would not be expected to do so.

The discussion between the two annotators resulted in the creation of two gold standards *StandardTimeML* and *LegalTimeML*:

1. **StandardTimeML** annotates all the TEs following the TimeML guidelines, and uses the *PRESENT_REF*, *PAST_REF* and *FUTURE_REF* tokens as usually done in the domain.

2. **LegalTimeML** annotates just the TEs relevant to the narratives of the judgment, following the particularities in the legal domain previously discussed (no dates in

[22]`https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml\_annguide_1.2.1.pdf`, last accessed 2021-01-31

[23]`https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf`

[24]European Court of Justice, case C-34/13, 10 September 2014

legal references, normalize to dates...). As per the *StandardTimeML* annotation set, it follows the guidelines but does not annotate all the expressions, being therefore a subset considering domain particularities.

The Inter-Annotator Agreement (IAA) between both gold standards is high (0.95), as well as Cohen's kappa [Cohen, 1960] (0.94) and Scott's Pi [Scott, 1955] (0.94), indicating that the normalization of the TE's that are included in both annotation sets have a high agreement. If we check differences between annotations, we find an average of 13.1 common TEs per document, 0.3 partial coincidences and about 16.2 TEs that are not contained in the *LegalTimeML* but appear in the *StandardTimeML*. The recall among both annotation sets is 0.44, while precision is 0.90, which confirms that a lot of TEs are not relevant for the case timeline (44% with regard to the ones annotated following the full TimeML standard), but that the way the annotators tag them is highly similar.

Tagging

Once the corpus was collected, the following temporal taggers were executed over our legal corpus, as they represent the different approaches available and are the most widely used in literature: HeidelTime [Strötgen et al., 2012], SUTime [Chang et al., 2012] GUTime (which is part of the TARSQI toolkit) [Verhagen et al., 2005], CAEVO [Chambers et al., 2014], ClearTK-TimeML [Bethard, 2013], SYNTime [Zhong et al., 2017], TERNIP [Northwood, 2010], TIPSem [Llorens et al., 2010], USFD2 [Derczynski et al., 2010] and UWTime [Lee et al., 2014]. These temporal taggers will be introduced in Section 5.1.5. HeidelTime was used in its configuration for narrative text. GUTime was used as a part of the TARSQI toolkit, using it alone with the preprocessor in the pipeline. Since the code available online was just able to annotate an specific corpus, USFD2 was slightly modified in order to annotate any input and to generate TIMEX3 tags as output[25]. All other taggers were used with default parametrization.

The output of the taggers, which generated offline annotations (such as GUTIME/-TARSQI) were modified in order to be comparable with the output of the rest of the taggers and ensure they were readable by GATE. These processes were executed using a new coded converter, which added the temporal annotations to the document and excluded non-temporal entities. Once the outputs of all the taggers were in the same format, they were loaded into the same GATE document, which contained twelve annotation sets (two for the manually-created gold standards and one for each of the ten temporal taggers).

Final Corpus

The final documents have been generated in several formats[26]. First, as GATE XML documents, that facilitate the storage of different annotation sets and also the visual and numerical comparison of the different sets. Second, a set of TimeML documents (TML) is provided for each of the manual gold standards. These documents contain the same annotations as in the correspondent annotation set in the corresponding GATE document, but makes the comparison with the output of other temporal taggers easier, as it is in the *official* TimeML format. Also a set of TML documents without any tag is provided

[25]The functionality and the rules were not modified.
[26]The final corpus can be downloaded at: `https://tempcourt.github.io/TempCourt/`

Table 5.5: Overview of temporal taggers

(*) Not all types are covered.

Temporal Tagger	Approach	Identification	Normalization	Events	Relations
HeidelTime (HE)	rule-based	✓	✓	-	-
SUTime (SU)	rule-based	✓	✓	-	-
GUTime (GU)	hybrid	✓	✓	✓	✓
CAEVO (CA)	hybrid	✓	✓	✓	✓
ClearTK (CL)	machine-learning	✓	-	✓	✓
SynTime (SY)	rule-based	✓	-	-	-
TERNIP (TE)	rule-based	✓	✓	-	-
TIPSem (TI)	hybrid	✓	✓	✓	✓
USFD2 (US)	hybrid	*	*	-	*
UWTime (UW)	hybrid	✓	✓	-	-

to facilitate testing. These TML documents have been validated using the TimeML validator[27] from TempEval-3[28], so it is guaranteed that they fulfill the guidelines of the TimeML standard. Finally, all original documents are stored as TXT-files; these documents are of similar size in terms of kilobyte and length in tokens as shown in Table 5.3.

5.1.5 Approach

Many of the temporal taggers described in the literature over the last few years are no longer available, not maintained, or just work for previous annotation schemas like the formerly mentioned TIMEX2. Some examples are DANTE [Mazur and Dale, 2009], TEA [Han et al., 2006], JU_CSE [Kolya et al., 2013] or ManTIME [Filannino and Nenadic, 2015]. Therefore, we focus on the most widely used active temporal taggers, which are often cited in literature and report good results on corpora from different domains, or have successfully participated in well-known temporal challenges, such as TempEval-3[29]. Table 5.5 provides an overview of the temporal taggers under investigation for which an implementation is freely available. The first column ("Temporal Tagger") is used to refer to particular temporal taggers later on.

The following aspects will be discussed for each tagger: supported languages, used approach, covered functionality, parametrization options, implementation language, availability, integration and interoperability with other software and dependencies on other resources and required installations.

[27]http://www.cs.york.ac.uk/semeval-2013/task1/data/uploads/timeml-validator-1.1a.tar.gz
[28]https://www.cs.york.ac.uk/semeval-2013/task1/, last accessed 2021-01-31
[29]https://www.cs.york.ac.uk/semeval-2013/task1/, last accessed 2021-01-31

Tasks of Temporal Taggers

The functionalities of temporal taggers can be classified into four categories as shown in Table 5.5. Some temporal taggers support all functionalities, while other taggers require additional tools.

- *Identification* means that the system is actually able to identify temporal expressions in a text compared to other systems, which are only used for normalization of already tagged texts.
- *Normalization* refers to the ability to represent temporal information in the written text into the corresponding standard value following the ISO 8601 norm, which can be further processed. For instance, expressions like "the next day" refer to the day before, which might be indicated with an explicit date in the text, and the temporal tagger is able to normalize this expression and assign the actual date as the value to the temporal annotation.
- *Events* are real-world situations at a particular time and are classified into seven categories, such as `OCCURENCE, STATE` or `REPORTING`, in the TimeML standard [Saurí et al., 2006].
- *Relations* indicate a certain connection between events, times or a mixture of both usually classified into temporal `TLINK`, subordination `SLINK` and aspectual `ALINK` links [Saurí et al., 2006].

The detection of temporal expressions in a text is based on different approaches. Some taggers use rules for both identification and normalization tasks, while others use machine learning for the former. Also hybrid approaches have been proposed in literature. Nevertheless, it must be noted that normalization is generally tackled using rules, even when the identification is done otherwise.

Rule-based Approach

Temporal information is detected based on manually created rules (e.g. regular expressions), which need to cover all possible variations of how temporal information might be expressed. Thoroughly created rules are expected to perform better than other approaches, but come with the disadvantage of being inflexible. A missing or erroneous rule will prevent the temporal tagger from finding a temporal expression.

HeidelTime [Strötgen et al., 2012] is a rule-based domain-sensitive temporal tagger. Available for more than 200 languages (just 13 of them based on manually developed resources, the rest of them being automatically created), it offers the option to select from four different text categories: *News, Narratives, Colloquial* and *Scientific*, the last two are only available for English texts. HeidelTime covers both TE identification and normalization, having different strategies for each domain. HeidelTime, implemented in Java, can be used as a standalone version[30], or integrated in other pipeline environments like the General Architecture for Text Engineering (GATE) [Cunningham et al., 2013] or a UIMA[31] pipeline. In spite of being one of the most popular temporal tagging tools, to the best of our knowledge, it has never been used in the legal domain.

[30]`https://github.com/HeidelTime/heideltime/`, last accessed 2021-01-31
[31]`https://uima.apache.org`, last accessed 2021-01-31

SUTime [Chang et al., 2012] is the Stanford CoreNLP [Manning et al., 2014] annotator for temporal expressions. SUTime is a rule-based temporal tagger built on the TokenRegex tool [Chang et al., 2014] (a pattern definition service also part of CoreNLP), able to both identify and normalize TEs. SUTime produces TimeML/TIMEX3 tags with new attributes not included in the standard, such an alternative value more flexible than the one covered by the standard. SUTime presents several related limitations, as analyzed by the authors themselves [Chang et al., 2012], and offers no domain adaptation. SUTime is available as part of the CoreNLP pipeline for different languages. The Java code[32] is available online, and also a GATE plugin and a Python wrapper have been developed[33].

SynTime [Zhong et al., 2017] is a rule-based tagger that proposes a *type-based* approach. It defines different types of tokens (*time tokens, modifiers* and *numerals*) with similar syntactic behaviour and builds heuristic rules on these types instead of doing it on strings or regular expressions. As the types are domain independent and the rules work on types, the system is designed to be domain and language independent. Nevertheless, in order to work in different domains or languages, more tokens need to be added for each type. SynTime only performs TEs identification, and does not normalize them. SynTime is written in Java and available online[34] and uses the Stanford CoreNLP library for Part of Speech (POS) disambiguation.

TERNIP (Temporal Expression Recognition and Normalisation in Python) [Northwood, 2010] is a rule-based Python 2.7 library that identifies and normalizes TEs. The rules used for both subtasks can be easily extended. It only covers English texts and provides no domain adaptations. TERNIP can be used as an API or be integrated as a GATE processing resource, via an XGAPP file (a GATE application file format) available with the code on github[35]. TERNIP relies on the Natural Language Toolkit library (NLTK) [Loper and Bird, 2002].

Machine learning-based Approach

In contrast to rule-based approaches machine learning-based temporal taggers do not rely on previously created rules to identify temporal expressions. Machine learning techniques make temporal taggers much more flexible and enables them to detect temporal expressions in an unexpected form, however it requires a good pretrained model based on a large annotated corpus that supports a variety of temporal expressions, which can be expected later in the document to be tagged with temporal expressions. A poor training set with missing variations of temporal expressions will result in a poor performance of the temporal tagger in terms of *precision*[36] and *recall*[37].

[32] `https://github.com/stanfordnlp/CoreNLP/tree/master/src/edu/stanford/nlp/time`, last accessed 2021-01-31
[33] `https://nlp.stanford.edu/software/sutime.shtml\#Extensions`, last accessed 2021-01-31
[34] `https://github.com/xszhong/syntime`, last accessed 2021-01-31
[35] `https://github.com/cnorthwood/ternip`, last accessed 2021-01-31
[36] Fraction of the results identified which were correct.
[37] Fraction of the results that should have been found which were correctly identified.

ClearTK-TimeML [Bethard, 2013] identifies temporal information in English texts using external machine learning tools. It uses specific annotators modelled as a BIO[38] token-chunking (for extent/identification of the expressions) or as a multiclass classification task (for types and attribute classification). The TIMEN normalization tool [Llorens et al., 2012] is suggested for the normalization task as this is not covered by ClearTK-TimeML. The features used are the ones proved to be the most successful in previous independent temporal taggers, and are extracted by a morpho-syntactic annotation pipeline with tools like OpenNLP and Apache. While ClearTK-TimeML does not offer domain-specific adaptions, the pipeline and the parameters can be customized by users. ClearTK-TimeML is written in Java and is available online[39].

Hybrid Approach

Hybrid approaches combine rules with machine learning. For instance, creating rules of a large corpus with machine learning techniques, which are manually refined afterwards.

GUTime [Mani et al., 2000] was developed at the Georgetown University originally for the temporal annotation of news. GUTime was subsequently incorporated into TARSQI, a modular system for automatic temporal annotation [Verhagen et al., 2005]. The approach of GUTime is different from the temporal taggers previously mentioned, as it does not only use rules to find temporal expressions, but it also applies a hybrid approach of rules and machine-learning techniques. The hand-crafted rules serve in GUTime as a basis for temporal annotations that are extended by additional machine-learning ones discovered using the C4.5 algorithm [Quinlan, 1993], i.e. rules to support term disambiguation. The TARSQI framework is also able to extract events and relations from English texts. TARSQI is written in Python[40] and well described.

CAEVO [Chambers et al., 2014] (CAscading EVent Ordering) is a sieve-based architecture, which uses twelve different classifiers based on rules and machine learning, pipelined in a cascade way, starting with the one with the highest precision. Even when these classifiers work individually, some transitivity constraints are imposed; also the order of the classifiers can be modified, and new sieves can be added. In contrast to other taggers, CAEVO focuses on the extraction of temporal relations for event ordering, producing *dense* temporal graphs where events and temporal expressions are heavily connected. CAEVO is an expansion of NavyTime [Chambers, 2013] and reuses part of the code of ClearTK-TimeML [Bethard, 2013] for part of its sieves. It works just for English texts and has no domain adaptations. CAEVO is written in Java and is available online in a github repository[41].

TIPSem [Llorens et al., 2010] (Temporal Information Processing based on Semantic information) is an hybrid temporal tagger able to extract temporal information from English and Spanish texts. TIPSem uses both Semantic Role Labeling [Gildea and

[38]Beginning of, Inside of, Outside of a time expression.
[39]`https://cleartk.github.io/cleartk/docs/module/cleartk_timeml.html`, last accessed 2021-01-31
[40]`https://github.com/tarsqi/ttk`, last accessed 2021-01-31
[41]`https://github.com/nchambers/caevo`, last accessed 2021-01-31

Jurafsky, 2002] and Conditional Random Field (CRF) [Lafferty et al., 2001] models. Different features are used by CRF recognition models, such as morphological or syntactic considerations at token level, along with semantic level ones such as the Role, the Governing Verb or Lexical Semantic information for each token. Similar features are used at tag level for classification. Finally, the relation extraction features differ depending on the type of relation. TIPSem therefore tackles all the temporal tasks. The Java code is available online[42], but it requires installation of additional software, and also optional libraries for certain languages, such as Spanish.

USFD2 [Derczynski et al., 2010] is a temporal tagger focusing on TEs and relations, using a rule-based approach for TEs and both rules and the NLTK's Maximum Entropy classifier for relations. USFD2 obtains a good recall with a smaller set of rules when compared with other taggers, since they consider specific heuristics for scpecific tags, such as `DATEs` and `DURATIONSs` as Temporal Expression types, that are the most common. USFD2 only supports the extraction of temporal information from English texts. The Python code of USFD2 is available online[43], but it must be noted that it is developed for the evaluation of specific datasets, so it must be slightly modified for custom use. This has been done for the results on our corpus.

UWTime [Lee et al., 2014] follows a hybrid approach, using a Combinatory Categorial Grammar (CCG) [Steedman et al., 2011] parser with hand-crafted rules and learning. UWTime just tackles the recognition and normalization of temporal expressions. It uses features such as surrounding tokens and POS, lexical and dependency information, and relies on techniques such as AdaBoost [Freund et al., 1999] for optimization. UWTime is only available in English with no domain particularities. It can be downloaded online[44], used as an API or as a server. UWTime relies on Stanford CoreNLP software.

5.1.6 Evaluation and Discussion

The final step of our research methodology involved a comparison of the effectiveness of all ten taggers on the two gold standards, along with the analysis of the results.

Evaluation Methodology

After having all documents annotated with the ten different temporal taggers, we evaluated the results for which we used the typical *precision*, *recall* and *F-measure* metrics, which are commonly used in literature for the evaluation of extraction and normalization of temporal annotations [Strötgen et al., 2012]. It is worth noting that we elected to provide both the *strict*-F-measure (which only considers completely correct and ignores partially correct annotations) and the *lenient*-F-measure, that admits partial annotations. The reason to do so is that while it is important to identify the complete temporal expression, it is also true that some taggers correctly normalize an expression even if they do not fully cover it. It also must be taken into account that in some cases

[42] `https://github.com/hllorens/otip`, last accessed 2021-01-31

[43] `https://github.com/leondz/usfd2`, `https://code.google.com/archive/p/usfd2/`, last accessed 2021-01-31

[44] `https://bitbucket.org/kentonl/uwtime-standalone`, last accessed 2021-01-31

the correct extent of a temporal expression is not clearly derivable from the guidelines. For this reason, we decided that providing both measures would allow for the evaluation of both the degree of support with respect to the guidelines and the actual detection capabilities.

The evaluation process was designed in a way to avoid a bias or preference towards a particular temporal tagger. Therefore, the results of all taggers are consolidated in a single document with individual annotation sets for each tagger containing the temporal annotations and respective features. Each evaluation involves a key set (the correct reference) and a response set (the annotations to evaluate). Since the goal is to create gold standards for the legal domain, the manually annotated temporal expressions in both annotation sets, *LegalTimeML* (LTML) and *StandardTimeML* (STML), serve as the key sets. The annotation sets of each tagger act as the response set for each evaluation run. We therefore evaluated each automatic tagger for all three sections of the corpus (i.e. the documents from the three different legal sources) against each of the manually created gold standards *LegalTimeML* and *StandardTimeML* and calculated the *lenient* and *strict* precision, recall and F-measure.

All the temporal taggers were applied to the corpus with the standard configuration and without domain-specific modifications to achieve better results specifically for the legal domain[45]. The standard configuration was chosen in order to evaluate the out-of-the-box performance of each temporal tagger and their suitability when applied to the legal domain. The average number of annotations per corpus in both gold standards (STML and LTML) and the various taggers are shown in Table 5.6, which illustrates the occurrences of different TIMEX3 annotation types (`DATE`, `DURATION`, `TIME`, `SET`) for each analyzed corpus. It is clearly shown that the most used annotation type in court decisions is DATE. This result is not surprising as the date is considered to be sufficient in most cases because the actual time of the day is not relevant. Furthermore, deadlines in the legal domain usually indicate the end of the day and it is not important if an action is taken in the morning or in the afternoon. It must also be noted that the pattern of appearances of each of the TIMEX3 types does not fit any of those of the domains analyzed by [Strötgen et al., 2012] (news, narratives, colloquial and scientific).

Table 5.7 clearly shows that most taggers perform well on the short ECHR subcorpus and tend to find the same number of annotations as contained in the gold standard. If we focus on the *lenient* figures, we can see that the errors are mostly in the extension of the tagging more than in its identification. In the ECJ and USSC subcorpora (Tables 5.8 and 5.9 respectively), the number of annotations by the taggers differs from the gold standards, especially HeidelTime draws attention to its annotations in the ECJ corpus. When looking into the documents, the reason for this significant difference becomes obvious. The designators of European legal acts such as regulations and directives follow a special scheme, which also includes the year when the legal act has been agreed. A typical designator of an EU directive is therefore, for instance ***2016/679***, which is considered to be a designator of a legal act but it is not a valuable temporal reference within a court decision.

[45]Except USFD2.

Table 5.6: Average number of annotation types per document for each corpus

D = Date, Dur = Duration, S = Set, T = Time

	ECHR				ECJ				USSC			
Tagger	**D**	**Dur**	**S**	**T**	**D**	**Dur**	**S**	**T**	**D**	**Dur**	**S**	**T**
StandardTimeML	11.6	1.3	1.0	0	31.5	4.3	2.0	2.7	35.7	5.6	3.5	4.0
LegalTimeML	10.1	1.3	1.0	0	16.8	4.3	1.5	3.0	9.1	5.4	1.5	0
HeidelTime	11.4	1.7	1.0	0	68.1	5.3	1.0	1.0	41.6	5.6	1.5	2.0
SUTime	11.3	2.0	0	0	39.1	3.9	1.3	1.3	46.9	7.9	1.5	2.7
GUTime	11.7	0	0	0	31.4	1.0	0	0	37.3	2.0	0	0
CAEVO	11.1	1.8	0	0	36.7	5.8	1.0	1.5	39.9	9.4	1.5	3.0
ClearTK	10.2	1.0	0	0	38.6	3.4	0	0	36.1	5.1	1.0	2.0
Syntime	11.5	0	0	0	39.1	0	0	0	47.8	0	0	0
TERNIP	11.7	1.7	0	0	30.3	3.6	0	0	33.3	5.6	1.0	0
TIPSem	13.0	1.0	0	0	38.4	2.6	0	0	-	-	-	-
USFD2	13.9	2.0	0	0	66.6	3.3	0	0	28.4	3.8	0	0
UWTime	11.0	2.5	0	0	-	-	-	-	-	-	-	-

Results

From the results shown in Tables 5.7 (ECHR), 5.8 (ECJ) and 5.9 (USSC), we can see that the performance of the individual temporal taggers is quite similar for each section of the corpus. Furthermore, the numbers for all three measures that have been calculated are unexpectedly high for some taggers in comparison to the application of temporal taggers (out of the box without any domain-specific modifications) in the case of non legal text. Nevertheless, they tend to be less performant than results previously reported by taggers in general evaluations[46] [Chang et al., 2012].

On the ECHR corpus most taggers perform equally well when *strictly* evaluated, while GUTime provides the best results, closely followed by TERNIP. On the contrary, TIPSem, USFD2 and UWTime are not as performant. This is because the ECHR uses fully qualified dates (e.g. "10 January 2017") and does not include many references to other court decisions. It also must be noted that most taggers (except of GUTime, SynTime and TERNIP) struggle with identifying dates denoting the birthdates of the persons involved in the cases and case numbers, with some also normalizing them. In addition, we want to note that big differences between *lenient* and *strict* values, such as those of UWTime and ClearTK-TimeML, do not only show differences in the extent of the annotation, but also impact the normalization results. For instance, if instead of marking up "October 13", just "October" is marked, the *lenient* score will count it as positive, the *strict* will not, but the normalization will for sure be wrong.

One outlier in the figures of the ECJ corpus can be spotted immediately, which is the precision of the HeidelTime annotations that is significantly different from other

[46] `https://github.com/HeidelTime/heideltime/wiki/Evaluation-Results`, last accessed 2021-01-31

Table 5.7: Evaluation results for the ECHR corpus for each temporal tagger

Both for identification (two first columns, *lenient* and *strict*) and normalization (two last columns, *lenient* and *strict*). The first row (in white) corresponds to results against the *StandardTimeML* gold standard, while the second (in gray) corresponds to the *LegalTimeML* gold standard. *P=Precision, R=Recall, F=F-score. Best results highlighted in boldface.*

	lenient			**strict**			**lenient+ value**			**strict+ value**		
A	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
HE	**0.99**	**0.99**	**0.99**	0.84	0.84	0.84	0.78	0.78	0.78	0.78	0.78	0.78
	0.88	**0.99**	**0.93**	0.71	0.80	0.75	0.67	0.75	0.71	0.64	0.72	0.68
SU	0.88	0.87	0.88	0.85	0.84	0.84	0.78	0.78	0.78	0.76	0.75	0.75
	0.76	0.85	0.80	0.71	0.80	0.76	0.66	0.74	0.79	0.64	0.72	0.68
GU	0.96	0.93	0.94	**0.95**	0.92	**0.93**	**0.86**	0.84	0.85	**0.86**	0.84	**0.85**
	0.84	0.92	0.88	**0.83**	**0.92**	**0.87**	0.74	0.82	0.78	**0.74**	0.82	**0.78**
CA	0.88	0.87	0.87	0.83	0.82	0.82	0.78	0.78	0.78	0.75	0.75	0.75
	0.75	0.85	0.80	0.70	0.79	0.74	0.65	0.74	0.69	0.64	0.72	0.67
CL	0.92	0.78	0.85	0.34	0.32	0.35	-	-	-	-	-	-
	0.80	0.77	0.78	0.33	0.32	0.33	-	-	-	-	-	-
SY	0.98	0.93	0.96	0.83	0.79	0.81	0	0	0	0	0	0
	0.86	0.93	0.90	0.70	0.76	0.73	0	0	0	0	0	0
TE	0.94	0.95	0.95	0.92	**0.93**	0.92	**0.86**	**0.88**	**0.87**	0.85	**0.86**	**0.85**
	0.83	0.95	0.89	0.80	**0.92**	0.85	**0.75**	**0.86**	**0.80**	0.72	**0.83**	0.77
TI	0.78	0.85	0.81	0.64	0.70	0.67	0.64	0.71	0.67	0.63	0.69	0.66
	0.69	0.86	0.76	0.62	0.77	0.69	0.64	0.79	0.71	0.61	0.76	0.68
US	0.73	0.61	0.67	0.69	0.58	0.63	0	0	0	0	0	0
	0.65	0.62	0.64	0.61	0.58	0.60	0	0	0	0	0	0
UW	0.90	0.53	0.67	0.51	0.30	0.38	0.55	0.33	0.41	0.51	0.30	0.38
	0.86	0.58	0.69	0.48	0.32	0.38	0.51	0.34	0.41	0.48	0.32	0.38

precision values. The much better performance of GUTime in the ECJ corpus can be explained by the fact that it does not annotate numbers referring to collections of judgments, in contrast to TIPSem and ClearTK-TimeML.

The USSC corpus is slightly different to ECHR and ECJ as it uses a different date format and it also repeats parts of the text in the judgment, which leads to poorer performance as incorrect annotations are also repeated.

Different date formats are a typical challenge when applying a temporal tagger to a corpus. Typically, dates found across all evaluated documents are fully qualified dates

Table 5.8: Evaluation results for the ECJ corpus for each temporal tagger

Both for identification (two first columns, *lenient* and *strict*) and normalization (two last columns, *lenient* and *strict*). The first row (in white) corresponds to results against the *StandardTimeML* gold standard, while the second (in gray) corresponds to the *LegalTimeML* gold standard. *P=Precision, R=Recall, F=F-score. Best results highlighted in boldface.*

	lenient			**strict**			**lenient+ value**			**strict+ value**		
A	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
HE	0.48	0.95	0.64	0.47	0.94	0.63	0.47	**0.94**	0.62	0.47	**0.93**	0.62
	0.27	0.97	0.42	0.26	**0.96**	0.41	0.26	**0.94**	0.40	0.26	**0.93**	0.40
SU	0.81	0.97	0.88	0.79	**0.95**	0.86	0.78	0.93	0.85	0.77	0.92	0.84
	0.44	0.95	0.60	0.43	0.93	0.58	0.41	0.90	0.57	0.41	0.89	0.56
GU	**0.97**	0.87	0.91	**0.97**	0.86	**0.91**	0.94	0.84	0.89	0.94	0.84	0.88
	0.51	0.82	0.63	0.50	0.82	0.62	0.48	0.78	0.60	0.48	0.78	0.60
CA	0.89	0.74	0.81	0.85	0.70	0.77	0.86	0.71	0.77	0.85	0.70	0.77
	0.49	0.74	0.59	0.46	0.70	0.56	0.46	0.70	0.56	0.46	0.69	0.55
CL	0.77	0.88	0.82	0.32	0.36	0.34	-	-	-	-	-	-
	0.42	0.88	0.57	0.18	0.37	0.24	-	-	-	-	-	-
SY	0.89	**0.99**	**0.93**	0.81	0.90	0.85	0	0	0	0	0	0
	0.49	**0.98**	0.65	0.46	0.92	0.61	0	0	0	0	0	0
TE	**0.97**	0.88	0.92	0.96	0.88	**0.91**	**0.96**	0.87	**0.91**	**0.95**	0.87	**0.91**
	0.54	0.89	**0.67**	**0.53**	0.88	**0.66**	**0.53**	0.88	**0.65**	**0.52**	0.87	**0.65**
TI	0.72	0.81	0.76	0.64	0.72	0.68	0.62	0.70	0.65	0.61	0.69	0.65
	0.41	0.83	0.54	0.37	0.75	0.49	0.35	0.71	0.47	0.34	0.70	0.46
US	0.31	0.54	0.39	0.29	0.51	0.37	0.02	0.04	0.03	0.02	0.03	0.02
	0.20	0.65	0.31	0.19	0.61	0.29	0.02	0.06	0.03	0.02	0.05	0.02
UW	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-

containing a day, the month in full and a year. The format in which these dates are provided are different for European and American sources of legal documents. The date in Europe is usually indicated in the format "Day, Month, Year" (e.g. "10 January 2017"), whereas the American date format is "Month, Day, Year" (e.g. "January 10, 2017"). This particular difference in the date format has been processed correctly by some taggers, such as HeidelTime and SUTime, annotating both versions as a single date. GUTime, however, was not reliable in this context, despite the fact that it is the best tagger in the other corpora. It either detected only one part of the American-formatted date (e.g. "January 10") or it treated both parts of the same date as two different annotations.

Table 5.9: Evaluation results for the USSC corpus for each temporal tagger

Both for identification (two first columns, *lenient* and *strict*) and normalization (two last columns, *lenient* and *strict*). The first row (in white) corresponds to results against the *StandardTimeML* gold standard, while the second (in gray) corresponds to the *LegalTimeML* gold standard. *P=Precision, R=Recall, F=F-score. Best results highlighted in boldface.*

	lenient			**strict**			**lenient+ value**			**strict+ value**		
A	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
HE	0.83	0.94	0.88	0.81	0.92	**0.86**	0.79	**0.90**	**0.84**	0.79	**0.89**	**0.83**
	0.29	0.97	0.44	0.26	0.88	0.40	0.20	0.67	0.31	0.19	0.64	0.29
SU	0.75	**0.99**	0.85	0.72	**0.95**	0.82	0.67	0.88	0.76	0.66	0.86	0.75
	0.25	**0.98**	0.40	0.23	**0.90**	0.36	0.18	0.72	0.29	0.17	**0.69**	0.28
GU	0.84	0.78	0.81	0.71	0.66	0.69	0.67	0.62	0.65	0.65	0.60	0.62
	0.25	0.69	0.36	0.16	0.45	0.23	0.12	0.34	0.18	0.10	0.27	0.14
CA	0.77	0.90	0.82	0.72	0.84	0.77	0.73	0.85	0.78	0.71	0.83	0.76
	0.23	0.82	0.36	0.21	0.72	0.32	0.21	**0.73**	0.33	0.20	**0.69**	0.30
CL	0.85	0.84	0.84	0.81	0.79	0.80	-	-	-	-	-	-
	0.30	0.89	0.45	0.26	0.78	0.39	-	-	-	-	-	-
SY	0.85	0.98	**0.91**	0.78	0.91	0.84	0	0	0	0	0	0
	0.28	**0.98**	0.44	0.24	0.84	0.37	0	0	0	0	0	0
TE	**0.93**	0.86	0.90	**0.90**	0.83	**0.86**	**0.86**	0.79	0.83	**0.85**	0.78	0.81
	0.32	0.90	**0.48**	**0.29**	0.81	**0.43**	**0.25**	0.69	**0.37**	**0.23**	0.64	**0.34**
TI	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-
US	0.50	0.30	0.37	0.39	0.23	0.29	0.08	0.02	0.08	0.02	0.01	0.02
	0.16	0.28	0.21	0.07	0.13	0.09	0.08	0.14	0.10	0.03	0.05	0.04
UW	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-

The performance of GUTime in terms of precision, recall and F-measure is pretty good over all three subcorpora. However, GUTime performs poorly on the USSC corpus. An inspection of the GUTime annotations in this corpus confirms the fact that GUTime has a hard time recognizing dates in the American format, as already pointed out above, an issue that is also reflected in normalization figures. In contrast to GUTime, TERNIP is able to maintain the performance.

In summary, although the results of the evaluation are promising, it is worth noting that legal documents, especially court decisions, have some particularities (such as those highlighted in Section 5.1.1), which cause some stumbling blocks for automatic temporal taggers being applied out of the box. An example of this would be the case of "dec.", a non-temporal expression that appears when citing *decisions on admissibility*[47] that most taggers (such as CAEVO or SUTime) normalize as *December*.

With regard to the comparison between the two reference standards, if we check the differences between the results and focus on the recall, we see that the best taggers remain more or less the same (GUTime, TERNIP, SUTime and HeidelTime, since although SynTime performs well in terms of recognition it does not provide a normalization value). Due to the fact that the taggers are not trained for the particularities of the LTML annotation set, the precision is not expected to be high and does not indicate the tagger's usefulness.

Comparative Analysis of Temporal Taggers

The thorough analysis of the corpus documents and the manual inspection of the most frequent errors of the taggers led to the synthesis of a collection of test cases that present the phrases prone to cause errors. The most salient results are described below, where the output of the tagger is represented in bold and the correct tag is underlined.

HeidelTime is able to identify temporal modifiers (e.g. "at least five years") automatically and add the feature to the annotation. However, it fails to detect the correct date format (e.g. DD/MM/YYYY vs MM/DD/YYYY) and fails to recognize the indication of the age of mentioned persons (e.g. "*a* ***62-year-old*** *woman*"). It does not normalize expressions like *today* and annotates them with the value `PRESENT_REF`. In legal texts, it tags references to other documents or IDs (e.g. *"No* ***1612/68****"*, *"No. 15-****1031****"*, *"See Pet. for Cert.* ***5-7****"*) as temporal expressions. It also has an interval option that does not work well in this kind of document.

SUTime also fails to identify the correct date representation form (e.g. DD/MM/YYYY vs MM/DD/YYYY). In addition, SUTime exhibits inconsistencies when parsing the same expression in different paragraphs, and it also wrongly annotates expressions like *"****fall****"*, *"****may****"* as temporal expressions although they refer to an action *"to fall"*, *"may"* instead of the season. SUTime also has some limitations with respect to ambiguity resolution or non-whole numbers recognition.

Although GUTime has a good performance in general, sometimes it does not normalize some expressions and has problems with some ways to represent hours (e.g., it does not recognize "*(...) between 12.15 and 18.45*", nor if it was expressed as "*12:15 and 18:45*", it just recognizes "***12h15*** *and* ***18h45***"). Also some DURATIONs are not recognized, series or dates neither (in "*15 and* ***16 December 2008***" it just recognizes the part in bold) , and sometimes it tags expressions that look like years, such as "*EUR* ***2000***".

CAEVO does normalize DATEs in the format DD/MM/YYYY as MM/DD/YYYY, so it does not even recognize the ones not fitting it, such as "*25/03/2016*". It also partially annotates expressions such as "*On the next* ***day***" (categorizing it as a DURATION) and tags separately "***once a week***", as a `PAST_REF DATE` and a `DURATION`, respectively. It also does not recognize "15" in "*15 and* ***16 December 2008***", and tags "***62-year-old*** *woman*", year-like expressions as *"§****1101****"* and time-like expressions as "*Order in No.* ***2:10****-cv-02698 (WD Tenn.)*". Finally, it also tags separately "*sentenced to* ***a year*** *and* ***a day*** *in prison*".

[47]`http://www.echr.coe.int/Documents/Note_citation_ENG.pdf`, last accessed 2021-01-31

Similarly to GUTime, ClearTK-TimeML does not recognize TIMEs when expressed as in "(...) between 12.15 and 18.45"; it does not either recognize expressions like "09/01/1981" as DATEs. Some DURATIONs are also not recognized (e.g. "at least five years"), and tags expressions such as "**May**" or "**62-year-old** woman". It annotates expressions such as "23 **January 2013**" or "once **a week**" partially and categorizes them as DURATION.

SynTime just normalizes to the current date when it is executed. Although it is able to recognize expressions such as "***15*** *and* ***16 December 2008***", it fails when it finds expressions such as "*as amended by Council Regulation (EC) No* ***1791/2006 of 20 November 2006***", where it annotates all in bold, not just the underlined correct part. It also seems to recognize all four-digit expressions as years (e.g.. "*See 10 U. S. C. §****1408****(c)(1).*", "*So. 3d* ***1264, 1269-1272***") and ambiguous expressions as "***may***", "***the second***" or "***fall***", but fails to fully annotate some temporal expressions (e.g. "*per* ***month***", "*May 15,* ***2017***").

TERNIP tags expressions such as "*EUR* ***2000***", "***may***", "***fall***", but fails to identify some DATEs and DURATIONs. It also does not identify "13" in "*13 and 27 October 2008*", but is on the other hand is able to recognize misspelled temporal expressions such as "***eighth months***" (even if it is not correctly normalized). It also tags "*303, 98 Stat.* ***2045,*** *21 U. S. C. §853(a)(1),*" as DATE expressions .

TIPSem is not able to annotate the USSC documents. Furthermore, it does not recognize the first DATE in the ECJ subset, which is expressed in the format `DD Month YYYY`. However, as it recognizes the dates in the remainder of the document without a problem, it is probably due to a lack of a syntactic/semantic context for the first date. TIPSem also wrongly tags expressions such as "*Directives* ***2004/83, 2005/85*** *and 2003/9]*", "***Article 5 of Directive 2008/115***", "**Directive 2001/42**" or "***the judgment of 28 February 2012***" as temporal expressions. It also tags expressions, for instance "***MON 810***" or random numbers and words, like "***4,285***" and "*(in* ***euros)i***", and tends to mark them as `FUTURE_REF`. Additionally, it does not recognize some dates, for example "*29/02/2016*", while other similar dates are correctly tagged, like "***28/09/2016***".

USFD2 is unable to parse some of the documents in the corpus, throwing errors when trying to normalize expressions it considers out of the range and warnings for some ASCII codes. It also tags some numbers randomly, such as in "*amending Regulation (EEC) No* ***1612/68*** *and repealing Directives 64/****221****/EEC, 68/****360****/EEC*" and always normalizes DATEs to the present day. It does not recognize straightforward dates and tags ambiguous words even when they are a part of another word, such as in "*Soto****may****or*". Moreover, TIME expressions are categorized as DATEs.

Finally, UWTime is not able to parse long legal sentences, throwing several errors because of the lack of head rules defined for some of the expressions it finds. In our corpus, it was not able to annotate even a third of the documents.

The most commonly occurring errors in which the taggers fail, whether because they happen frequently in the text or because they contain several temporal expressions, are the following:

- Separation of whole `SET` expressions as "*Once a week*" into "*Once*" and "*a week*", converting one `SET` into a `PAST_REF`, `DATE` and a `DURATION`.

- Not recognizing series of `DATE`s such as "*15 and 16 December*", but detecting the last `DATE` of such a series only.
- Separation of `DURATION`s, for example "*One year and one day*", into two different `DURATION`s.
- In some documents, as it also happens in other kinds of legal texts, such as in the previously mentioned transactional ones, some information is put into brackets, such as in "*before the expiry of a period of* ***[48] hours***". Usually, generic temporal taggers are not able to detect them, for instance tagging just "*hours*" in this concrete example.
- Tagging general ambiguous expressions, like "*fall*" or "*may*" or specific ambiguous ones such as the previously described case of "*dec.*", as temporal expressions.
- Tagging year-like expressions such as "*No* ***1612****/68*" or "*§****1408***". Most taggers tag every four-digit number as a year.
- Problems with dates expressed in the format "DD/MM/YYYY", frequently in identification but in some cases also in normalization.
- Identification of a currency value as a year, for instance "*EUR* ***2000***".
- Tagging of generic expressions such as "*62-year-old*".
- Most taggers do not take modifiers (`mod`) into account, probably because of the low ratio of appearance of `SET`s in other domains, despite the fact that they are extremely important in legal documents. Namely, HeidelTime correctly tagged[48] 17 out of 28 modifiers, while TERNIP tagged 10 out of 28 correctly. The remaining taggers tagged no modifiers with the exception of UWTime, which tagged modifiers in ECHR documents, but not correctly.
- The case of quantifier (`quant`) and frequency (`freq`) attributes is similar. While HeidelTime marks correctly 2 out of 11 `quant`, and marks incorrectly two `freq` as `1` when it should be `1X`. TERNIP only marks one `quant` (incorrectly as the quantifier is indicated in capital letters) out of 11 and no `freq`.

Despite the errors of non-domain specific temporal taggers, the results show an overall good performance for the extraction of temporal expressions from court decisions. Such temporal expressions are also important in a broader context when it comes to the extraction of legal events from court decisions.

5.2 Event Extraction from Court Decisions

Courts elaborate on the facts of a case, involved parties, interpretations of the circumstances, the applicable law and legal principles, and finally the legal assessment leading to the decision. Legal professionals constantly extract, interpret and reason with and about prior cases whilst arguing for a decision in a current, undecided case. However, court decision texts can be long and complex and thus time-consuming to read. Therefore, it would be beneficial to find a means to provide a quick overview of

[48]Some cases, such as distinctions between *EQUAL_OR_LESS* / *LESS_THAN* (for UWTime) and *LATE* / *END* and *EARLY* / *START* (for TERNIP) were counted as errors.

a case, thereby helping to turn decisions into operational, consumable and actionable legal knowledge.

In this section, we focus specifically on Natural Language Processing (NLP) techniques to automatically extract the essence of a court case. Besides extracting general legal rules from individual cases, we aim at providing a quick overview of what happened, who was involved and when the event took place. In the terminology of NLP, event extraction can be treated as a *text classification task* with the goal to assign text fragments (typically, paragraphs, sentences or smaller parts of documents) to predefined (event) classes [Sebastiani, 2002]. Another, related NLP task is Named Entity Recognition (NER), which extracts entities referred to in texts and classifies them into categories [Grishman and Sundheim, 1996], for instance people, places and organizations. Moreover, named entities can also be domain-specific, for instance, courts or laws. Event extraction can benefit from NER, since it can be used to enrich events with relevant information, such as the parties involved. In this section, we focus on the extraction of events and their components from court decisions of the European Court of Human Rights (ECHR)[49] based on a sample thereof.

5.2.1 Corpus and Annotation

Similarly to the extraction of temporal information from court decisions, there is no gold standard corpus with annotated legal events available, which is why we introduce a new corpus used to extract events and the annotation process.

Corpus

The corpus consists of 30 decisions of the ECHR. The ECHR decisions were chosen because they contain: (i) different types of time-related events concerning different actors in comparison with the decisions of the Court of Justice of the EU [María Navas-Loro, 2018]; and (ii) a standard structure in which different legal events are embedded. ECHR decisions are divided into several sections containing specific information according to Rule 74 of the Rules of the Court [Registry of the Court, 2020]: the *preamble* and the *introduction* are followed by *facts*, which contain information about the formal procedure and the circumstances of the case providing details about what happened. The following *law* section describes the legal situations and states the alleged violation(s). The document concludes with the *decision* section. For the extraction of legal events, we use the mentioned document structure excluding the *law* section and focus on the procedure, circumstances and decision.

Annotation

The corpus was annotated by two legal experts in several iterations. The experts annotated independently and then met with a third person to reach a consensus on the disagreements. As we focus on event extraction for the automated timeline generation, we are interested in information that is relevant for the search or extraction of time-related information, such as events, processes, temporal information, and the parties involved in

[49] `https://echr.coe.int/`, last accessed 2021-01-31

these events. As time-related events of cases are linguistically expressed, we annotated the most salient candidate passages thereof. The decisions were manually annotated following the frame "who-when-what". In order to illustrate the applicability thereof, we make use of an annotated paragraph of the case Altay v. Turkey (no. 2), no. 11236/09, 9 April 2019 (a case referring to respect of private life):

> "*On 29 May 2008 the applicant lodged an application with the Edirne Enforcement Court for the restriction on the conversations between him and his lawyer to be lifted.*"[50]

"*Who*" corresponds to the subject of the event, which can be a subject, but also an object (e.g. an application). In the example shown, the subject is "(the) applicant". "*When*" refers to the date of the event, or to any temporal reference thereto; in the paragraph considered, the "when" is the "29 May 2008". "*What*" usually corresponds to the main verb reflecting the baseline of all the paragraph, which in this case is "lodged". Additionally, we include thereto a *complementing* verb or object whenever the core verb is not self-explicit or requires an extension to attain a sufficient meaning. For example, in the paragraph considered, the "what" is "lodged an application", another example is "dismiss an action". "*Event*" relates to the extent of text containing contextual event-related information. The *type* of such annotations can be either *circumstance* – meaning that the event correspond to the facts under judgment; or *procedure*– wherein the event belongs to the procedural dimension of the case. This includes court procedures (legal proceedings stricto sensu), but also actions that trigger procedural effects. A further analysis of this distinction can be found in Section 5.1 and previous literature [María Navas-Loro, 2018]. In the paragraph at stake, we annotated the whole sentence as an event, which is of type "procedure".

Further, we have annotated events and their temporal dimension (related-time events) with concrete guidelines:

- Extension of *what* event element

  One *what* event element can also include two or more close-related verbs, e.g. "divorced" and "agree on custody", instead of annotating two connected verbs autonomously. Moreover, whenever there is some causal relationship between events, we annotate merely one, e.g. "they drink water and they felt unwell".

- Repeated events

  When there is reference to events happening in several dates (e.g. "the dates of birthday of three applicants, respectively"), we annotate solely one event as the *what*, and add just one annotation that covers all the related dates.

- Non-dated events

  Events that are not dated, though semantically expressing an implicit time reference, are then annotated as "when". Examples for non-dated events are time expressions like "the same date", "this afternoon", "on unspecified dates" and "in a number of occasions".

- Non-annotated events

  Some events are not considered relevant to be depicted in a timeline, and therefore not annotated, e.g. the fact that "X was born in X" seems irrelevant.

[50] European Court of Human Rights, Altay v. Turkey (no. 2), no. 11236/09, 9 April 2019.

- Factuality

  Events that are mentioned in the text but do not occur, are yet annotated with the feature "factuality", but not included in the timeline. When events are negated, this feature equals to "NOT", for instance, a party does not appeal against a decision.

- Difficult and blurred annotations

  During the annotation process, some events were difficult to tag, and others sparked a discussion about how to do it, challenging the stipulated guidelines and evidencing how complex and subjective annotating tasks can be. One example that triggered the discussion on the type of events between procedure/circumstance is the sentence "*On 26 February 2014 the Deputy Town Prosecutor carried out an inspection of remand prison SIZO-6.*"[51]. The issue in this sentences relates to the semantics attributed to the role "Deputy Town Prosecutor", which renders the idea of being a court magistrate, and as such, it would be deemed as a procedural event. Herein, the function instead refers to an inspection task, without procedural effect.

5.2.2 Approach

Herein we describe different methods used in our experiments for the extraction of events and their components in the ECHR court decisions. The applied approaches include deep learning and embedding based, conditional random fields and rule-based methods. The corpus and the code is available on Github[52].

The task of assigning one or multiple classes from a set of classes to a text fragment is called text classification [Sebastiani, 2002]. Fragments in our context are typically sentences that are classified into the types "procedure", "circumstance" or neither. Hence we deal with a multi-class classification problem. The extraction of the event components is similar to a NER problem.

Deep learning

Similar to the extraction of legal entities described in Section 4.2, we again use the state-of-the-art NLP library Flair[53], which uses contextualized string embeddings (called FlairEmbeddings) that capture the semantics and the context, and therefore, produce different context dependent embeddings for the same words [Akbik et al., 2018]. The pretrained transformer models (BERT, DistilBERT) are provided by the Huggingface library [Wolf et al., 2019] and can be easily imported into Flair. We use the pretrained state-of-the-art models as a baseline and compare it further with additional approaches selected upon their results on legal texts (cf. [Chalkidis et al., 2019, Shaheen et al., 2020, Tuggener et al., 2020]). As there is no pretrained legal model available, we apply the common approach of *fine-tuning* a Universal Language Model for Text Classification (ULMFiT) [Howard and Ruder, 2018], which takes a generic model and fine-tunes it with a domain-specific corpus (called transfer learning). The Flair ECHR model is created using the Flair library, and fine-tuning of the BERT and DistilBERT models is also based on the transformers library by Huggingface. In terms of preprocesing, we

[51] European Court of Human Rights, Yermakovich v. Russia, case no. 35237/14, 28 May 2019
[52] `https://mnavasloro.github.io/EventsMatter/`, last accessed 2021-01-31
[53] `https://github.com/flairNLP/flair`, last accessed 2021-01-31

remove very short sentences from the dataset, for instance headings such as "II THE LAW". The models are:

Flair and Flair fine-tuned. We use the pretrained generic *news-forward-fast* language model from the Flair embedding approach [Akbik et al., 2018], which is pretrained on an English corpus with one billion words ([Chelba et al., 2013]), as our baseline model. We also fine-tune the pretrained Flair model with the documents from our corpus for one epoch to obtain the Flair fine-tuned model. The process of fine-tuning the language model is very time-intensive and took more than seven hours.

Flair ECHR. There are no specific legal pretrained models available that we could use for our experiments. On a different classification task, we made good experiences in prior work with using a domain specific model trained on a small corpus of EU legal documents outperforming generic models in a multi-label text classification task (cf. Section 4.3). Therefore, we also train a model on a corpus of 13,000 ECHR court decisions acquired from the European Court of Human Rights OpenData project [Quemy, 2018] for four epochs.

BERT and BERT fine-tuned. The Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019] is a language model learning the context of words in a bidirectional way and is applicable to many tasks. We use a BERT model (*bert-base-cased*) pretrained on English Wikipedia and a book corpus consisting of around 11,000 unpublished books[54]. We fine-tune the pretrained model with the ECHR corpus for two epochs.

DistilBERT and DistilBERT fine-tuned. DistilBERT is a lightweight version of BERT that makes use of a teacher-student setup to distill the knowledge of the large model (BERT) to the student (DistilBERT) [Sanh et al., 2019]. We use the pretrained *distilbert-base-cased* model, which is based on the BERT base model. Additionally, we fine-tune the pretrained model for two epochs with our ECHR corpus to obtain the DistilBERT fine-tuned model.

Conditional Random Fields

Conditional Random Fields (CRF) are used for the mapping of sequences based on probabilistic models to label sequences [Lafferty et al., 2001]. CRF have already been applied in similar tasks in the legal domain for extracting specific legal entities, such as lawyers, courts and legal literature [Dozier et al., 2010, Cardellino et al., 2017b, Leitner et al., 2019]. A CRF model uses features of a token, for instance casing, position of the token and subsequences, to calculate the probability that it is preceded or followed by a particular other token. It also takes the probabilities into account that a specific named entity, for instance a temporal information, is followed by a subject. In contrast to the extraction of legal entities in Section 4.2, here we use CRF for the extraction of events and their components.

[54] `https://huggingface.co/bert-base-cased`, last accessed 2021-03-15

Rules

Unlike the previous approaches, implemented as a classification task, the rule-based approach is an annotation task based on a search for specific patterns of events in the form of frames. Our approach consists of two steps: 1. The collection of frames, which is done before the annotation; and 2. the event extraction that uses the frames in order to annotate a text.

1. Frame collection. We list all "what" event components in the training set, and then identify the main verb, its type and the dependency relations, for which we use the CoreNLP dependency parser [Chen and Manning, 2014], within the "what", and towards the subject ("who"), including the object for both possible active and passive voices since they are very different. When there are several mentions of the same main verb, all information is gathered and combined into a single frame. Once all the "what" elements are processed, they are stored for later use by the extraction algorithm.

2. Event extraction. Based the previously obtained frames, we look for the relevant events in the text. Since there are events that can appear many times in a text, we just consider events that have a date attached. In order to find dates and their normalized value (to be able to build a timeline), we adapt the Añotador software [Navas-Loro and Rodríguez-Doncel, 2020]. Then we use the information from the frames to search for the main verb of the event and for the previously identified dependency relations, as well as some Part-of-Speech considerations (using also CoreNLP). Additionally, some specific events that tend to appear always in the same form in the text, for instance the final decision, are identified using regular expressions.

5.2.3 Evaluation and Discussion

In this section, we present results of our experiments. All models have been trained with the same settings of a maximum of 150 epochs, patience of 3 and an anneal factor set to 0.5. The training is automatically stopped when the learning rate is too small. The results are evaluated using the metrics Precision (P), Recall (R) and F-score (F).

The documents have an average size of 2,302 tokens without the legal section (legal framework). Each document includes on average 21 different events, divided into 10 "procedure" and 11 "circumstance" events on average. The number of "who" occurrences amounts to 13.9 on average, while the number of temporal information annotations ("when") to 17.6, and the number of "what" annotations to 24. We split the dataset into training, testing and validation set on a document level applying 5-fold cross-validation (in the deep learning-based approach) such that the training set consists of 24, and the test and validation sets of three documents each. The results represent the average of all splits. The results for all approaches are presented in Table 5.10. When comparing different approaches on event (component) extraction, we can observe that more advanced language models based on the transformer architecture [Vaswani et al., 2017] (BERT and DistilBERT), in general, provide a better result compared to the standard embedding models (Flair). Furthermore, we can see that the application of the ULMFiT approach to fine-tune generic language models, with a domain-specific corpus, leads to improved results between less than 1% (Flair pretrained to Flair fine-tuned for "who") and 25%

Table 5.10: Evaluation results for event classification and event components

P=Precision, R=Recall, F=F-score. Best results highlighted in boldface.

		Event Types		Event Components		
		Procedure	Circumstance	What	When	Who
CRF	P	0.8239	0.6878	0.8510	0.8930	0.8909
	R	0.8026	0.4788	0.7691	0.8446	0.7038
	F	0.8080	0.5478	**0.8050**	0.8658	0.7834
Flair pretrained	P	0.8332	0.5721	0.5641	0.9050	0.8993
	R	0.7895	0.3264	0.4550	0.7965	0.7649
	F	0.8031	0.4057	0.5010	0.8435	0.8230
Flair fine-tuned	P	0.8707	0.5888	0.6012	0.9087	0.9163
	R	0.8157	0.5112	0.5179	0.8002	0.8371
	F	0.8413	0.5333	0.5558	0.8487	0.8744
Flair ECHR	P	0.7678	0.4193	0.5794	0.8200	0.4048
	R	0.7121	0.1312	0.1569	0.5788	0.1187
	F	0.7386	0.1792	0.2328	0.6688	0.1823
BERT pretrained	P	0.8195	0.6670	0.6045	0.8588	0.8637
	R	0.8079	0.4923	0.6117	0.8822	0.8990
	F	0.8056	0.5431	0.6078	0.8698	0.8805
BERT fine-tuned	P	0.9144	0.7681	0.6558	0.8945	0.8888
	R	0.9020	0.7894	0.6626	0.9101	0.9222
	F	0.9055	0.7759	0.6583	**0.9022**	**0.9044**
DistilBERT pretrained	P	0.8391	0.5653	0.5958	0.8187	0.8667
	R	0.8357	0.5163	0.5745	0.8635	0.8573
	F	0.8326	0.5326	0.5841	0.8395	0.8609
DistilBERT fine-tuned	P	0.9164	0.8161	0.6279	0.8731	0.8992
	R	0.9327	0.7865	0.6206	0.8933	0.9012
	F	**0.9238**	**0.7975**	0.6237	0.8823	0.8998

		Event				Event Components					
		Identification		Type		What		When		Who	
		Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict
Rules	P	0.8571	0.8000	0.4714	0.4286	0.8026	0.2368	0.7759	0.7241	0.7500	0.6875
	R	0.7792	0.7273	0.4286	0.3896	0.6932	0.2045	0.6338	0.5915	0.6316	0.5789
	F	0.8163	0.7619	0.4490	0.4082	0.7439	0.2195	0.6977	0.6512	0.6857	0.6286

(DistilBERT for "circumstance"). The average increase in performance with fine-tuning is 8% for recognizing "procedure" and 21% for "circumstance" events, resp. The results of the CRF approach for the "what" component is unexpected, as it outperforms the more advanced methods by approximately 20%. The results for the extraction of the event components show that recognizing temporal information ("when") of an event yields better results than the "what" of an event by 27% and the subject ("who") by 21% (averaged over all approaches). The performance increase for the extraction of the event components of fine-tuned models, compared to generic models, is with 5% ("what"), 3% ("when") and 4% ("who") lower compared to the results for event types.

We see that results of the event type detection are within approx. 20% over all approaches, with the worst result being achieved by the Flair ECHR approach (F 73.86%), and the best result is achieved by the DistilBERT fine-tuned approach with an F-score of 92.38%. The results for the "circumstance" event type show a bigger spread between the worst result of the Flair ECHR approach with an F-score of only 17.92%, while the best result is achieved by DistilBERT fine-tuned model (F 79.75%). For the "circumstance" event types we see generally lower results than for "procedure" type detection. We attribute this to the fact that the linguistic variety of the "procedure" events is narrower as they refer to a restricted set of ways of how to express them. The performance of the Flair ECHR model showed the least performance, due to being trained only on 13,000 ECHR documents, while it is common to train language models on much larger corpora to capture the basics of a language.

The performance differences between the "procedure" and "circumstance" event classes are evident with the latter results being worse by 29% on average. "Procedure" events capture formal processes throughout a legal trail and the ways to formulate the same events is somewhat restricted, for instance "the court upheld the judgment". In the description of the "circumstance"s of a case, however, the English language is potentially used in its entirety. Similarly, we observe the same behavior with the results for the event components with the results for "when" and "who" being better than the results for "what". We attribute this to the fact that absolute temporal information (e.g. a date) contained in the court decisions under investigation always follows the structure of "day month year", and the number of acting subjects is also limited to a certain range of persons (e.g. applicant, judge, prosecutor), authorities (e.g. Supreme Court, housing authority) or things (e.g. application, appeal). Relative temporal information (e.g. "X days later", "between X and Y" or "until X") is also expressed in a few ways only.

Overall, we can say that fine-tuning an existing language model trained on a large corpus, that captures the basic features of a language with a domain-specific corpus, performs better than training a new language model with a rather small domain-specific corpus. Moreover, the more restricted the variety of class candidates for classification is, the better the results. The same applies to the information adhering to a specific format, i.e. temporal information in the form of dates.

Regarding the rule-based approach, the evaluation is slightly different. In the deep learning approach (first table) the number of named entities reflect the results of finding the event arguments *only* in those sentences where there is an event. On the contrary, the rule-based approach (second table) finds the events and the arguments in the same algorithm, so the results of the argument are contingent upon the event results. Additionally, we provide both *strict* and the *lenient* results, meaning that either the extent of our annotation match exactly to the one by the annotators or that it only overlaps (adding or omitting some words), resp. Also, the event evaluation includes finding the extent of the event, and then, over this finding, decide its type. The annotation and evaluations for the rule-based approach were done with the software GATE [Cunningham et al., 2013].

From the results of the rule-based approach, we see that in the event finding task we got good results, both in the strict and lenient case, meaning that most of the events are correctly found and with the correct extent. Generally speaking, we identify about

4 out of 5 relevant events, and additionally some that were not marked as relevant (although this does not mean they are not events). Regarding event types, the results for rule-based approaches are not very promising, mainly due to the fact that the same verb can often represent both circumstantial or procedural events, depending on surrounding information that the current rule-based implementation is not able to identify.

Results for detecting event arguments with the rule-based approach, on the other hand, are very different. The "what" event component has very bad strict results, mainly due to the difficulty to determine the extent of the relevant modifiers of a verb. The "who" and the "when" show very good results, finding most of them correctly (e.g., 68.57% of the "who" taken into account that the limit was less than the 81.63% of the events) and almost always with the correct extent. The lenient results of the "what" component, similar to the ones from the other arguments, demonstrates that besides the extent, the identification is correct.

5.3 Related Work

Temporal tagging is a mature area of research that has been applied in different contexts, but scarcely in the legal domain. This section reviews several corpora with temporal annotations, along with the work previously done in temporal annotation of legal texts and in other domains.

The temporal information of a text document can be represented in structured, ad-hoc formats such as TIDES TIMEX2 [Ferro et al., 2005] or TimeML [Pustejovsky et al., 2003b]. TimeML is the ISO standard[55] for time and event markup and annotation. Other general-purpose annotation standards can also be used to represent TEs, such as the W3C Web Annotations[56] or the NLP Interchange Format[57] (NIF) [Hellmann et al., 2012]. TimeML uses TIMEX3 tags (modelled on previously mentioned TIMEX2) for marking TEs, and distinguishes between different `types` (namely, `DATE, DURATION, TIME` and `SET`, the latter being the type associated with sets of recurrent times). Other attributes in TIMEX3 tags allows for the expression of temporal information as a normalized value, for instance the actual date instead of relative expressions such as `yesterday`, following the ISO 8601 standard (`value`). TIMEX3 can also mark the presence of modifiers (`mod`) such as `END` or `LESS_THAN`, or specific information for each type, such as the frequency (`freq`) for `SET`.

Thus, for the analysis of temporal expressions, the following three domains received the most attention: medical texts (e.g. the THYME corpus [Styler IV et al., 2014]), news (e.g. the Timebank corpus [Pustejovsky et al., 2003a] and the MEANTIME corpus [Minard et al., 2016]) and historical documents (e.g. the Wikiwars corpus [Mazur et al., 2010]). Corpora have also included texts in different language registers, such as tweets [Tabassum et al., 2016], colloquial texts [Strötgen et al., 2012] or scientific abstracts [Strötgen et al., 2012]. However, to the best of the authors' knowledge, there are no temporally annotated legal corpora publicly available that relate to English language court decisions. Although annotation challenges (both in general and also in different

[55] ISO 24617-1 Language Resource Management - Semantic Annotation Framework (SemAF) - Time and Events (SemAF Time and ISO-TimeML)

[56] `https://www.w3.org/TR/annotation-model/`, last accessed 2021-01-31

[57] `http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core`, last accessed 2021-01-31

specific domains) have been identified in literature [Ji et al., 2014, Strötgen et al., 2012, Styler IV et al., 2014], very little work has been conducted in connection with the legal domain. A description of the different approaches adopted by existing temporal taggers, including the identification of several state-of-the-art temporal taggers, can be found in Section 5.1.5.

In the legal domain, previous research work by [Schilder, 2005] already pointed out the relevance of the temporal dimension of information in legal documents. In this work, an analysis of the different types of legal documents and the temporal information that can be found in the legal documents was outlined. Schilder distinguished between dates in transactional documents (namely, documents written by lawyers for specific transactions, such as contracts or agreements), constraints in statutes or regulations, and legal narratives in case law. While the first two types of documents received dedicated attention, narratives in case law were assimilated to narratives present in news. An alternative approach proposed by [Isemann et al., 2013], used both Named Entity Recognition (NER) and temporal processing to extract temporal dependencies from regulations with no narrative-structure. The authors also described some of the recurrent pitfalls temporal taggers have to deal with, such as the confusion between legal references (e.g. "Directive *2009*/28/EC") and actual dates or the distinction between *episodic* and *generic* events. The former referring to a specific moment (e.g. "the rescission of the contract was done on 7 December 2017") and the latter referring to an event in general truths, laws, rules or expectations (e.g. "Every rescission implies the following actions"). A finding that we can confirm in our work, for instance shown in Table 5.2 for mistakenly tagged legal references. Other approaches in the legal domain include works on transactional documents by [Naik et al., 2011], where a first framework for dealing with temporal information in that kind of texts is proposed. Also additional efforts focused on reasoning with legal evidence (burden of proof) and coherence of narratives (e.g. plausibility and completeness) were made [Vlek et al., 2013], using temporal information but without extracting it from scratch.

Works in other fields, such as the medical domain, are also of interest since they share common requirements. They also confirm the need of domain knowledge for identifying specific events[58] and for dealing with the existence of several timelines in the same text, among others. The analysis by [Styler IV et al., 2014] in the clinical domain identifies the need of specific guidelines for temporal annotation, which require domain-specific temporal knowledge and the definition of general phases in clinical processes (some kind of commonsense domain knowledge). Furthermore, new tags not included in the temporal annotation standard TimeML, commonsense information and events are defined in the same work, along with annotation needs and different timelines (such as discussions with other colleagues and notes about risks in treatment) were redesigned for fitting the medical particularities. We work under the assumption that most of these considerations and challenges can also be present in a similar form in legal documents, requiring therefore a dedicated approach.We conclude that one of the primary limitations of existing work is the fact that no special consideration has to date been given to both the narrative structure and the particularities of the legal domain (see Section 5.1.1 for additional details).

[58]For instance, *diagnosis* such as *tumors* or medical tests are relevant events that should appear in a timeline of a medical doctor, as stated by Styler et al. [Styler IV et al., 2014], but not in other types of texts. Similarly, specific legal events such as *preliminary rulings* (explained in Section 5.1.1) in European judgments are always relevant to lawyers, although they never appear in other kinds of texts.

Recent advances in NLP are often based on embedding text in multidimensional vector space, with neural network architectures being trained on such numeric representations. Such methods yield in re-usable, publicly available language models trained on large corpora of texts, where embeddings can be created on different levels, for instance words, sentences and documents. While pretraining models on large corpora of generic texts is a very time-consuming process [Howard and Ruder, 2018], adapting (aka fine-tuning) such generic models to domain-specific language is often less demanding.

Overviews on diverse automated event extraction approaches in the general domain can be found in literature [Hogenboom et al., 2011, Xiang and Wang, 2019]. Specifically in the legal domain [María Navas-Loro, 2018], existing work usually involves searching for *ad hoc* definitions of events, ignoring general event annotation schemas such as the ACE 2005 model [ACE, 2005]. Several approaches tend to be supported by patterns and use manually crafted rules or semantic role labeling techniques [Kiyavitskaya et al., 2008, Maxwell et al., 2009, Lagos et al., 2010, María Navas-Loro, 2018]. Other approaches do not search for events specifically, but target legal case factors, which are descriptions of facts that occur in many court decisions [Wyner and Peters, 2010].

The automated generation of timelines out of annotated documents could help to get a better and faster understanding of the content of a document. Existing work focusing on this task include Linea [Etiene et al., 2015], a system that is able to build and navigate timelines from unstructured text, and TimeLineCurator [Fulda et al., 2016], which is a system that is primarily designed to allow journalists to generate temporal stories and can be used to produce a timeline from any free text or URL. Furthermore, the creation of timelines has also been investigated in other domains, such as medicine [Styler IV et al., 2014, Jung et al., 2011] and journalism [Tannier and Vernier, 2016]. We refer to [Fulda et al., 2016] for further details on the respective approaches.

5.4 Summary and Future Directions

In this chapter, we focused on temporal information contained in court decisions and the approaches to extract temporal information, which can be put into the context of legal events.

In Section 5.1, we have analyzed a corpus of court decisions with regard to temporal information. Furthermore, we investigated how the particularities of court decisions as well as legal English influence the extraction of temporal expressions from court decisions of the European Court of Justice, the European Court of Human Rights and the United States Supreme Court. This analysis showed that court decisions follow a certain structure that vary from court to court but is consistent for all decisions from the same court. Moreover, the analysis showed that court decisions contain text sequences as parts of case references or domain-specific abbreviations being picked up by temporal taggers by mistake. The TimeML standard to represent temporal information in a structured format turned out to be incomplete to cover all kinds of temporal expressions used in the legal domain, for instance uncertain durations or to indicate legal domain specific points in time. Furthermore, we introduced a categorization of three temporal dimensions. As to the best of our knowledge, there was no gold standard legal domain temporal corpus available. Hence, we created a new gold standard corpus consisting of thirty court decisions from three different courts with manually annotated temporal expressions. The proposed corpus has been annotated with generic (*StandardTimeML*) and legal domain specific (*LegalTimeML*) annotations which have been considered to be important for the facts of the case. We used this new corpus to compare the performance of ten non-domain specific state-of-the-art temporal taggers. They achieved good results in terms of detecting all possible temporal expressions but the performance decreases when looking on legal domain specific temporal expressions only. The results to the experiment also show that there is no temporal tagger outperforming all others compared across the decisions from the three courts being a result from the the different formats of how temporal information is expressed by the individual courts.

We see that future work towards the extraction of temporal information from legal documents is necessary. The experiments conducted on English documents only, already show that the different formats of temporal expressions lead to a very different performance of the compared temporal taggers. The lack of available temporal taggers has already been picked up and a domain-specific temporal tagger “Añotador” capable of processing Spanish and English legal texts [Navas-Loro and Rodríguez-Doncel, 2020] has been proposed. Still, we see the need to extend research area and also include less structured court decisions and multiple languages, which will presumably lead to the creation of language or even court specific temporal taggers.

In Section 5.2, we investigated the possibility of extracting legal events from court decisions, which builds on the work about temporal information. We manually annotated a corpus of thirty court decisions from the European Court of Human Rights resulting in a new gold standard corpus. We introduced the notion of legal events and described their components, in particular we fragment events into the temporal information, the subject and the object. We showed that this fragmentation serves as the basis for the creation of timelines, which could be used to get a quick overview of the events contained in a case in a consecutive order. Furthermore, we conducted a comparison of different approaches to extract events and their components using rules, conditional random fields and deep learning-based approaches with pretrained and fine-tuned language models. The evaluation of the different approaches showed that there is not a single approach performing best for the event classification task and the event component extraction rather that the best approach should be chosen on the task to be carried out. Moreover, classic rule-based approaches also provide a good performance. Furthermore, we compared language models pretrained on generic texts such as Wikipedia and news articles with generic language models fine-tuned with legal documents. The results prove that fine-tuning on domain-specifics text increases the performance.

While we showed that the extraction of events from court decisions is possible, it would be interesting to investigate how these approaches perform on less structured court decisions from other courts and other languages. With respect to an integration of temporal expressions and legal events into the legal knowledge graph, we see two possible integration strategies, which would be worthwhile investigating. The first strategy aims at implementing an annotation system that automatically annotates a document with legal events “on the fly” on user request. The legal events are not stored and only generated when requested. The second strategy is to integrate legal events into the legal knowledge graph be making them persistent using an extension of the legal knowledge graph with additional classes and properties to cater for the representation of legal events. A future research direction along this line would include an analysis of existing ontologies for the legal domain as well as ontologies covering temporal aspects, which could be used for this purpose. However, an annotation of RIS documents was not performed due to the lack of resources in terms of human annotators for experiments with RIS documents caused by the end of the project with the Austrian ministry. Finally, it would be interesting to investigate how the event extraction could be used for an automatic compliance check with respect to the abidance by the law.

CHAPTER 6

Legal Knowledge Graph Integration

Our goal is to integrate the Austrian legal knowledge graph with legal knowledge graphs from other countries, which is also in-line with the objectives laid out in the ELI and ECLI proposals. It is therefore necessary that other countries also provide legal data at first and, ideally, also participate in the ELI and ECLI initiatives enabling others to use their legal data in an automated fashion. In this chapter, we provide a comparative analysis of the current situation regarding the provision of linked legal data across the EU member states. We focus on the used ontologies and features of the available legal information systems provided by the governments of these countries in Section 6.1. Besides the governmental initiatives, there are also non-governmental initiatives based on ELI and ECLI with the goal to ease access to legal information, for example by extracting information from or interlinking of legal documents. The efforts of these non-governmental initiatives are described in Section 6.2. In Section 6.3, we analyze and demonstrate the benefits of an integrated legal knowledge graph including legal data from Austria as well as other EU member states and answer the sample questions described in Section 3.1. We provide a roadmap towards a linked legal knowledge graph summarizing the challenges we came across during the creation process of the Austrian legal knowledge graph. Finally, Section 6.5 summarizes the chapter and includes a view on possible future research directions.

6.1 Governmental Initiatives

We include the EU member states without the United Kingdom and EU candidate countries in our analysis of whether and how they make legal information available in a machine-readable form. We use the EU e-Justice portal[1] as a starting point for our research process, which includes overview pages on which EU member states can provide additional information about their implementation. These information pages are available for all EU member states for ELI[2] and ECLI[3]. While the country-specific ECLI information page contains all EU member states, the ELI information page only

[1]`https://e-justice.europa.eu/`, last accessed 2021-03-20
[2]`https://eur-lex.europa.eu/eli-register/implementation.html`, last accessed 2021-03-20
[3]`https://e-justice.europa.eu/content_european_case_law_identifier_ecli-175-en.do?init=true`, last accessed 2021-03-20

has information for 17 countries. Typically, an explanation of the current state and examples are included as well as links to national legal databases. Some countries provide detailed information about their deployed ELI/ ECLI structure, while others do not provide any information or, respectively, only in the national language, which needs to be translated using a translation service. When available, we followed the links provided, otherwise we used a search engine to manually find additional national legal databases and examples for legislative and judiciary documents (cf. Tables A.1 and A.2 (Appendix A.1) for links to databases and examples). In the first step, we examine whether ELI/ECLI identifiers are visible in the document. In the second step, we also scan the source code of the (HTML) document and search in the metadata for keywords such as “eli”, “ontology”, “dc”, “dcterms”, “creator” and “date”. Where we find metadata embedded in the document, we parse the URL of this document with EasyRdf[4] to automatically retrieve RDF triples per document. We also check whether countries use national Named Authority Lists (NALs), i.e. determine whether national information pages about the used NAL are provided. In addition to this search process on the national level, we also query the EU Open Data Portal[5] for national legal data. We also record the type of available search interfaces, available document formats, languages and availability of judiciary documents in the EU ECLI search engine per country.

Table 6.1 provides a comprehensive overview of the national ELI and ECLI implementation initiatives of the EU member states with a focus on the ELI/ ECLI implementation status. The columns “Implementation ELI” and “Implementation ECLI” describe the implementation status. The keyword “Identifier” refers to the situation where documents are given an ELI identifier and “Identifier/Metadata” indicates that the particular country also provides metadata for their documents. The general assumption is that all countries use the ELI ontology for legislative documents and ECLI for judiciary documents respectively, but some countries provide national extensions in order to represent legal information based on national requirements. These additional ontology extensions are indicated in brackets, for instance Finland defined its own extensions for ELI in the *Semantic Finlex Legislation Ontology (SFL)*[6] and in the *Semantic Finlex Case Law Ontology (SFCL)*[7] for judiciary documents. Luxembourg also uses an additional ontology called *JOLUX*[8] in their *Casemates* project[9] incorporating and extending the ELI ontology. Special cases are Latvia and Slovenia who do not participate in the ELI and therefore also do not assign ELI identifiers to their legislative documents. However, they do provide a basic set of metadata, which is less than and different to ELI, based on the *Open Graph Protocol (OGP)*[10]. Portugal assigns an ECLI identifier to judiciary documents, but uses OGP for the metadata. The Netherlands use the *dcterms* and *Overheid* ontologies for their legislative documents. We can see that 11 out of 27 countries implemented at least the first pillar of the ELI ontology, i.e. assigning an ELI identifier to the documents. Participation/Implementation is better in terms of ECLI, where 19 countries assign an ECLI identifier to judiciary documents, but the number of countries providing machine-readable metadata (i.e. 3) is lower compared to ELI (i.e.

[4] `http://www.easyrdf.org/`, last accessed 2021-03-20
[5] `https://data.europa.eu/euodp/en/home`, last accessed 2021-03-20
[6] `http://data.finlex.fi/schema/sfl/`, last accessed 2021-03-20
[7] `http://data.finlex.fi/schema/sfcl/`, last accessed 2021-03-20
[8] `https://data.public.lu/en/datasets/legilux-journal-officiel-du-grand-duche-de-luxembourg/`, last accessed 2021-03-20
[9] `http://www.legilux.lu/editorial/casemates`, last accessed 2021-03-20
[10] `https://ogp.me/`, last accessed 2021-03-20

Table 6.1: Linked legal data feature comparison of EU member states

(+) indicates the usage of additional or other ontologies.

Country	**Implementation ELI**	**Implementation ECLI**	**Data Availability**	**Information ELI / ECLI / NAL**	**Thesaurus**
Austria	Identifier	Identifier	-	✓/ - / -	✓
Belgium	Identifier	Identifier	-	✓/ ✓/ -	-
Bulgaria	-	Identifier	-	- / ✓/ -	-
Croatia	Identifier	Identifier	-	✓/ - / -	-
Cyprus	-	-	-	- / ✓/ -	-
Czech Republic	-	Identifier	-	- / ✓/ -	-
Denmark	Identifier/Metadata	-	RDF	✓/ ✓/ ✓	-
Estonia	-	Identifier	-	- / ✓/ -	✓
Finland	Identifier/Metadata (+)	Identifier/Metadata (+)	RDF	✓/ ✓/ ✓	✓
France	Identifier/Metadata	Identifier	RDFa	✓/ ✓/ -	-
Germany	-	Identifier/Metadata	-	- / ✓/ -	-
Greece	-	Identifier	-	-	-
Hungary	-	-	-	✓/ - / -	-
Ireland	Identifier/Metadata	-	RDFa, RDF	✓/ ✓/ -	-
Italy	Identifier/Metadata	Identifier	RDFa, RDF	✓/ ✓/ ✓	-
Latvia	- (+)	Identifier	-	- / ✓/ -	-
Lithuania	-	-	-	- / ✓/ -	✓
Luxembourg	Identifier/Metadata (+)	-	RDFa	✓/ - / ✓	-
Malta	-	Identifier	-	✓/ - / -	-
Netherlands	- (+)	Identifier/Metadata	RDFa, RDF	- / ✓/ -	-
Poland	-	-	-	-	-
Portugal	Identifier/Metadata	Identifier (+)	RDFa	✓/ - / -	-
Romania	-	Identifier	-	- / ✓/ -	-
Slovakia	-	Identifier	-	- / ✓/ -	✓
Slovenia	- (+)	Identifier	-	- / ✓/ -	-
Spain	Identifier/Metadata	Identifier	RDFa	✓/ - / ✓	-
Sweden	-	-	-	-	-

9). Compared to a study conducted in 2017 [van Opijnen et al., 2017b], the participation in ECLI increased in the last years. Now, seven additional countries participate in ECLI and assign at least an ECLI identifier to their judiciary documents. The column "Data Availability" describes how the data is provided to the public. As shown, the majority of participating countries opts to use the RDFa format and embed the metadata in the source code of the document. Denmark, Finland, Ireland and Italy also allow users to download the data in RDF, either from a national website or the European Open Data Portal. The Netherlands provide a web service[11] that can be used to download the data in RDF. In the column "Information ELI/ECLI/NAL", we indicate whether information about the national implementation of ELI and ECLI as well as the usage of NAL is provided. This information can be provided either using dedicated pages on the EU e-Justice portal or a national website. A thesaurus, such as EuroVoc or a national index of legal terms, is used by five countries as indicated in the column "Thesaurus".

In more detail, Table 6.2 contains an overview of the used ELI properties for the EU member states, which use ELI for their legal documents. We also include

[11] `https://linkeddata.overheid.nl/front/portal/services`, last accessed 2021-03-20

Table 6.2: Overview of used ELI properties of countries providing metadata using ELI including non-governmental initiatives

This table also includes the properties used by non-governmental initiatives (see Section 6.2), which are highlighted gray. Mandatory properties are highlighted in boldface.

ELI Property	**Austria**	**Denmark**	**Finland**	**France**	**Greece**	**Ireland**	**Italy**	**Luxemburg**	**Portugal**	**Spain**
			Finlex		*Nomothesia*					
Data based on	*LKG*	*RDF*	*SPARQL Endpoint*	*RDFa*	*SPARQL Endpoint*	*RDF*	*RDFa, RDF*	*RDF*	*RDFa*	*RDFa*
eli:amended_by	✓		✓							
eli:amends	✓		✓		✓			✓		
eli:applied_by										
eli:applies										
eli:based_on		✓				✓		✓		
eli:basis_for		✓								
eli:changed_by	✓	✓	✓							
eli:changes	✓	✓	✓		✓	✓				
eli:cited_by			✓							
eli:cited_by_case_law	✓									
eli:cited_by_case_law_reference										
eli:cites			✓	✓	✓			✓	✓	
eli:commenced_by										
eli:commences										
eli:consolidated_by		✓								✓
eli:consolidates	✓	✓						✓		
eli:corrected_by										✓
eli:corrects								✓		
eli:date_applicability								✓		
eli:date_document	✓	✓		✓		✓	✓	✓		✓
eli:date_no_longer_in_force	✓	✓	✓					✓		
eli:date_publication	✓		✓	✓	✓		✓	✓	✓	✓
eli:description						✓		✓	✓	
eli:embodies	✓	✓	✓	✓		✓		✓	✓	✓
eli:first_date_entry_in_force	✓		✓		✓			✓		
eli:format	✓	✓	✓	✓		✓	✓	✓		✓
eli:has_another_publication										
eli:has_member	✓		✓							✓
eli:has_part			✓		✓	✓				
eli:has_translation										
eli:id_local	✓	✓	✓	✓	✓		✓	✓	✓	✓
eli:implemented_by										
eli:implements										
eli:in_force	✓	✓	✓						✓	
eli:is_about	✓		✓						✓	
eli:is_another_publication_of										
eli:is_embodied_by	✓	✓	✓	✓		✓	✓	✓		✓
eli:is_exemplified_by										
eli:is_member_of	✓		✓					✓		✓
eli:is_part_of			✓	✓				✓		
eli:is_realized_by	✓	✓	✓			✓	✓		✓	✓
eli:is_translation_of										
eli:jurisdiction	✓									✓
eli:language	✓	✓	✓	✓		✓	✓	✓	✓	✓
eli:legal_value		✓		✓		✓			✓	
eli:licence						✓		✓	✓	
eli:media_type										
eli:number	✓					✓			✓	✓
eli:passed_by		✓	✓	✓			✓			
eli:published_in		✓		✓	✓	✓				
eli:published_in_format						✓		✓	✓	
eli:publisher				✓		✓	✓	✓	✓	✓
eli:publisher_agent							✓		✓	
eli:publishes										
eli:realized_by	✓				✓					
eli:realizes	✓	✓	✓	✓		✓	✓	✓	✓	✓
eli:related_to			✓			✓				
eli:relevant_for	✓	✓				✓				
eli:repealed_by	✓		✓							
eli:repeals	✓		✓					✓		
eli:responsibility_of				✓				✓	✓	
eli:responsibility_of_agent								✓	✓	
eli:rights								✓		
eli:rightsholder						✓				
eli:rightsholder_agent								✓	✓	
eli:title	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
eli:title_alternative	✓	✓	✓					✓		
eli:title_short	✓	✓								
eli:transposed_by	✓				✓					
eli:transposes	✓	✓	✓			✓	✓	✓		
eli:type_document	✓	✓	✓	✓		✓	✓	✓	✓	✓
eli:uri_schema		✓				✓	✓		✓	
eli:version	✓		✓			✓	✓			✓
eli:version_date			✓							✓

Table 6.3: Overview of used ECLI properties of countries providing metadata using ECLI

This table also includes the properties used by non-governmental initiatives (see Section 6.2), which are highlighted gray. Mandatory properties are highlighted in boldface.

ECLI Property	Austria	Finland	Germany	Netherlands
Data based on	*LKG*	*Finlex SPARQL Endpoint*	*RDFa*	*RDFa*
dcterms:abstract		✓		✓
dcterms:accessRights	✓		✓	
dcterms:contributor	✓	✓		
dcterms:coverage	✓		✓	
dcterms:creator	✓	✓	✓	✓
dcterms:date	✓	✓	✓	
dcterms:description		✓		
dcterms:identifier	✓		✓	✓
dcterms:isReplacedBy				
dcterms:issued		✓		✓
dcterms:isVersionOf	✓	✓	✓	
dcterms:language	✓	✓	✓	✓
dcterms:publisher	✓	✓	✓	✓
dcterms:references	✓			
dcterms:subject	✓			✓
dcterms:title				✓
dcterms:type	✓		✓	✓

the non-governmental initiatives (highlighted in gray) for Finland and Greece. We can see that the usage of the ELI properties to represent national legislative documents differs from country to country, which is not a problem as the majority of the ELI properties is optional. However, there are six mandatory ELI properties (highlighted in boldface), which are generally used with a few exceptions. There are some properties, for instance `eli:applies` (`eli:applied_by`) and its predecessor `eli:implements` (`eli:implemented_by`), `eli:commences` (`eli:commenced_by`) and `eli:has_translation` (`eli:is_translation_of`), which are currently not used by any ELI participant. However, they are available in case other EU member states implementing ELI decide to use them. The property `eli:cited_by_case_law` is only used by our legal knowledge graph to establish a link from a legislative to a judiciary documents.

Table 6.3 provides an overview of the used ECLI properties of four countries, of which two are from non-governmental initiatives (Austria and Finland). Similarly to the ELI properties, we can see that the usage of the individual properties is dependent on the national requirements but also that not all implementations also use the nine mandatory ECLI properties. Three properties stand out as they are currently not used at all (`dcterms:isReplacedBy`) or are only used by a single country (`dcterms:description` and `dcterms:references`). However, it needs to be taken into account that both tables for the ELI properties (Table 6.2) and ECLI properties (Table 6.3) only take the ELI and ECLI properties into account. That is why it might look like that the Austrian legal knowledge graph is the only one interlinking legislative and judiciary documents. As shown in Table 6.1, other countries might use additional ontologies,

Table 6.4: Overview of the used NAL in different countries for legislative and judiciary documents

This table also includes the NAL used by non-governmental initiatives (see Section 6.2), which are highlighted gray.

NAL for property	**Austria**	**Denmark**	**Finland**	**France**	**Italy**	**Luxembourg**	**Netherlands**	**Portugal**	**Spain**
Data based on	*LKG*	*RDF*	*Finlex SPARQL Endpoint*	*RDFa*	*RDF*	*RDF*	*RDFa*	*RDFa*	*RDFa*
`dcerms:type`	✓	-	-	-	-	-	✓	-	-
`dcterms:subject`	✓	-	-	-	-	-	✓	-	-
`eli:is_about`	✓	-	-	-	-	✓	-	-	-
`eli:jurisdiction`	✓	-	-	-	-	-	-	-	✓
`eli:language`	✓	-	-	✓	✓	✓	-	-	✓
`eli:passed_by`	-	✓	✓	-	-	-	-	-	-
`eli:publisher_agent`	-	-	-	-	-	✓	-	✓	-
`eli:relevant_for`	✓	✓	-	-	-	-	-	-	-
`eli:responsibility_of_agent`	-	-	-	-	-	✓	-	✓	-
`eli:rightsholder_agent`	-	-	-	-	-	✓	-	✓	-
`eli:type_document`	✓	✓	-	-	✓	✓	-	✓	✓
`eli:version`	-	-	-	-	✓	-	-	-	✓

which could include this information. For example, the finish SFCL ontology uses `sfcl:refToLegislation` instead of `dcterms:references` to link judiciary with legislative documents.

Table 6.4 provides an overview over the used NALs for selected countries. We notice that there are more countries using NALs, however they do not all provide an information page. Some properties in the legal knowledge graph are very suitable to be used with the already available NALs provided by the EU, for instance `eli:language`, `eli:type_document` and `eli:jurisdiction`. However, as these NALs are provided on an EU level they are directed to EU institutions and might not be directly applicable in individual member states. For instance, the items in a NAL might not be appropriate to represent national requirements such that a national list needs to be created.

We show the features of the EU member states' legal databases in Table 6.5. Central search interfaces are very convenient as users can find all the required information in the same place. However, as legal systems are typically divided into a legislation and judiciary branch, the information for both branches falls under the responsibility of different authorities and therefore might be provided at distinct places. The column "Central Interface" shows if there is a central interface available that enables users to access legislation as well as judiciary documents from different authorities, even if they are stored in separated backend systems. The EU e-Justice portal contains an ECLI search engine[12], which enables users to search for ECLI identifiers and keywords in judiciary documents from multiple countries. However, not all countries assigning an ECLI identifier are also participating in the ECLI search engine. The column "Search Interface" indicates how the search process can be performed by users. As we can see, the majority of countries provides a keyword-based search interface, which might be enhanced with additional filters, for instance to restrict dates to a certain time frame or select only special types of documents. Faceted search interfaces are implemented by a minority of countries only, "Both" means that one legal database provides a keyword-based search and the other legal database supports faceted search. We can also see that Finland and

[12] `https://e-justice.europa.eu/content_ecli_search_engine-430-en.do`, last accessed 2021-03-20

Table 6.5: Features of legal databases of EU member states

** denotes a subset.*

Country	Central Interface	ECLI Search	Search Interface	Document Format	Languages
Austria	✓	-	Keyword	HTML, PDF, RTF, XML	DE, EN*
Belgium	-	✓	Keyword	HTML	FR, NL, DE
Bulgaria		✓	Keyword	HTML, PDF	BG
Croatia	-	✓	Keyword	HTML	HR
Cyprus	✓	-	Keyword	PDF	EL
Czech Republic	-	✓	Keyword	PDF	CZ
Denmark	-	-	Faceted	HTML, PDF	DK
Estonia	✓	✓	Keyword	HTML, PDF, TXT, XML	EE, EN*
Finland	✓	✓	Keyword, SPARQL	HTML	FI, SE
France	✓	✓	Keyword	HTML, PDF	FR, EN*, DE*, IT*, ES*
Germany	-	✓	Keyword	HTML	DE, EN*
Greece	-	✓	Keyword	PDF	EL
Hungary	-	-	Keyword	HTML	HU, EN*
Ireland	-	-	Keyword	HTML, PDF	EN
Italy	-	✓	Keyword	HTML	IT
Latvia	-	✓	Keyword	HTML, PDF	LV, EN*, RU*
Lithuania	-	-	Faceted	HTML, PDF	LT
Luxembourg	-	-	Faceted, SPARQL	HTML, PDF, XML, RDF	FR
Malta	-	-	Keyword	PDF	MT, EN
Netherlands	-	✓	Both	HTML, PDF, RDF	NL, FR, EN*
Poland	-	-	Keyword	PDF	PL
Portugal	✓	✓	Faceted	HTML, PDF	PT, EN*
Romania	-	-	Keyword	HTML	RO
Slovakia	✓	-	Keyword	HTML, PDF	SK
Slovenia	-	✓	Keyword	HTML, PDF, DOCX	SI, EN*
Spain	-	✓	Both	HTML, PDF, XML, EPUB	ES
Sweden	-	-	Keyword	HTML	SE

Luxembourg provide a publicly accessible SPARQL endpoint, which allows users to run structured queries on the data directly. The standard and most commonly used way to represent legal documents on the web is HTML as shown in column "Document Format". Even though the content is displayed using HTML, the majority of legal information systems also allow users to download documents in PDF format. However, some countries provide documents in PDF only. A popular structured format is XML, supported by Austria, Estonia, Luxembourg and Spain. The EPUB format, a format for the distribution of digital publications and documents [W3C Community Group, 2021], is only used in Spain. As indicated in the column "Languages", it is obvious that countries provide their documents in their official language(s). What is more, Austria, Estonia, France, Germany, Hungary, Latvia, Netherlands, and Slovenia also publish a subset of their documents in additional languages, mainly English. These are typically documents, which are considered to be of high importance in a legal system, like the constitution or the civil code.

Table 6.6: Non-governmental initiatives using ELI and ECLI.

Project	Type	Using ELI / ECLI	Extension ELI / ECLI	Data Availability	Thesaurus	Open Data Linking	SPARQL
Legal Knowledge Graph	Linking	✓/✓	LKG / LKG	RDF	EuroVoc, Other	✓	✓
Semantic Finlex	Linking	✓/✓	SFL / SFCL	RDF	EuroVoc, Other	✓	✓
Nomothesia	Linking	✓/ -	Nomothesia / -	RDF	-	✓	✓
EUCases	Linking	✓/✓	-	-	EuroVoc, Other	-	-
Lynx	Linking	✓/ -	Lynx-LKG	RDF	EuroVoc, Other	✓	✓
GDPRtEXT	Linking	✓/ -	GDPRtEXT	RDF	-	-	✓
Linkoln	Extraction	✓/ -	-	-	-	-	-
BO-ECLI	Extraction	- /✓	-	-	-	-	-

6.2 Non-Governmental Initiatives

Besides linked legal data initiatives driven by governments, there are also efforts by academia and industry in this direction, which are often conducted in collaboration with and funded by governments. We are particularly interested in non-governmental initiatives working with ELI and ECLI providing a linked legal data framework or focusing on special legal areas.

Table 6.6 shows an overview of several non-governmental initiatives across Europe based on the information provided by the project websites, publications or namespaces used in RDF data retrieved via a SPARQL endpoint. The column "Project" shows the title of the project. We classify the projects, as indicated in column "Type", into the classes "linking", which means that this project aims to link legal data with other legal other data or external knowledge bases, and "extraction", which means that the project focuses on the extraction of specific information contained in legal documents. The column "Using ELI / ECLI" indicates whether a project uses ELI, ECLI or both. In cases where the project results in extensions to the ELI and ECLI ontologies, the name of these extensions is listed in column "Extension ELI / ECLI". In cases where data is made available for download, the format is shown in column "Data Availability". The column "Thesaurus" indicates whether the European thesaurus EuroVoc or other thesauri (e.g. a national thesaurus) is used. When the data used in the project is linked with other external data, such as DBpedia or Geonames, this is indicated in the column "Open Data Linking". The column "SPARQL" shows whether a SPARQL endpoint is available to retrieve the data from that project.

The *Legal Knowledge Graph* project that aims to integrate legal data from disparate legal databases into a knowledge graph is described in in this thesis. The *Semantic Finlex Project*[13] [Oksanen et al., 2019] carried out by the University of Aalto is, similar to our Austrian research project, based on the national legal database of Finland, which contains legislative and judiciary documents and transforms the data into linked legal data based on the ELI and ECLI ontologies. The results of this Finnish project are also visible in Table 6.1 as they are available to the public via the official Finlex website[14], as well as via a SPARQL endpoint[15]. Finlex extends the ELI with the Semantic Finlex

[13] `https://seco.cs.aalto.fi/projects/lawlod/`, last accessed 2021-03-20
[14] `https://data.finlex.fi/`, last accessed 2021-03-20
[15] `https://www.ldf.fi/sparql-services.html`, last accessed 2021-03-20

Legislation Ontology[16] (SFL) and ECLI with the Semantic Finlex Case Law ontology[17] (SFCL). The greek project *Nomothesia* [Chalkidis et al., 2017] by the University of Athens focuses on legislation only. It uses legal documents published in PDF format, which are transformed into linked legal data based on the ELI, which is incorporated in the Nomothesia ontology[18]. The data produced by the Nomothesia project is available for download as well as via a SPARQL endpoint[19] and includes DBpedia as an external knowledge base, for instance to link persons that are mentioned in legal acts. In the *EUCases* project [Boella et al., 2015], a first effort effort was made trying to link national and EU legislation and case law. Unfortunately, this project is no longer accessible because a login is required and there is no response to email requests[20]. This project also includes a proposal to link legal documents with the EuroVoc thesaurus and incorporates the Legal Taxonomy Syllabus (LTS) [Ajani et al., 2007]. The EU funded *Lynx* project[21] aims at creating a legal knowledge graph with a special focus on compliance [Montiel-Ponsoda et al., 2017]. This project includes Spanish legislation and jurisdiction as well as documents from selected countries and extends ELI and ECLI with the Lynx-LKG ontology[22]. The Lynx data can also be accessed via a SPARQL endpoint[23]. A legal domain-specific work is *GDPRtEXT*[24], extending the ELI to provide the General Data Protection Regulation (GDPR)[25] as a linked data resource together with a taxonomy of GDPR terms using SKOS [Pandit et al., 2018]. The linked legal data version of the GDPR extends the ELI ontology with the GDPRtEXT ontology. The data and the ontology are available for download[26] and can be accessed via a SPARQL endpoint[27]. The Italian *Linkoln project* focuses on the automatic extraction of references from legal documents of the Italian Senate and is also able to extract ELI references [Bacci et al., 2019]. The EU funded *BO-ECLI*[28] project focused on the ECLI and investigated the implementation of the ECLI in selected countries, which resulted in a proposal of a new version of the ECLI due to discovered drawbacks [van Opijnen et al., 2017a].

6.3 Benefits of an Integrated Knowledge Graph

We can revisit the example questions raised in Section 3.1 again and demonstrate the benefits of an integrated legal knowledge graph by underpinning them with example SPARQL queries providing answers to such questions.

Q 1 *Which documents are referenced in a specific court decision?*

Court decisions are based on the law and therefore reference legal provisions but also other court decisions and legal rulings. Users nowadays typically need to query the

[16] `http://data.finlex.fi/schema/sfl/`, last accessed 2021-03-20
[17] `http://data.finlex.fi/schema/sfcl/`, last accessed 2021-03-20
[18] `http://legislation.di.uoa.gr/data/ontology`, last accessed 2021-03-20
[19] `http://legislation.di.uoa.gr/endpoint`, last accessed 2021-03-20
[20] `http://www.eucases.eu`, last accessed 2021-03-20
[21] `http://www.lynx-project.eu/`, last accessed 2021-03-20
[22] `http://lynx-project.eu/doc/lkg/`, last accessed 2021-03-20
[23] `http://sparql.lynx-project.eu/`, last accessed 2021-03-20
[24] `https://openscience.adaptcentre.ie/projects/GDPRtEXT/`, last accessed 2021-03-20
[25] `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, last accessed 2021-03-20
[26] `https://old.datahub.io/dataset/gdprtext`, last accessed 2021-03-20
[27] `http://openscience.adaptcentre.ie/sparql`, last accessed 2021-03-20
[28] `https://bo-ecli.eu/`, last accessed 2021-03-20

Listing 6.1: Query example question 1

```
SELECT DISTINCT ?Reference ?Text ?Type
WHERE {
  ?justiz rdfs:label "10Ob60/17x" .
  ?justiz dcterms:references ?ref .
  {
    ?ref rdf:type lkg:LegalProvision ;
         rdfs:label ?Reference ;
         eli:is_realized_by ?realization .
    ?realization lkg:has_text ?Text .
    ?ref eli:type_document ?type_document .
    ?type_document skos:prefLabel ?Type .
  } UNION {
    ?ref rdf:type lkg:JudicialResource ;
         dcterms:type av:jud_rs ;
         rdfs:label ?Reference ;
         lkg:has_text ?Text .
    av:jud_rs skos:prefLabel ?Type .
  } UNION {
     ?ref rdf:type lkg:JudicialResource ;
         dcterms:type av:jud_te ;
         rdfs:label ?Reference ;
         lkg:has_text ?Text .
     av:jud_te skos:prefLabel ?Type .
  }
    FILTER (lang(?Type) = 'de')
}
ORDER BY ?type
```

Table 6.7: Result example question 1

Reference	**Text**	**Type**
"§ 1333 ABGB"	"§ 1333. (1) Der Schaden, den der Schuldner [...]"	"Bundesgesetz"
"§ 28a KSchG"	"§ 28a. (1) Wer im geschäftlichen Verkehr [...]"	"Bundesgesetz"
"5Ob145/11a"	"Im Rahmen der Verbandsklage hat die Auslegung [...]"	"Rechtssatz"
...	...	...

respective database, e.g. the law database for legal provisions, and manually search the referenced document in order to get the content. In a knowledge graph, we can combine several involved steps into a single query that returns a court decision with all referenced documents, their texts, plus types of the documents. This leads to a more **efficient legal information search process**. In order to enable such a query, we need to extract the referenced documents from the court decision and replace them with the respective URIs as well as a schema of document types.

Listing 6.1 shows the convenience of such a query for the court decision with case number "10Ob60/17x". The lawyer gets all referenced documents with their text and sorted by their types as a result, as illustrated in Table 6.7. In particular, the added links between the documents enable a query across application boundaries (judiciary documents, legislative documents) present in RIS and also to obtain the actual document text with the query. The query shows that a court decision is searched by the case number specified via `rdfs:label` uses the predicate `dcterms:references` to get the referenced legal provisions (`?ref rdf:type lkg:LegalProvision`) and judicial resources (`?ref rdf:type lkg:JudicialResource`), which are either decision texts (`dcterms:type av:jud_te`) or legal rules (`dcterms:type av:jud_rs`). We are also able to obtain the document text via `lkg:has_text` for each reference.

Listing 6.2: Query example question 2

```
select ?court where {
  ?geo gn:name "Krieglach" .
  ?jd lkg:judicial_district_member ?geo ;
      lkg:court_having_jurisdiction ?c .
    ?c rdfs:label ?court
}
```

Table 6.8: Result example question 2

Court
"Bezirksgericht Mürzzuschlag"

Q 2 *Over which districts does a court have competent jurisdiction?*

Legal databases are typically domain-specific and focus on legal matters only without additional contextual references that would be useful to be included for scoping search (such as explicit spatio-temporal references). For instance, a lawyer has a client who is facing a lawsuit regarding a property. Therefore, the lawyer needs to know which court has spatial competent jurisdiction (as required by law), in order to find related cases in a regional context. At the moment, this information is not made explicit in the legal information system and the lawyer would need to look through various websites of the authorities to find out about the regionally competent jurisdiction. This problem can be addressed by integrating external data in our legal knowledge graph and leads to **enriched information content and better user experience**. In our knowledge graph, we have readily linked the information about the Austrian courts and the judicial districts from the respective authorities with a geospatial hierarchy. Therefore, we can easily provide such information again by a straightforward SPARQL query also taking the court hierarchy into account.

As shown in Listing 6.2, the lawyer is now able to query the court having competent jurisdiction for a specific geospatial entity, just by providing the name of a community, for instance "Krieglach". At first, the query finds the spatial entity with the name (`gn:name`) "Krieglach" and uses this spatial entity to obtain the linked judicial district (`lkg:judicial_district_member`) and the associated court (`lkg:court_having_jurisdiction`). Finally, the lawyer finds the "Bezirksgericht Mürzzuschlag" (district court), which has competent jurisdiction as shown in Table 6.8.

Q 3 *What are the national transpositions of a specific EU directive?*

Legal systems differ across countries but still we need to consider legal information from other countries from time to time. Especially in an European context with the EU's harmonization activities through issuing common regulations, but also directives, which need to be transposed into national legislation. For companies wanting to expand their business abroad, it is necessary to know the legal situation and standards in these foreign countries. So far, a lawyer needs to search for the legal information system of the other country and find out how a particular directive, that is relevant for the company, has been transposed.[29] Also, the EUR-Lex search interface is not always helpful here, because it

[29] Further tedious search would be needed to find out about and compare respective jurisdictions across countries.

Listing 6.3: Query example question 3

```
select ?country ?title ?document where {
   VALUES ?format {
      <http://www.iana.org/assignments/media-types/text/html>
      <http://www.iana.org/assignments/media-types/application/html> }
  ?n ?p <http://data.europa.eu/eli/dir/2014/92/oj> ;
        eli:relevant_for ?c ;
        eli:is_realized_by ?r .
   ?r eli:title ?title ;
     eli:is_embodied_by ?document .
   ?document eli:format ?format .
   ?c skos:prefLabel ?country .
  FILTER (lang(?country) = 'en')
}
```

Table 6.9: Result example question 3

Country	Title	Document
"Ireland"	"European Union (Payment Accounts) Regulations 2016."	Document 1
"Austria"	"Bundesgesetz, mit dem ein Bundesgesetz über [...]"	Document 2
"Austria"	"Verordnung der Finanzmarktaufsichtsbehörde (FMA) über [...]"	Document 3
...	...	

Document 1: `http://www.irishstatutebook.ie/eli/2016/si/482/made/en/html`
Document 2: `https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2016_I_35/BGBLA_2016_I_35.html`
Document 3: `https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2018_II_60/BGBLA_2018_II_60.html`

does not provide the transposed texts. Integrating legal data across countries in a legal knowledge graph thus would enable **cross-jurisdictional search of legal information.**

In Listing 6.3, we demonstrate how this can be achieved across countries that follow the proposed ELI and ECLI standards for legal data. As shown, the company lawyer is able to find the concrete national transpositions of a given directive "2014/92/EU" with the actual transposed texts, across national legislations, again with a single query. Starting with a particular EU directive under investigation, in this example "2014/92/EU", the national documents are found which are usually linked via `eli:transposes`. This interlinking is the crucial part to establish the connection between the national and EU legal data. The remainder of the query gets the metadata for each national transposition like the title (`eli:title`), the actual jurisdiction (`eli:relevant_for`) and the actual document (`eli:is_embodied_by`). Further integrating and harmonizing existing legal knowledge graphs across countries, as discussed in Section 6.1, would further enable comparison of the respective jurisdiction for a particular directive. The results are shown in Table 6.9, which includes the national transpositions for Austria and Ireland with a direct link to the respective documents.

Q 4 *Which legal documents regulate a specific legal area searched with keywords in a foreign language?*

Legal systems are not only different in their structure but legal documents are typically penned in the official language(s) of a country. This puts an additional language barrier in the legal information search process. Additional sources such as the EuroVoc thesaurus, ideally aligned with national thesauri, which contain terms in multiple languages to the legal knowledge graph enables **multi-lingual search of legal information**. Linking legal documents with concepts instead of language-specific labels allows users to search in their language for documents written in another language.

Listing 6.4: Query example question 4

```
select ?law ?legalprovision ?document where {
   ?ev skos:prefLabel "protezione del consumatore"@it .
   ?austrovoc rdfs:seeAlso ?ev .
  ?lp eli:is_about ?austrovoc ;
      eli:jurisdiction <http://publications.europa.eu/resource/authority/country/AUT> ;
      eli:in_force eli:InForce-inForce ;
      eli:is_realized_by ?le ;
      lkg:has_number_paragraph ?number ;
      rdfs:label ?legalprovision .
   ?le eli:title_alternative ?law ;
      eli:is_embodied_by ?document .
   ?document eli:format <http://www.iana.org/assignments/media-types/application/html>
}
ORDER BY ASC(?law) ASC(?number)
```

Table 6.10: Result example question 4

Law	Legal Provision	Document
"KSchG"	"§ 1 KSchG"	Document 1
"KSchG"	"§ 42 KSchG"	Document 2
"VKrG"	"§ 1 VKrG"	Document 3
...	...	

Document 1: https://www.ris.bka.gv.at/Dokumente/Bundesnormen/NOR12041200/NOR12041200.html
Document 2: https://www.ris.bka.gv.at/Dokumente/Bundesnormen/NOR40050352/NOR40050352.html
Document 3: https://www.ris.bka.gv.at/Dokumente/Bundesnormen/NOR40117826/NOR40117826.html

In Listing 6.4 an Italian lawyer is, for instance, researching in a lawsuit covering another country. The lawyer is interested in the Austrian legal provisions covering the specific legal area “protezione del consumatore” and is able to search in his language. For this case, the EuroVoc thesaurus and the language versions for each concept contained therein can be utilized. A concept has labels (`skos:prefLabel`) for each EU language and the language independent concept is used for the further query. As shown, the EuroVoc concept is linked (`rdfs:seeAlso`) to the AustroVoc concept, which is used to retrieve the respective legislative documents related to the EuroVoc concept initially search in Italian. The results for this query are shown in Table 6.10 and point the lawyer to the consumer protection law (“KSchG”) and consumer credit law (“VKrG”). Different languages are a barrier and supporting multi-lingual search is a step towards improved, more transparent access to legal information.

Q 5 *Which events are mentioned in a court decision and could be used for a quick overview of the case?*

Court decisions are potentially very long documents containing information about the case, the legal proceedings and the legal assessment together with the actual verdict. Depending on the structure of such court decisions, related information might be presented together in a particular section or is spread across the entire document. In order to get an overview about a case, it is necessary to read the court decision, which easily turns into a very time-consuming process when such a document consists of multiple pages. Extracting and classifying legal events including temporal information, the subjects, for instance the acting persons, and a description of the event in a structured format, is beneficial for the reader to get a **quick overview about what happened when** in a case.

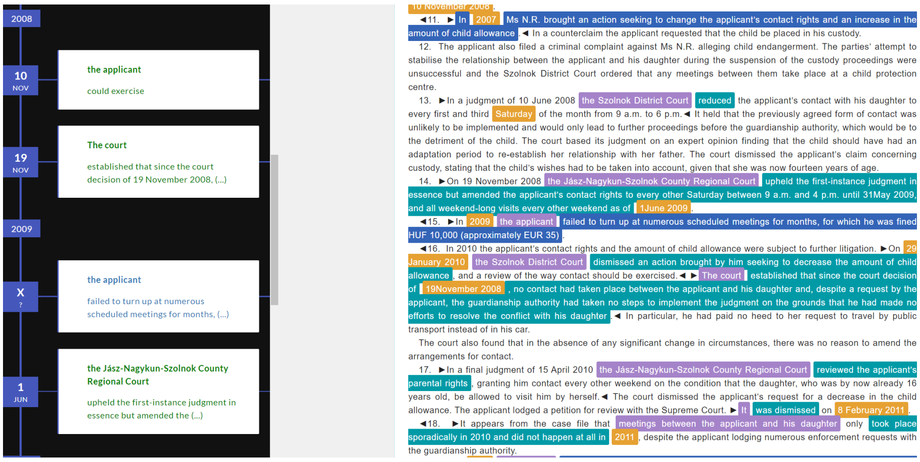

Figure 6.1: Example of event extraction from a court decision.

An example of such a summary of case events is shown in Figure 6.1, which is not a query as such, but can be generated on the fly from a given court decision. Note that we use an ECHR court decision for demonstration as the model has been trained on ECHR documents (cf. Section 5.2). The example shows an ECHR court decision with a timeline on the left hand side and the legal events highlighted in the text. In more detail, we can see that temporal expressions are highlighted in orange and the subjects, persons and things, and highlighted in purple. The core of the events describing what actually happened are colored green when referring to a procedural events and blue when describing circumstances.

6.4 Roadmap towards a Linked Legal Knowledge Graph

The current situation towards a truly interconnected legal knowledge graph on a European level looks promising, with many good starting points, but some challenges lie ahead to be addressed. On the one hand, providers of legal information, typically governments, would need to help to ease the access to law and support non-governmental initiatives to provide and obtain legal information. On the other hand, these providers are confronted with resource restrictions and other priorities, which slows down this process. We discuss some of the related challenges in the following.

6.4.1 Licensing and Access Policies

The publication of and access to legal information might be hindered by licensing and access policies, or lack thereof. Open (government) data is a goal of the European Union as laid out in the PSI-Directive[30], which stipulates that documents from the public sector should be made available free of charge in machine-readable and open formats as well as the possibility for a mass download. The PSI directive goes hand in hand

[30]EU 2019/1024, `http://data.europa.eu/eli/dir/2019/1024/oj`, last accessed 2021-03-20

with the *8 Open Government Data Principles*[31] to provide data in a machine-readable, license free, complete and accessible format in a timely manner. Following open government data publication methodologies, such as COMSODE [Kucera et al., 2015], helps governments to set up respective publication strategies. The terms and conditions should be communicated in a clear manner and data provided ideally under a permissive license, which also allows private initiatives to use the data for their business model by providing additional services, e.g. build on the data and restrict access to certain parts of the knowledge graph such as linked legal commentaries.

6.4.2 Support of Linked Legal Data Initiatives

Our analysis of the legal landscape (cf. Section 6.1) shows that documents are provided in various formats with structured formats being the minority. The problem of having documents in an unstructured format as a starting point (e.g. [Chalkidis et al., 2017]) might slow down the process of the providing linked legal data. It is therefore desirable that legal documents are provided in a structured format from the very beginning in order to enable the transition to and participation in an EU-wide linked legal data ecosystem. Hence, following the Linked Data Principles together with using appropriate linked data formats such as JSON-LD [W3C Community Group, 2012], RDF serializations or XML standards for legal documents, such as Akoma-Ntoso[32], enables easy access to the data for linked legal data initiatives. The EU can help member states in activities towards the provision of linked legal data by providing detailed guidelines on how to use the proposed ELI and ECLI standards or software tools supporting the transition. Furthermore, the provision of dedicated vocabularies in addition to the existing named authority lists and EuroVoc thesaurus, which do not really fit the requirements of member states, are beneficial as it reduces the barrier of participating in ELI and ECLI.

We emphasize here, that despite the resulting documents are typically plain text documents, in many countries – including Austria – the legal document preparation process is regulated by clearly defined processes where, as opposed to extracting unambiguous metadata on hindsight only - such metadata and linked data creation could and should be directly included into these processes.

6.4.3 Information Provision

The lack of coordination in terms of ELI and ECLI implementation concerns the European Union as well as EU member states. Currently, it is a very time-consuming task to find any information about ELI and ECLI implementation in different countries. At the moment, the information is cluttered with some countries using the EU e-Justice portal or others providing respective information only on national websites. Furthermore, implementation details can often only be inferred from studying the source code of example documents, rather than by available documentation. Positive examples of countries providing extensive information are, for instance, Denmark[33], Finland[34], and Luxembourg[35], which run national websites with implementation information about the

[31] `https://public.resource.org/8_principles.html`, last accessed 2021-03-20
[32] `http://www.akomantoso.org/`, last accessed 2021-03-20
[33] `https://www.retsinformation.dk/eli/about`, last accessed 2021-03-20
[34] `https://data.finlex.fi/en/datamodeling`, last accessed 2021-03-20
[35] `http://www.legilux.lu/editorial/casemates`, last accessed 2021-03-20

ELI. The same applies to the usage of NAL, which is encouraged by the ELI and ECLI ontologies. Without additional information about the used NAL it is a tedious task for outsiders to find information which NAL are used. In addition to missing information websites about the NAL, some countries use NAL but these NAL cannot be retrieved from the internet or dereferenced. As argued herein, aligning the ELI and ECLI pages at EU level, hence integrating ELI into the EU e-Justice portal, and providing templates for member states about their ELI and ECLI implementation status as well as the usage of national NAL could be highly beneficial. More consistent best practices would also help other, not yet participating countries to investigate what and how to implement ELI and ECLI in an overall more aligned manner, which in turn might lower the barrier to participate.

6.4.4 Search Interfaces

Access to legal information should be as easy as possible for end users as well as data processing professionals. Centralized web search interfaces serving as a *one-stop-shop* with a graphical user interface enabling the access to legal documents from various authorities eases the search process for the end user, citizens and legal professionals. Linked legal data initiatives enable such centralized aggregation of legal information, and can also support common application programming interfaces (API) – such as, e.g. access through the SPARQL protocol – as well as indexes to access and retrieve legal data for subsequent processing.

6.4.5 Multilinguality

Legal data is typically presented in the official language(s) of the respective country, some of the legal information systems provide some laws (e.g. civil code and the constitution) in English. As demonstrated herein, one approach to enable better multi-lingual search is to link national indexes with the multi-lingual EuroVoc thesaurus, which then acts as a connecting point between legal information provided in different countries and languages. Yet, we also emphasize the importance of national extensions (such as the proposed AustroVoc thesaurus) to cover countrywise specifics, or for keeping ambiguous language use in different legislations/jurisdictions (e.g. Germany and Austria) separate. We envision the creation of similar national extensions, for instance “SpainVoc” or “IrishVoc”, by other member states. Another emerging approach to the multilinguality challenge is to create graph-based Linked Data native dictionaries that include lexical knowledge and overcome the disadvantages of tree-based dictionaries [Gracia et al., 2017]. Others enrich the underlying ontology with linguistic information, for instance as proposed by the Ontolex-lemon model ([McCrae et al., 2017, W3C Ontology-Lexica Community Group, 2016]). Finally, multilinguality could be further supported by adding linguistic and lexical information to enable NLP applications working with this information contained in an ontology.

6.4.6 Modeling Standards

In order to achieve the overarching ELI and ECLI goals, EU member states should follow the modeling standards outlined in these proposals. Both ELI and ECLI describe a minimum set of non-country specific metadata and are therefore very well suited for national extensions where needed. Our comparison of the linked legal data features in the EU member states (cf. Table 6.1) shows that most of the participating countries follow the proposed modeling standards. Some countries, for instance Luxembourg provide their JOLUX ontology in their own as well as the ELI format. Individual deviations from these standards undermine the fundamental ideas of easier access to legal information across borders. One of the drawbacks of the current modeling standard, is the need to write queries in order to retrieve certain data as shown by [Francesconi et al., 2015]. The proposed solution, which involves decoupling the ELI and FRBR ontologies, needs to be approached and initiated in a centralized manner, for instance via a stakeholder engagement process whereby national experts who know their legal system and experts from the responsible EU institutions work together in order to shape future ELI and ECLI enhancements.

6.5 Summary and Future Directions

In this chapter, we explored the linked legal data landscape focusing on EU member states and the availability of their legal data in a structured format such that it could be integrated in a legal knowledge graph.

In Section 6.1, we analyzed at the availability of legal information provided by the governments of EU member states. The focus was on the implementation of the proposed ELI and ECLI ontologies, but we also investigated the current situation regarding publicly available legal information systems and which information could be found there. The findings regarding the implementation of ELI and ECLI were a bit disappointing. Only 11 out 27 EU member states implemented the first ELI pillar (assign an ELI identifier to legislative documents) but nine countries also provide metadata for their legal documents, which is a good sign, although sometimes the government is supported by non-governmental initiatives. The results regarding the implementation of the ECLI are slightly better with 19 countries assigning ECLI identifiers but only three countries also provide metadata for their judicial documents. Furthermore, a few countries extend the ELI and ECLI ontologies with national ontologies in order to properly represent national requirements of their legal system, while a few others chose to use a different ontology, for instance OGP. A comparison of the legal information systems of the EU member states revealed that only seven countries provide a central search interface for legal information, while it is more common that each authority publishes – if at all – documents on its own website and format. Regarding the search process, a keyword-based search interface was found to be most common, while a SPARQL endpoint is provided by two countries only. Not surprising, legal documents are mainly available in HTML and PDF format in the national language of the respective country. The "most important" laws, usually the civil code, constitution and alike, are also translated into English and other official languages of the country.

Non-governmental initiatives are covered in Section 6.2, which are mostly projects carried out by academia. We found two groups of projects focusing on the interlinking of and the extraction of references in legal documents. The majority of these projects deals with ELI and provides data in RDF format. However, they mostly focus on specific aspects, for instance a certain area of law for their project or the extraction of specific legal entities.

In Section 6.3, we demonstrated the benefits of an integrated legal knowledge graph by showing how we can solve the example questions described in the introduction with SPARQL queries. These questions were not possible to be solved with single queries in a traditional legal information system but with the help of a legal knowledge graph. Moreover, we also showed that legal data from other countries can be easily integrated in the legal knowledge graph and extending the search space for a user when it is provided by the country. We showed that it would be possible to create a European legal knowledge graph when countries participate in the ELI and ECLI proposals.

Unfortunately, not all countries are interested in providing their legal data in a structured and aligned way with other countries. That is why we provided a roadmap towards a linked legal knowledge graph describing the key challenges preventing countries to provide easier access to law in Section 6.4. Such challenges comprise several aspects, might they be (i) legal regarding licensing and access policies for the provision of publicly available and free of charge accessible legal information; (ii) administrative to get several so far independent authorities to agree on publication ways; or (iii) financial to provide the required resources in order to publish linked legal data.

Future work includes the monitoring of linked legal data activities across the EU member states and, when a larger number of countries participates, comparing the national extensions to find common classes or properties. These findings could be used for further updates of the ELI and ECLI ontologies. Furthermore, an integrated legal knowledge graph could also be used to explore the semantic meaning, differences and ambiguities of legal terms across countries and languages.

CHAPTER 7

Conclusion

This thesis aimed at enhancing the legal search process by turning a traditional legal information system into a legal knowledge graph. To this end, we first created the model of a legal knowledge graph in the form of an ontology that combined existing schemata and ontologies with novel, specific extensions for our use case - the Austrian legislation and jurisdiction - in a middle-out fashion. One the one hand, we had to take the proposals of the European Union for the European Law Identifier (ELI) and the European Case Law Identifier (ECLI) into account, both providing a minimum set of metadata and an identifier for legislative documents and judiciary documents respectively. On the other hand, we had to consider the available data in the Austrian legal information system. We found that the proposed ELI and ECLI ontologies provide a basic set of metadata but lacked some information available in the Austrian legal information system RIS. The resulting Legal Knowledge Graph (LKG) ontology contains Austrian specific classes and properties to properly represent Austrian specific requirements of the legal system by extending the ELI and ECLI ontologies. Furthermore, since ELI and ECLI encourage member states to set up their own lists for certain properties, we created the AustroVoc thesaurus containing Austrian specific legal terms as an example of such a specific national extension of the general terminology provided by the EuroVoc vocabulary.

Next, in Chapter 4 we populated this Austrian legal knowledge graph with actual data. For this purpose, we introduced different population methods based on the available data and target properties. In general, we distinguished between the population from structured data and from unstructured data. For the population from structured data, we used the metadata from the Austrian legal information system. For the population from unstructured data, we extracted the required information from the legal documents. In more detail, the population from structured data, presented in Section 4.1, was subdivided into direct population, indirect population and population by interlinking with external sources. The direct population allowed us to transfer the data directly, without any manipulation, from the legal information system into the legal knowledge graph. Other properties required a little syntactical manipulation, for instance changing the date formats. We also showed that we can use available data to interlink it with external sources, such as Geonames. For the population from unstructured data, we employed different NLP methods to extract additional information from the actual document texts, for instance decision texts of the Austrian Supreme Court. We showed that a population from unstructured data could be achieved by the extraction of legal entities from the documents as well as by the classification of documents into given categories of a thesaurus. In order to train and evaluate our information extraction pipeline, we manually annotated a corpus of 50 Austrian Supreme Court decision texts with legal entities and used this new corpus to compare the performance of traditional and state-of-the-art automatic approaches. Furthermore, we proposed a new approach to boost the performance for document classification approaches by exploiting the class hierarchy of the EuroVoc thesaurus and evaluated this method with several machine learning and deep learning approaches and different language models.

In Chapter 5, we investigated the extraction of temporal information from court decisions. Temporal information might have an impact on the court proceedings and the applicable law. Here, we analyzed the special properties of legal English used in court decisions and introduced three different temporal dimensions, which can occur within court decisions. We compared and evaluated the performance of ten state-of-the-art temporal taggers on a new corpus of in total 30 court decisions. The corpus is composed of ten court decisions each, from the European Court of Human Rights, the European Court of Justice and the United States Supreme Court, which again we annotated manually. Moreover, the documents in this corpus were annotated with two annotation sets, one following the TimeML standard including all temporal expressions and the other containing only temporal expressions relevant for the legal domain.

As the results of the evaluations in Chapters 4 and 5 show, in the field of legal text analysis we can rarely rely on existing benchmark datasets, which is why we hope that the respective annotated corpora we created in these chapters as such provide a valuable contribution for others to work with and extend.[1,2,3]

In a final step, we described the integration of legal data in the legal knowledge graph with similar initiatives from other EU member states in Chapter 6. For this purpose, we conducted an in-depth analysis of the current situation regarding the provision of legal data across the EU member states in Section 6.1. We focused on the participation of the EU member states in the ELI and ECLI initiatives and provided an overview for each country about the ELI and ECLI implementation status as well as additional or other ontologies might have been used. Furthermore, we investigated the availability of (linked) legal data, for instance for download in RDF or embedded in HTML using RDFa, and the available background information about the implementation, used ontologies and named authority lists. Moreover, we also analyzed the availability of traditional legal information systems in the EU member states and provided an overview along multiple criteria, for example the available document formats to view/ download the documents and languages in which the documents are available. Besides the governmental initiatives, we looked into non-governmental initiatives in Section 6.2, which used ELI and ECLI to interlink or to extract information from legal documents. These initiatives were also classified based on different categories. After having obtained the information about the initiatives, we were able to integrate the Austrian data with the data from other EU member states. In Section 6.3, we demonstrated the benefits of a legal knowledge graph with example SPARQL queries, which showed the possibility to receive legal information, also including data from external knowledge bases and across borders, with a single query that was either not possible before or required multiple search queries across different legal databases. However, the participation in ELI and ECLI could be improved, which is why we proposed a roadmap in Section 6.4 containing the – in our opinion – most important points towards the provision of linked legal data and ultimately

[1]RIS, `https://github.com/efiltz/legal-knowledge-graph`, last accessed 2021-03-12
[2]TempCourt, `https://tempcourt.github.io/TempCourt/`, last accessed 2021-03-12
[3]EventsMatter, `https://mnavasloro.github.io/EventsMatter/`, last accessed 2021-03-12

reaching the goal of an integrated legal knowledge graph containing data of all EU member states.

7.1 Assessment of Research Questions

The summarized contributions above span the creation of a legal knowledge graph, from the modeling to the population and integration of external legal data, to finally demonstrating how it is now possible to seamlessly answer a variety of questions, that typically occur in research workflows by legal practitioners. From the perspective of our original research questions stated in Section 1.1, we can conclude the following results.

RQ 1 *What is required in order to construct a legal knowledge graph from an existing legal information system?*

The answer to this research question is presented in Chapter 3. We compare the available classes and properties of the ELI and ECLI ontologies and extend them by applying a middle-out approach. We see that the driving factors for the ontology extension are the available data in the legal information system and the classes as well as the properties from the ontologies as described in Section 3.3. Furthermore, in cases where the ELI and ECLI ontologies suggest the creation of national schemes, we created the AustroVoc thesaurus in Section 3.4.

RQ 2 *Which approaches can be followed in order to populate the legal knowledge graph from different data sources in an automated fashion?*

To answer this research question we investigated three different cases, which are expressed with the three sub-research questions:

RQ 2.1 *Which approaches are available for the population of the legal knowledge graph from structured data and how effective are they?*

In Section 4.1, we propose three approaches, which have in common that the data required for the population of the legal knowledge graph is available in a structured format. As an answer to this question, we can say that the most convenient way to populate the knowledge graph is by direct population, which is mainly applicable to simple strings, for instance the document title.

RQ 2.2 *Which approaches are available for the population of the legal knowledge graph from text sources (i.e. legal documents) and how effective are they?*

The population of the legal knowledge from text sources involves the application of NLP tools and techniques. In Section 4.2, we compare the performance of rule-based, machine learning and deep learning-based legal entity extraction approaches applied on a new, manually annotated corpus of 50 Austrian Supreme Court decisions. The results show that there is no approach outperforming all other approaches but the performance depends on the complexity of the entity that needs to be extracted. The more complex a legal entity is, the better is the performance of deep learning-based approaches. However, all results of the compared approaches are close together.

In Section 4.3, we propose an approach that exploits the concept hierarchy of the EuroVoc thesaurus to boost the classification results with a very little information loss. We compare machine learning and deep learning approaches and see that an approach using neural networks is able to outperform classic machine learning approaches. However, the performance overhead is not as high as expected, which is caused by the large number of classes in the EuroVoc thesaurus.

RQ 2.3 *Which approaches are available for the extraction of events from legal documents and how effective are they?*

Legal events can be used to summarize what happened in court decisions. For this purpose, we present a comparison of ten state-of-the-art non-domain specific temporal taggers on their performance of extracting temporal information from a new and manually annotated corpus of 30 court decisions. The compared temporal taggers use rule-based, machine learning and hybrid approaches. The results show that the performance depends on the structure and wording of the text, but in general, rules provide solid results. However, legal English causes some problems as citations are often mistaken for, or wordings are wrongly interpreted as temporal expressions.

In the a next step, we propose the extraction of legal events from court decisions in Section 5.2. We introduce the concept of legal events, which consist of temporal expression, subject and an event core. The extraction of legal events is evaluated on a newly created and manually annotated corpus of 30 court decisions using rules, machine learning and deep learning approaches. In general, we can say that using deep learning approaches with fine-tuned language models to the legal domain provide the best results. However, there is a connection between the structure of the event components and the evaluation results, which can be summarized as the more structured, the better the results.

The overall answer to this research question is that the population of the legal knowledge graph in an automated fashion is possible, but the selection of the best approach highly depends on the available data and resources. When the application of NLP tools and techniques is required, it is very likely that state-of-the-art approaches perform at least equally good compared to traditional approaches. However, the required effort to create training data, which is hardly available for the legal domain and the reason why we created three different corpora of legal documents, needs to be taken into account. Furthermore, the cost-benefit ratio in terms of required performance and available human and computational resources is subject to an individual assessment.

RQ 3 *In how far is it possible to enhance the legal inquiry and search process by linking legal data?*

The answer to this research question is presented in Section 6.3. We show that a newly created legal knowledge graph with national Austrian data and integrating legal data from external sources is highly beneficial. In more detail, the benefits of an integrated legal knowledge graph are demonstrated with sample questions that cannot be answered with traditional legal information systems but with a legal knowledge graph.

7.2 Future Research Directions

While we have demonstrated the benefits of an integrated legal knowledge graph over a traditional legal information system, we also identified future research directions that would be worthwhile investigating.

Improvement of ELI and ECLI Ontologies

In the light of the goals laid out in the ELI and ECLI proposals [Council of the European Union, 2012, Council of the European Union, 2011] it would be worthwhile to continue tracking of the advances in the implementation of the ELI and ECLI standards across the EU member states, especially with regards to the actual implementation and modeling approaches, which we already discovered in the course of our work (cf. Chapter 6). With more participants and available data it might be possible to revise the ELI and ECLI ontologies based on found overlaps in the national extensions towards a better alignment of legal data representation. Furthermore, this also enables improvements to reduce the query complexity as already proposed by [Francesconi et al., 2015].

Extension of the Legal Knowledge Graph

While, ELI and ECLI aim at the interlinking of legal documents it would certainly be interesting to include – apart from temporal event and spatial references – even more additional factual knowledge encoded in those documents, such as tax rates or penalty ranges. Furthermore, the data contained in a legal knowledge graph would fit nicely into providing legal case knowledge in an accessible way on particular regulations and compliance assessments in relation to related court cases. Unfortunately, recent studies ([Leone et al., 2019, de Oliveira Rodrigues et al., 2019]) show that the majority of legal ontologies are pre-date and are not referenced by ELI and ECLI. However, it would be interesting to investigate whether these prior ontologies are worth to be integrated in the legal knowledge graph and what effort would be required. Another future work is the extension of the legal knowledge graph with ontologies that are not directly related to the legal domain, but could be used to represent information contained in legal documents. In Chapter 5, we focused specifically on temporal expressions and legal events, which could also be integrated into the legal knowledge graph based on a "temporal ontology". A starting point could be to turn existing standards (TIDES TIMEX [Ferro et al., 2005] and TimeML [Pustejovsky et al., 2003b]) into ontologies or to join current efforts in this direction, for instance the Time Ontology for OWL W3C Candidate Recommendation [Spatial Data on the Web Working Group, 2020].

Provision of Legal Resources

An area needing further attention is the provision of legal resources, which can be used for experiments with legal documents. Such resources do not only encompass annotated legal datasets but also extensive language models trained on large corpora of legal documents. Available resources mainly contain English legal documents from the European Union, for instance [Chalkidis et al., 2019, Chalkidis et al., 2020b], while there is clearly a lack of resources from national jurisdictions. It is therefore desirable to also provide appropriate resources, hence annotated datasets and language models, from national jurisdictions. We could only fill this gap partially by connecting resources from Austria's national jurisdiction. More of such focused national efforts, driven by and requiring expert knowledge in the respective national legal system, and document publishing process, seem necessary to arrive at a more complete picture of the trans-national legal landscape across Europe and beyond.

Improvement/Adaption of NLP Techniques specifically for Legal Texts

Despite the good results of our entity extraction and document classification experiments (cf. Chapter 4, Chapter 5), there is still room for improvement. Especially the document classification setting with a large number of classes and skewed distribution needs attention. Particularly, since for legal text classification we often face cold-start problems with little or not at all available training data. In order to overcome this problem, zero-shot learning [Romera-Paredes and Torr, 2015] could be a research direction worthwhile investigating. A starting point to tackle this problem for the legal domain using, among others, the EUR-Lex dataset, has been presented by [Chalkidis et al., 2020a].

Support of Multilinguality

Last but not least, a major barrier in cross-border activities is still the language, which is not so much a problem with legal documents issued by the European Union but with national documents, which are mainly provided in the official languages of the member states only. A legal knowledge graph helps to find the relevant foreign legal information, while the actual content of the found legal documents might remain a mystery without proper language skills. The simplest solution would be that the issuing authorities provide the documents in all languages, however, we know that this is unfeasible from a practical point of view. Furthermore, the legal language might differ from the ordinary language in terms of the used words but also in their meaning such that simply translating texts while retaining the exact meaning might not be possible and also cause legal issues [Stolze, 2001, Felici, 2010]. From this point of view, we can imagine that integrating linguistic and lexicographical information into the knowledge graph could act as a starting point.

APPENDIX A

Appendix

A.1 Legal Databases and Example Documents

Table A.1: Overview of legal databases and example documents for legislation

URLs of example documents are shortened and last accessed on 2020-07-15.

Country	Legislation	Example Document
Austria	https://ris.bka.gv.at/	https://bit.ly/2UEq4E9
Belgium	http://www.ejustice.just.fgov.be/	https://bit.ly/30zHeGx
Bulgaria	https://dv.parliament.bg/	https://bit.ly/2MQ7q83
Croatia	http://nn.hr/	https://bit.ly/3hnXy34
Cyprus	http://www.cylaw.org/	https://bit.ly/3hew1Bc
Czech Republic	https://aplikace.mvcr.cz/sbirka-zakonu/	https://bit.ly/2XVLajg
Denmark	https://www.retsinformation.dk/	https://bit.ly/2YwzBhs
Estonia	https://www.riigiteataja.ee/	https://bit.ly/2XUxLIf
Finland	https://www.finlex.fi/	https://bit.ly/2UEbRXA
France	https://www.legifrance.gouv.fr/	https://bit.ly/2XUy4Tp
Germany	http://www.bgbl.de/	https://bit.ly/3cV7dLh
Greece	http://www.et.gr/	https://bit.ly/2B4bApT
Hungary	http://njt.hu/	https://bit.ly/3d2iQQN
Ireland	http://www.irishstatutebook.ie/	https://bit.ly/2XUhVxd
Italy	https://www.normattiva.it/	https://bit.ly/30zls5Z
Latvia	http://www.likumi.lv/	https://bit.ly/2UDUJBl
Lithuania	https://www.e-tar.lt/	https://bit.ly/2XVquIj
Luxembourg	http://legilux.public.lu/	https://bit.ly/30ycd5Q
Malta	https://legislation.mt/	https://bit.ly/2XSrgpq
Netherlands	https://www.officielebekendmakingen.nl/	https://bit.ly/2Ooq5IV
Poland	http://isip.sejm.gov.pl/	https://bit.ly/3hkOc8i
Portugal	https://dre.pt/	https://bit.ly/3gOtNrn
Romania	http://legislatie.just.ro/	https://bit.ly/37lNJhA
Slovakia	https://www.slov-lex.sk/	https://bit.ly/2XUz4a7
Slovenia	http://www.pisrs.si/Pis.web/	https://bit.ly/3cVWu2Y
Spain	https://boe.es/	https://bit.ly/2AjLPCk
Sweden	http://rkrattsbaser.gov.se/	https://bit.ly/3d2jQV3

Table A.2: Overview of legal databases and example documents for jurisdiction

URLs of example documents are shortened and last accessed on 2020-07-15.

Country	Judiciary	Example Document
Austria	https://ris.bka.gv.at/	https://bit.ly/37maTo6
Belgium	http://jure.juridat.just.fgov.be/	https://bit.ly/2AjPLTB
Bulgaria	https://legalacts.justice.bg/	https://bit.ly/3hpfdI2
Croatia	https://sudskapraksa.vsrh.hr/home	https://bit.ly/3fiabv0
Cyprus	http://www.cylaw.org/	https://bit.ly/30BXxD0
Czech Republic	http://www.nsoud.cz/	https://bit.ly/2Ywztyu
Denmark	https://domstol.dk/	https://bit.ly/2MQrqan
Estonia	https://www.riigiteataja.ee/	https://bit.ly/2UFChIK
Finland	https://www.finlex.fi/	https://bit.ly/3cS3w93
France	https://www.courdecassation.fr/	https://bit.ly/30CQjyq
Germany	http://www.bundesverfassungsgericht.de/	https://bit.ly/2BXbHUN
Greece	http://www.adjustice.gr/	https://bit.ly/3feYwwT
Hungary	https://birosag.hu/birosagi-hatarozatok-gyujtemenye/	Direct download
Ireland	https://beta.courts.ie/	https://bit.ly/2YpEwRm
Italy	http://www.italgiure.giustizia.it/	Registration required
Latvia	https://manas.tiesas.lv/eTiesas/	https://bit.ly/30Bettm
Lithuania	https://www.lat.lt/	https://bit.ly/30BHfdc
Luxembourg	https://justice.public.lu/	https://bit.ly/3fnuJ5n
Malta	https://justice.gov.mt/	https://bit.ly/2XVMmmK
Netherlands	https://data.rechtspraak.nl/	https://bit.ly/3hlxaqN
Poland	http://orzeczenia.nsa.gov.pl/	https://bit.ly/2zuAq22
Portugal	https://jurisprudencia.csm.org.pt/	https://bit.ly/3dQNspr
Romania	http://www.rolii.ro/	https://bit.ly/2YtigWL
Slovakia	https://obcan.justice.sk/	https://bit.ly/2MO0BDX
Slovenia	http://www.sodnapraksa.si/	https://bit.ly/2XRJEia
Spain	http://www.poderjudicial.es/	https://bit.ly/3fm2ALX
Sweden	https://rattsinfosok.domstol.se/	https://bit.ly/3fckCjq

Bibliography

[ACE, 2005] (2005). The ace 2005 (ace 05) evaluation plan. evaluation of the detection and recognition of ace entities, values, temporal expressions, relations, and events. https://www.semanticscholar.org/paper/The-ACE-2005-(-ACE-05-)-Evaluation-Plan-Evaluation-NTRODUCTION/3a9b136ca1ab91592df36f148ef16095f74d009e.

[Agarwal et al., 2018] Agarwal, S., Steyskal, S., Antunovic, F., and Kirrane, S. (2018). Legislative compliance assessment: framework, model and gdpr instantiation. In *Annual Privacy Forum*, pages 131–149. Springer.

[Aggarwal and Zhai, 2012] Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 163–222. Springer.

[Ajani et al., 2016] Ajani, G., Boella, G., Caro, L. D., Robaldo, L., Humphreys, L., Praduroux, S., Rossi, P., and Violato, A. (2016). The european taxonomy syllabus: A multi-lingual, multi-level ontology framework to untangle the web of european legal terminology. *Applied Ontology*, 11(4):325–375.

[Ajani et al., 2007] Ajani, G., Lesmo, L., Boella, G., Mazzei, A., and Rossi, P. (2007). Terminological and ontological analysis of european directives: Multilinguism in law. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, page 43–48, New York, NY, USA. Association for Computing Machinery.

[Akbik et al., 2018] Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

[Alkhatib et al., 2017] Alkhatib, W., Rensing, C., and Silberbauer, J. (2017). Multi-label text classification using semantic features and dimensionality reduction with autoencoders. In Gracia, J., Bond, F., McCrae, J. P., Buitelaar, P., Chiarcos, C., and Hellmann, S., editors, *Language, Data, and Knowledge - First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings*, volume 10318 of *Lecture Notes in Computer Science*, pages 380–394. Springer.

[Alkhatib et al., 2018] Alkhatib, W., Sabrin, S., Neitzel, S., and Rensing, C. (2018). Towards ontology-based training-less multi-label text classification. In Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., and Meziane, F., editors, *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, volume 10859 of *Lecture Notes in Computer Science*, pages 389–396. Springer.

[Allahyari et al., 2017] Allahyari, M., Pouriyeh, S. A., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *CoRR*, abs/1707.02919.

[Altinel and Ganiz, 2018] Altinel, B. and Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Inf. Process. Manage.*, 54(6):1129–1153.

[Amit Singhal, 2012] Amit Singhal (2012). Introducing the Knowledge Graph: things, not strings. `https://blog.google/products/search/introducing-knowledge-graph-things-not/`. Online; accessed 2020-05-17.

[Angelidis et al., 2018] Angelidis, I., Chalkidis, I., and Koubarakis, M. (2018). Named entity recognition, linking and generation for greek legislation. In Palmirani, M., editor, *Legal Knowledge and Information Systems - JURIX 2018: The Thirty-first Annual Conference, Groningen, The Netherlands, 12-14 December 2018*, volume 313 of *Frontiers in Artificial Intelligence and Applications*, pages 1–10. IOS Press.

[Athan et al., 2013] Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., and Wyner, A. (2013). OASIS LegalRuleML. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, pages 3–12, New York, NY, USA. ACM.

[Austrian Supreme Court, 2016] Austrian Supreme Court (2016). 10Ob60/17x. `https://www.ris.bka.gv.at/Dokumente/Justiz/JJT_20180220_OGH0002_0100OB00060_17X0000_000/JJT_20180220_OGH0002_0100OB00060_17X0000_000.html`. accessed 2021-03-12.

[Baars and Kemper, 2008] Baars, H. and Kemper, H. (2008). Management support with structured and unstructured data - an integrated business intelligence framework. *Inf. Syst. Manag.*, 25(2):132–148.

[Bacci et al., 2019] Bacci, L., Agnoloni, T., Marchetti, C., and Battistoni, R. (2019). Improving public access to legislation through legal citations detection: The linkoln project at the italian senate. *Knowledge of the Law in the Big Data Age*, 317:149.

[Berners-Lee, 2006] Berners-Lee, T. (2006). Linked data design issues. `https://www.w3.org/DesignIssues/LinkedData.html`. Accessed: 2020-03-15.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific american*, 284(5):34–43.

[Bethard, 2013] Bethard, S. (2013). ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Proceedings of the Workshop SemEval 2013*, pages 10–14. ACL.

[Bhatia et al., 2015] Bhatia, K., Jain, H., Kar, P., Varma, M., and Jain, P. (2015). Sparse local embeddings for extreme multi-label classification. In *Advances in neural information processing systems*, pages 730–738.

[Biebow et al., 1999] Biebow, B., Szulman, S., and Clément, A. J. (1999). Terminae: A linguistics-based tool for the building of a domain ontology. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 49–66. Springer.

[Bizer, 2009] Bizer, C. (2009). The emerging web of linked data. *IEEE Intell. Syst.*, 24(5):87–92.

[Bizer et al., 2009] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.

[Boella et al., 2015] Boella, G., Caro, L. D., Graziadei, M., Cupi, L., Salaroglio, C. E., Humphreys, L., Konstantinov, H., Marko, K., Robaldo, L., Ruffini, C., Simov, K. I., Violato, A., and Stroetmann, V. (2015). Linking legal open data: breaking the accessibility and language barrier in european legislation and case law. In Sichelman, T. and Atkinson, K., editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015*, pages 171–175. ACM.

[Boella et al., 2016] Boella, G., Caro, L. D., Humphreys, L., Robaldo, L., Rossi, P., and van der Torre, L. (2016). Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law. *Artif. Intell. Law*, 24(3):245–283.

[Boella et al., 2019] Boella, G., Caro, L. D., and Leone, V. (2019). Semi-automatic knowledge population in a legal document management system. *Artif. Intell. Law*, 27(2):227–251.

[Boella et al., 2012] Boella, G., Caro, L. D., Lesmo, L., Rispoli, D., and Robaldo, L. (2012). Multi-label Classification of Legislative Text into EuroVoc. In Schäfer, B., editor, *Legal Knowledge and Information Systems - JURIX 2012: The Twenty-Fifth Annual Conference, University of Amsterdam, The Netherlands, 17-19 December 2012*, volume 250 of *Frontiers in Artificial Intelligence and Applications*, pages 21–30. IOS Press.

[Boer et al., 2002] Boer, A., Hoekstra, R., Winkels, R., Van Engers, T., and Willaert, F. (2002). Metalex: Legislation in XML. *Legal Knowledge and Information Systems (Jurix 2002)*, pages 1–10.

[Boer et al., 2008] Boer, A., Winkels, R., and Vitali, F. (2008). Metalex XML and the Legal Knowledge Interchange Format. In *Computable models of the law*, pages 21–41. Springer.

[Bonatti et al., 2018] Bonatti, P. A., Decker, S., Polleres, A., and Presutti, V. (2018). Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). *Dagstuhl Reports*, 8(9):29–111.

[Bonatti et al., 2020] Bonatti, P. A., Kirrane, S., Petrova, I. M., and Sauro, L. (2020). Machine understandable policies and gdpr compliance checking. *KI-Künstliche Intelligenz*, 34(3):303–315.

[Bontas et al., 2005] Bontas, E. P., Mochol, M., and Tolksdorf, R. (2005). Case studies on ontology reuse. In *Proceedings of the IKNOW05 International Conference on Knowledge Management*, volume 74, page 345. Citeseer.

[Borst, 1997] Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, University of Twente, Enschede, Netherlands.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

[Breuker et al., 2009] Breuker, J., Casanovas, P., Klein, M. A. C., and Francesconi, E. (2009). The Flood, the Channels and the Dykes: Managing Legal Information in a Globalized and Digital World. *Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, 188:3.

[Cardellino et al., 2017a] Cardellino, C., Teruel, M., Alemany, L. A., and Villata, S. (2017a). A low-cost, high-coverage legal named entity recognizer, classifier and linker. In Keppens, J. and Governatori, G., editors, *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017*, pages 9–18. ACM.

[Cardellino et al., 2017b] Cardellino, C., Teruel, M., Alemany, L. A., and Villata, S. (2017b). Ontology population and alignment for the legal domain: Yago, wikipedia and LKIF. In Nikitina, N., Song, D., Fokoue, A., and Haase, P., editors, *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017*, volume 1963 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Chalkidis et al., 2020a] Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020a). An empirical study on large-scale multi-label text classification including few and zero-shot labels. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7503–7515. Association for Computational Linguistics.

[Chalkidis et al., 2020b] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020b). LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

[Chalkidis et al., 2019] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., and Androutsopoulos, I. (2019). Large-scale multi-label text classification on EU legislation. *CoRR*, abs/1906.02192.

[Chalkidis et al., 2017] Chalkidis, I., Nikolaou, C., Soursos, P., and Koubarakis, M. (2017). Modeling and querying greek legislation using semantic web technologies. In Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., and Hartig, O., editors, *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I*, volume 10249 of *Lecture Notes in Computer Science*, pages 591–606.

[Chambers, 2013] Chambers, N. (2013). Navytime: Event and time ordering from raw text. In *Proceedings of the Workshop SemEval 2013*, volume 2, pages 73–77.

[Chambers et al., 2014] Chambers, N. et al. (2014). Dense event ordering with a multi-pass architecture. *Transactions of the ACL*, 2:273–284.

[Chang et al., 2012] Chang, A. X. et al. (2012). Sutime: A library for recognizing and normalizing time expressions. In *Proceedings of LREC 2012*.

[Chang et al., 2014] Chang, A. X. et al. (2014). TokensRegex: Defining cascaded regular expressions over tokens. Technical Report CSTR 2014-02, Department of Computer Science, Stanford University.

[Chelba et al., 2013] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., and Koehn, P. (2013). One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005.

[Chen and Manning, 2014] Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

[Chen and Lin, 2012] Chen, Y.-N. and Lin, H.-T. (2012). Feature-aware label space dimension reduction for multi-label classification. In *Advances in Neural Information Processing Systems*, pages 1529–1537.

[Chiticariu et al., 2013] Chiticariu, L., Li, Y., and Reiss, F. R. (2013). Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 827–832. ACL.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[Collins Dictionary, 2021] Collins Dictionary (2021). Language. `https://www.collinsdictionary.com/dictionary/english/language`. Online; accessed 2021-02-28.

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

[Corcho et al., 2003] Corcho, Ó., Fernández-López, M., Gómez-Pérez, A., and López-Cima, A. (2003). Building legal ontologies with METHONTOLOGY and webode. In Benjamins, V. R., Casanovas, P., Breuker, J., and Gangemi, A., editors, *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, volume 3369, pages 142–157. Springer Berlin Heidelberg.

[Council of the European Union, 2011] Council of the European Union (2011). Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52011XG0429(01)`. Online; accessed 2020-05-17.

[Council of the European Union, 2012] Council of the European Union (2012). Council conclusions inviting the introduction of the European Legislation Identifier (ELI). `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52012XG1026(01)`. Online; accessed 2020-05-17.

[Council of the European Union, 2017] Council of the European Union (2017). Council conclusions of 6 November 2017 on the European Legislation Identifier. `https://eur-lex.europa.eu/legal-content/GA/TXT/?uri=CELEX:52017XG1222(02)`. Online; accessed 2020-05-17.

[Cristani and Cuel, 2005] Cristani, M. and Cuel, R. (2005). A survey on ontology creation methodologies. *Int. J. Semantic Web Inf. Syst.*, 1(2):49–69.

[Cunningham et al., 1999] Cunningham, H., Cunningham, H., Maynard, D., Maynard, D., Tablan, V., and Tablan, V. (1999). JAPE: a Java Annotation Patterns Engine.

[Cunningham et al., 2013] Cunningham, H. et al. (2013). Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLOS Computational Biology*, 9(2):1–16.

[de Oliveira Rodrigues et al., 2019] de Oliveira Rodrigues, C. M., de Freitas, F. L. G., Barreiros, E. F. S., de Azevedo, R. R., and de Almeida Filho, A. (2019). Legal ontologies over time: A systematic mapping study. *Expert Systems with Applications*, 130:12 – 30.

[De Vos et al., 2019] De Vos, M., Kirrane, S., Padget, J., and Satoh, K. (2019). Odrl policy modelling and compliance checking. In *International Joint Conference on Rules and Reasoning*, pages 36–51. Springer.

[Dell'Orletta et al., 2012] Dell'Orletta et al. (2012). The splet–2012 shared task on dependency parsing of legal texts. In *Semantic Processing of Legal Texts (SPLeT-2012) Workshop Programme*, page 42.

[Derczynski et al., 2010] Derczynski, L. et al. (2010). Usfd2: Annotating temporal expresions and tlinks for tempeval-2. In *Proceedings of the Workshop SemEval*, pages 337–340. ACL.

[Devlin et al., 2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

[Distinto et al., 2016] Distinto, I., d'Aquin, M., and Motta, E. (2016). LOTED2: an ontology of european public procurement notices. *Semantic Web*, 7(3):267–293.

[Dozier and Haschart, 2000] Dozier, C. and Haschart, R. (2000). Automatic extraction and linking of person names in legal text. In Mariani, J. and Harman, D., editors, *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 2000, 6th International Conference, College de France, France, April 12-14, 2000. Proceedings*, pages 1305–1321. CID.

[Dozier et al., 2010] Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. (2010). Named entity recognition and resolution in legal text. In Francesconi, E., Montemagni, S., Peters, W., and Tiscornia, D., editors, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, volume 6036 of *Lecture Notes in Computer Science*, pages 27–43. Springer.

[Dragoni et al., 2017] Dragoni, M., Villata, S., Rizzi, W., and Governatori, G. (2017). Combining natural language processing approaches for rule extraction from legal documents. In Pagallo, U., Palmirani, M., Casanovas, P., Sartor, G., and Villata, S., editors, *AI Approaches to the Complexity of Legal Systems - AICOL International Workshops 2015-2017: AICOL-VI@JURIX 2015, AICOL-VII@EKAW 2016, AICOL-VIII@JURIX 2016, AICOL-IX@ICAIL 2017, and AICOL-X@JURIX 2017, Revised Selected Papers*, volume 10791 of *Lecture Notes in Computer Science*, pages 287–300. Springer.

[Ehrlinger and Wöß, 2016] Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. In Martin, M., Cuquet, M., and Folmer, E., editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Elluri et al., 2018] Elluri, L., Nagar, A., and Joshi, K. P. (2018). An integrated knowledge graph to automate GDPR and PCI DSS compliance. In Abe, N., Liu, H., Pu, C., Hu, X., Ahmed, N. K., Qiao, M., Song, Y., Kossmann, D., Liu, B., Lee, K., Tang, J., He, J., and Saltz, J. S., editors, *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 1266–1271. IEEE.

[Etiene et al., 2015] Etiene, T., Pagliosa, P. A., and Nonato, L. G. (2015). Linea: Building Timelines from Unstructured Text. In *28th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2015*, pages 234–241. IEEE Computer Society.

[European Court of Justice, 2014] European Court of Justice (2014). ECLI:EU:C:2014:2336. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX\%3A62013CJ0166`. accessed 2021-03-12.

[Färber et al., 2018] Färber, M., Bartscherer, F., Menne, C., and Rettinger, A. (2018). Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO. *Semantic Web*, 9(1):77–129.

[Felici, 2010] Felici, A. (2010). Translating eu law: legal issues and multiple dynamics. *Perspectives: Studies in Translatology*, 18(2):95–108.

[Fernandez et al., 1997] Fernandez, M., Gomez-Perez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 spring symposium series on ontological engineering*, pages 33–40. Stanford, USA.

[Ferro et al., 2005] Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2005). Tides 2005 standard for the annotation of temporal expressions. Technical report, Technical report, MITRE.

[Filannino and Nenadic, 2015] Filannino, M. and Nenadic, G. (2015). Temporal expression extraction with extensive feature type selection and a posteriori label adjustment. *Data & Knowledge Engineering*, 100:19 – 33.

[Filtz, 2017] Filtz, E. (2017). Building and processing a knowledge-graph for legal data. In Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., and Hartig, O., editors, *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part II*, volume 10250 of *Lecture Notes in Computer Science*, pages 184–194.

[Filtz et al., 2018] Filtz, E., Kirrane, S., and Polleres, A. (2018). Interlinking legal data. In Khalili, A. and Koutraki, M., editors, *Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems co-located with the 14th International Conference on Semantic Systems (SEMANTiCS 2018), Vienna, Austria, September 10-13, 2018*, volume 2198 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Filtz et al., 2021] Filtz, E., Kirrane, S., and Polleres, A. (2021). The linked legal data landscape: linking legal data across different countries. *Artificial Intelligence and Law*, pages 1–55.

[Filtz et al., 2019] Filtz, E., Kirrane, S., Polleres, A., and Wohlgenannt, G. (2019). Exploiting eurovoc's hierarchical structure for classifying legal documents. In Panetto, H., Debruyne, C., Hepp, M., Lewis, D., Ardagna, C. A., and Meersman, R., editors, *On the Move to Meaningful Internet Systems: OTM 2019 Conferences - Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21-25, 2019, Proceedings*, volume 11877 of *Lecture Notes in Computer Science*, pages 164–181. Springer.

[Filtz et al., 2020] Filtz, E., Navas-Loro, M., Santos, C., Polleres, A., and Kirrane, S. (2020). Events matter: Extraction of events from court decisions. In Serena, V., Harasta, J., and Kremen, P., editors, *Legal Knowledge and Information Systems - JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*, volume 334 of *Frontiers in Artificial Intelligence and Applications*, pages 33–42. IOS Press.

[Francart et al., 2018] Francart, T., Dann, J., Pappalardo, R., Malagon, C., and Pellegrino, M. (2018). The european legislation identifier. In Peruginelli, G. and Faro, S., editors, *Knowledge of the Law in the Big Data Age, Conference 'Law via the Internet 2018', Florence, Italy, 11-12 October 2018*, volume 317 of *Frontiers in Artificial Intelligence and Applications*, pages 137–148. IOS Press.

[Francesconi et al., 2015] Francesconi, E., Küster, M. W., Gratz, P., and Thelen, S. (2015). The ontology-based approach of the publications office of the EU for document accessibility and open data services. In Ko, A. and Francesconi, E., editors, *Electronic Government and the Information Systems Perspective - 4th International Conference, EGOVIS 2015, Valencia, Spain, September 1-3, 2015, Proceedings*, volume 9265 of *Lecture Notes in Computer Science*, pages 29–39. Springer.

[Francesconi et al., 2010] Francesconi, E., Montemagni, S., Peters, W., and Tiscornia, D., editors (2010). *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, volume 6036 of *Lecture Notes in Computer Science*. Springer.

[Freund et al., 1999] Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

[Fulda et al., 2016] Fulda, J., Brehmer, M., and Munzner, T. (2016). TimeLineCurator: Interactive Authoring of Visual Timelines from Unstructured Text. *IEEE Trans. Vis. Comput. Graph.*, 22(1):300–309.

[Gangemi, 2005] Gangemi, A. (2005). Ontology design patterns for semantic web content. In Gil, Y., Motta, E., Benjamins, V. R., and Musen, M. A., editors, *The Semantic Web - ISWC 2005, 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005, Proceedings*, volume 3729 of *Lecture Notes in Computer Science*, pages 262–276. Springer.

[Gangemi, 2007] Gangemi, A. (2007). Design Patterns for Legal Ontology Constructions. *LOAIT*, 2007:65–85.

[Gangemi and Presutti, 2009] Gangemi, A. and Presutti, V. (2009). Ontology design patterns. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, pages 221–243. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Genesereth, 2018] Genesereth, M. (2018). Computational Law. The Cop in the Backseat. `http://logic.stanford.edu/publications/genesereth/complaw.pdf`. accessed: 01.11.2020.

[Genesereth and Nilsson, 1988] Genesereth, M. R. and Nilsson, N. J. (1988). *Logical foundations of artificial intelligence*. Morgan Kaufmann.

[Ghosh et al., 2016] Ghosh, M. E., , Naja, H., Abdulrab, H., and Khalil, M. (2016). Towards a middle-out approach for building legal domain reference ontology. *International Journal of Knowledge Engineering*, 2(3):109–114.

[Gildea and Jurafsky, 2002] Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

[Gracia et al., 2017] Gracia, J., Kernerman, I., and Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pages 19–21.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message understanding conference- 6: A brief history. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pages 466–471.

[Gruber, 1993] Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge acquisition*, 5(2):199–220.

[Guarino et al., 2009] Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 1–17. Springer.

[Guo et al., 2020] Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., and He, Q. (2020). A survey on knowledge graph-based recommender systems. *CoRR*, abs/2003.00911.

[Han et al., 2006] Han, B., Gates, D., and Levin, L. (2006). From language to time: A temporal expression anchorer. In *Temporal Representation and Reasoning, 2006. TIME 2006. Thirteenth International Symposium on*, pages 196–203. IEEE.

[Heist et al., 2020] Heist, N., Hertling, S., Ringler, D., and Paulheim, H. (2020). Knowledge graphs on the web - an overview. *CoRR*, abs/2003.00719.

[Hellmann et al., 2012] Hellmann, S., Lehmann, J., and Auer, S. (2012). NIF: An ontology-based and linked-data-aware NLP Interchange Format. Working Draft.

[Hirschberg and Manning, 2015] Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

[Hitzler et al., 2010] Hitzler, P., Krötzsch, M., and Rudolph, S. (2010). *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press.

[Hoekstra et al., 2007] Hoekstra, R., Breuker, J., Bello, M. D., and Boer, A. (2007). The LKIF core ontology of basic legal concepts. In Casanovas, P., Biasiotti, M. A., Francesconi, E., and Sagri, M., editors, *Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques June 4th, 2007, Stanford University, Stanford, CA, USA*, volume 321 of *CEUR Workshop Proceedings*, pages 43–63. CEUR-WS.org.

[Hogan et al., 2020] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J. E. L., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngomo, A.-C. N., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2020). Knowledge graphs.

[Hogenboom et al., 2011] Hogenboom, F., Frasincar, F., Kaymak, U., and De Jong, F. (2011). An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011)*, volume 779, pages 48–57. Koblenz, Germany: CEUR-WS. org.

[Hotho et al., 2005] Hotho, A., Nürnberger, A., and Paass, G. (2005). A brief survey of text mining. *LDV Forum*, 20(1):19–62.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

[International Federation of Library Associations and Institutions, 2009] International Federation of Library Associations and Institutions (2009). Functional Requirements for Bibliographic Records. `https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf`. Online; accessed 2020-11-27.

[International Federation of Library Associations and Institutions, 2016] International Federation of Library Associations and Institutions (2016). Definition of FRBRoo. A Conceptual Model for Bibliographic Information in Object-Oriented Formalism. `https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf`. Online; accessed 2020-11-27.

[Iqbal et al., 2013] Iqbal, R., Murad, M. A. A., Mustapha, A., Sharef, N. M., et al. (2013). An analysis of ontology engineering methodologies: A literature review. *Research journal of applied sciences, engineering and technology*, 6(16):2993–3000.

[Isemann et al., 2013] Isemann, D. et al. (2013). *Temporal Dependence in Legal Documents*, pages 497–504. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Jacovi et al., 2018] Jacovi, A., Shalom, O. S., and Goldberg, Y. (2018). Understanding convolutional neural networks for text classification. In Linzen, T., Chrupala, G., and Alishahi, A., editors, *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 56–65. Association for Computational Linguistics.

[Ji et al., 2014] Ji, H., Cassidy, T., Li, Q., and Tamang, S. (2014). Tackling representation, annotation and classification challenges for temporal knowledge base population. *Knowledge and Information Systems*, 41(3):611–646.

[Jones et al., 1998] Jones, D., Bench-Capon, T., and Visser, P. (1998). Methodologies for ontology development. *IFIP world computer congress*, pages 62–75.

[Jones, 2004] Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5):493–502.

[Jung et al., 2011] Jung, H., Allen, J., Blaylock, N., De Beaumont, W., Galescu, L., and Swift, M. (2011). Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 workshop*, pages 146–154. ACL.

[Junior et al., 2020] Junior, A. C., Orlandi, F., Graux, D., Hossari, M., O'Sullivan, D., Hartz, C., and Dirschl, C. (2020). Knowledge graph-based legal search over german court cases. In Harth, A., Presutti, V., Troncy, R., Acosta, M., Polleres, A., Fernández, J. D., Parreira, J. X., Hartig, O., Hose, K., and Cochez, M., editors, *The Semantic*

Web: ESWC 2020 Satellite Events - ESWC 2020 Satellite Events, Heraklion, Crete, Greece, May 31 - June 4, 2020, Revised Selected Papers, volume 12124 of *Lecture Notes in Computer Science*, pages 293–297. Springer.

[Junior et al., 2019] Junior, A. C., Orlandi, F., O'Sullivan, D., Dirschl, C., and Reul, Q. (2019). Using mapping languages for building legal knowledge graphs from XML files. In Samavi, R., Consens, M. P., Khatchadourian, S., Nguyen, V., Sheth, A. P., Giménez-García, J. M., and Thakkar, H., editors, *Proceedings of the Blockchain enabled Semantic Web Workshop (BlockSW) and Contextualized Knowledge Graphs (CKG) Workshop co-located with the 18th International Semantic Web Conference, BlockSW/CKG@ISWC 2019, Auckland, New Zealand, October 27, 2019*, volume 2599 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Kajiwara et al., 2016] Kajiwara, T. et al. (2016). Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016: Technical Papers*, pages 1147–1158.

[Kiyavitskaya et al., 2008] Kiyavitskaya, N., Zeni, N., Breaux, T. D., Antón, A. I., Cordy, J. R., Mich, L., and Mylopoulos, J. (2008). Automating the extraction of rights and obligations for regulatory compliance. In *International Conference on Conceptual Modeling*, pages 154–168. Springer.

[Kolya et al., 2013] Kolya, A. K., Kundu, A., Gupta, R., Ekbal, A., and Bandyopadhyay, S. (2013). Ju_cse: A crf based approach to annotation of temporal expression, event and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 64–72.

[Kucera et al., 2015] Kucera, J., Chlapek, D., Klímek, J., and Necaský, M. (2015). Methodologies and best practices for open data publication. In Necaský, M., Pokorný, J., and Moravec, P., editors, *Proceedings of the Dateso 2015 Annual International Workshop on DAtabases, TExts, Specifications and Objects, Neprivec u Sobotky, Jicin, Czech Republic, April 14, 2015*, volume 1343 of *CEUR Workshop Proceedings*, pages 52–64. CEUR-WS.org.

[Lafferty et al., 2001] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Brodley, C. E. and Danyluk, A. P., editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

[Lagos et al., 2010] Lagos, N., Segond, F., Castellani, S., and O'Neill, J. (2010). Event extraction for legal case building and reasoning. In *International Conference on Intelligent Information Processing*, pages 92–101. Springer.

[Lample et al., 2016] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition.

[Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nat.*, 521(7553):436–444.

[Lee et al., 2014] Lee, K. et al. (2014). Context-dependent semantic parsing for time expressions. In *Proceedings of the 52nd Annual Meeting of the ACL*, volume 1, pages 1437–1447.

[Lehmann et al., 2015] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

[Leitner et al., 2019] Leitner, E., Rehm, G., and Schneider, J. M. (2019). Fine-grained named entity recognition in legal documents. In Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., and Sure-Vetter, Y., editors, *Semantic Systems. The Power of AI and Knowledge Graphs - 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings*, volume 11702 of *Lecture Notes in Computer Science*, pages 272–287. Springer.

[Leone et al., 2019] Leone, V., Di Caro, L., and Villata, S. (2019). Taking stock of legal ontologies: a feature-based comparative analysis. *Artificial Intelligence and Law*.

[Lilleberg et al., 2015] Lilleberg, J., Zhu, Y., and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In Ge, N., Lu, J., Wang, Y., Howard, N., Chen, P., Tao, X., Zhang, B., and Zadeh, L. A., editors, *14th IEEE International Conference on Cognitive Informatics & Cognitive Computing, ICCI*CC 2015, Beijing, China, July 6-8, 2015*, pages 136–140. IEEE Computer Society.

[Liu and Chen, 2015] Liu, S. M. and Chen, J. (2015). A multi-label classification based approach for sentiment classification. *Expert Syst. Appl.*, 42(3):1083–1093.

[Llorens et al., 2010] Llorens, H. et al. (2010). Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the Workshop SemEval*, pages 284–291. ACL.

[Llorens et al., 2012] Llorens, H. et al. (2012). Timen: An open temporal expression normalisation resource. In *Proceedings of LREC 2012*, pages 3044–3051.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. *CoRR*, cs.CL/0205028.

[Loza Mencía and Fürnkranz, 2010] Loza Mencía, E. and Fürnkranz, J. (2010). Efficient multilabel classification algorithms for large-scale problems in the legal domain. In [Francesconi et al., 2010], pages 192–215.

[Mani et al., 2000] Mani, I. et al. (2000). Robust temporal processing of news. In *Proceedings of the 38th annual meeting on ACL*, pages 69–76. ACL.

[Manning et al., 2014] Manning, C. D. et al. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the ACL 2014, System Demonstrations*, pages 55–60.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

[Mansouri et al., 2008] Mansouri, A., Affendey, L. S., and Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344.

[María Navas-Loro, 2018] María Navas-Loro, C. S. (2018). Events in the legal domain: first impressions. 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 45–57.

[Maxwell et al., 2009] Maxwell, K. T., Oberlander, J., and Lavrenko, V. (2009). Evaluation of semantic events for legal case retrieval. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 39–41. ACM.

[Mazur and Dale, 2009] Mazur, P. and Dale, R. (2009). The dante temporal expression tagger. In Vetulani, Z. and Uszkoreit, H., editors, *Human Language Technology. Challenges of the Information Society*, pages 245–257, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Mazur et al., 2010] Mazur, P. et al. (2010). Wikiwars: A new corpus for research on temporal expressions. In *Proceedings of EMNLP 2010*, pages 913–922. ACL.

[McCrae et al., 2017] McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference, 2017*, pages 19–21.

[Mikheev et al., 1999] Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 1–8. The Association for Computer Linguistics.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

[Minard et al., 2016] Minard, A.-L., Speranza, M., et al. (2016). Meantime, the newsreader multilingual event and time corpus. In *Proceedings of LREC 2016, European Language Resources Association.*

[Montiel-Ponsoda et al., 2017] Montiel-Ponsoda, E., Rodríguez-Doncel, V., and Gracia, J. (2017). Building the legal knowledge graph for smart compliance services in multilingual europe. In Rodríguez-Doncel, V., Casanovas, P., and González-Conejero, J., editors, *Proceedings of the 1st Workshop on Technologies for Regulatory Compliance co-located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017), Luxembourg, December 13, 2017*, volume 2049 of *CEUR Workshop Proceedings*, pages 15–17. CEUR-WS.org.

[Muñoz et al., 2009] Muñoz, S., Pérez, J., and Gutiérrez, C. (2009). Simple and efficient minimal RDFS. *J. Web Semant.*, 7(3):220–234.

[Muñoz-Soro et al., 2016] Muñoz-Soro, J. F., Esteban, G., Corcho, Ó., and Serón, F. J. (2016). Pproc, an ontology for transparency in public procurement. *Semantic Web*, 7(3):295–309.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Naik et al., 2011] Naik, V., Vanitha, G., and Inturi, S. (2011). Reasoning in legal text documents with extracted event information. *International Journal of Computer Applications*, 28:8—13.

[Navas-Loro et al., 2019] Navas-Loro, M., Filtz, E., Rodríguez-Doncel, V., Polleres, A., and Kirrane, S. (2019). Tempcourt: evaluation of temporal taggers on a new corpus of court decisions. *The Knowledge Engineering Review*, 34:e24.

[Navas-Loro and Rodríguez-Doncel, 2020] Navas-Loro, M. and Rodríguez-Doncel, V. (2020). Annotador: a temporal tagger for spanish. *Journal of Intelligent & Fuzzy Systems*, 39:1979–1991. 2.

[Neumaier and Polleres, 2019] Neumaier, S. and Polleres, A. (2019). Enabling spatio-temporal search in open data. *Journal of Web Semantics*, 55:21–36.

[Nivre et al., 2007] Nivre, J. et al. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL 2007*.

[Noble, 2006] Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

[Northwood, 2010] Northwood, C. (2010). TERNIP: temporal expression recognition and normalisation in Python. Master's thesis, University of Sheffield.

[Noy et al., 2001] Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology.

[Oksanen et al., 2019] Oksanen, A., Tamper, M., Tuominen, J., Mäkelä, E., Hietanen, A., and Hyvönen, E. (2019). Semantic finlex: Transforming, publishing, and using finnish legislation and case law as linked open data on the web. *Knowledge of the Law in the Big Data Age*, 317:212–228.

[Oltramari et al., 2018] Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T. B., Russell, N. C., Story, P., Reidenberg, J. R., and Sadeh, N. M. (2018). Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web*, 9(2):185–203.

[Palmirani et al., 2011] Palmirani, M., Governatori, G., and Contissa, G. (2011). Modelling Temporal Legal Rules. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ICAIL '11, pages 131–135. ACM.

[Palmirani et al., 2018] Palmirani, M., Martoni, M., Rossi, A., Bartolini, C., and Robaldo, L. (2018). Pronto: Privacy ontology for legal reasoning. In Ko, A. and Francesconi, E., editors, *Electronic Government and the Information Systems Perspective - 7th International Conference, EGOVIS 2018, Regensburg, Germany, September 3-5, 2018, Proceedings*, volume 11032 of *Lecture Notes in Computer Science*, pages 139–152. Springer.

[Palmirani and Vitali, 2011] Palmirani, M. and Vitali, F. (2011). Akoma-ntoso for legal documents. In Sartor, G., Palmirani, M., Francesconi, E., and Biasiotti, M. A., editors, *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*, pages 75–100. Springer Netherlands, Dordrecht.

[Pandit et al., 2018] Pandit, H. J., Fatema, K., O'Sullivan, D., and Lewis, D. (2018). Gdprtext - GDPR as a linked data resource. In Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., and Alam, M., editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 481–495. Springer.

[Paulheim, 2017] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508.

[Pellow et al., 2014] Pellow, D. et al. (2014). An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 84–93.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M. A., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

[Pinto and Martins, 2004] Pinto, H. S. A. N. P. and Martins, J. P. (2004). Ontologies: How can they be built? *Knowl. Inf. Syst.*, 6(4):441–464.

[Prabhu and Varma, 2014] Prabhu, Y. and Varma, M. (2014). Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–272. ACM.

[Presutti and Gangemi, 2008] Presutti, V. and Gangemi, A. (2008). Content ontology design patterns as practical building blocks for web ontologies. In Li, Q., Spaccapietra, S., Yu, E. S. K., and Olivé, A., editors, *Conceptual Modeling - ER 2008, 27th International Conference on Conceptual Modeling, Barcelona, Spain, October 20-24, 2008. Proceedings*, volume 5231 of *Lecture Notes in Computer Science*, pages 128–141. Springer.

[Publications Office of the European Union, 2020a] Publications Office of the European Union (2020a). Common Data Model (CDM). `https://op.europa.eu/en/web/eu-vocabularies/cdm/`. Online; accessed 2020-11-27.

[Publications Office of the European Union, 2020b] Publications Office of the European Union (2020b). European Legislation Identifier (ELI). `https://op.europa.eu/en/web/eu-vocabularies/eli/`. Online; accessed 2020-11-27.

[Publications Office of the European Union, 2021] Publications Office of the European Union (2021). Eurovoc thesaurus. `https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc?target=Documentation`. Online; accessed 2021-01-02.

[Pustejovsky et al., 2003a] Pustejovsky, J. et al. (2003a). The Timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

[Pustejovsky et al., 2003b] Pustejovsky, J. et al. (2003b). TimeML: Robust Specification of Event and Temporal Expressions in Text. In Maybury, M. T., editor, *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium*, pages 28–34. AAAI Press.

[Quaresma and Gonçalves, 2010] Quaresma, P. and Gonçalves, T. (2010). Using linguistic information and machine learning techniques to identify entities from juridical documents. In [Francesconi et al., 2010], pages 44–59.

[Quemy, 2018] Quemy, A. (2018). European court of human right open data project. *CoRR*, abs/1810.03115.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[Registry of the Court, 2020] Registry of the Court (2020). `https://www.echr.coe.int/documents/rules_court_eng.pdf`. accessed 2020-09-14.

[Rodríguez-Doncel et al., 2016] Rodríguez-Doncel, V., Delgado, J., Llorente, S., Rodríguez, E., and Boch, L. (2016). Overview of the MPEG-21 media contract ontology. *Semantic Web*, 7(3):311–332.

[Rodríguez-Doncel et al., 2013] Rodríguez-Doncel, V., Suárez-Figueroa, M. C., Gómez-Pérez, A., and Poveda-Villalón, M. (2013). License linked data resources pattern. In Gangemi, A., Gruninger, M., Hammar, K., Lefort, L., Presutti, V., and Scherp, A., editors, *Proceedings of the 4th Workshop on Ontology and Semantic Web Patterns co-located with 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013*, volume 1188 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Romera-Paredes and Torr, 2015] Romera-Paredes, B. and Torr, P. H. S. (2015). An embarrassingly simple approach to zero-shot learning. In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2152–2161. JMLR.org.

[Rotmensch et al., 2017] Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., and Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):1–11.

[Sanh et al., 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

[Santos et al., 2016] Santos, C., Pruski, C., Silveira, M. D., Rodríguez-Doncel, V., Gangemi, A., van der Torre, L., and Casanovas, P. (2016). Complaint ontology pattern - COP. In Hammar, K., Hitzler, P., Krisnadhi, A., Lawrynowicz, A., Nuzzolese, A. G., and Solanki, M., editors, *Advances in Ontology Design and Patterns [revised and extended versions of the papers presented at the 7th edition of the Workshop on Ontology and Semantic Web Patterns, WOP@ISWC 2016, Kobe, Japan, 18th October 2016]*, volume 32 of *Studies on the Semantic Web*, pages 69–83. IOS Press.

[Saurí et al., 2006] Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). TimeML annotation guidelines. *Version*, 1(1):31.

[Schilder, 2005] Schilder, F. (2005). Event Extraction and Temporal Reasoning in Legal Documents. In *Annotating, Extracting and Reasoning about Time and Events, International Seminar, Dagstuhl Castle. Revised Papers*, pages 59–71.

[Scott, 1955] Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

[Shaheen et al., 2020] Shaheen, Z., Wohlgenannt, G., and Filtz, E. (2020). Large-scale legal text classification using transformer models. In vor der Brück, T., editor, *SEMAPRO 2020 The Fourteenth International Conference on Advances in Semantic Processing, 2020, Nice, France, October 25-29, 2020.*, pages 7–17. IARIA.

[Simperl and Luczak-Rösch, 2014] Simperl, E. and Luczak-Rösch, M. (2014). Collaborative ontology engineering: a survey. *Knowl. Eng. Rev.*, 29(1):101–131.

[Spatial Data on the Web Working Group, 2020] Spatial Data on the Web Working Group (2020). Time Ontology in OWL W3C Candidate Recommendation 26 March 2020. `https://www.w3.org/TR/owl-time/`. Online; accessed 2021-03-12.

[Stadler et al., 2012] Stadler, C., Lehmann, J., Höffner, K., and Auer, S. (2012). Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354.

[Steedman et al., 2011] Steedman, M. et al. (2011). *Combinatory Categorial Grammar*, chapter 5, pages 181–224. Wiley-Blackwell.

[Steinberger et al., 2012] Steinberger, R., Ebrahim, M., and Turchi, M. (2012). JRC eurovoc indexer JEX - A freely available multi-label categorisation tool. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 798–805. European Language Resources Association (ELRA).

[Steinberger et al., 2006] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 2142–2147. European Language Resources Association (ELRA).

[Stolze, 2001] Stolze, R. (2001). Translating legal texts in the eu. *Perspectives: studies in translatology*, 9(4):301–311.

[Strötgen et al., 2012] Strötgen, J. et al. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of LREC 2012*, volume 12, pages 3746–3753.

[Strötgen and Gertz, 2016] Strötgen, J. and Gertz, M. (2016). Domain-sensitive temporal tagging. *Synthesis Lectures on Human Language Technologies*, 9(3):1–151.

[Studer et al., 1998] Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data Knowl. Eng.*, 25(1-2):161–197.

[Styler IV et al., 2014] Styler IV, W. et al. (2014). Temporal annotation in the clinical domain. *Transactions of the Association of Computational Linguistics*, 2(1):143–154.

[Tabassum et al., 2016] Tabassum, J. et al. (2016). Tweetime: A minimally supervised method for recognizing and normalizing time expressions in twitter. *arXiv preprint arXiv:1608.02904*.

[Tannier and Vernier, 2016] Tannier, X. and Vernier, F. (2016). Creation, Visualization and Edition of Timelines for Journalistic Use. In *Proceedings of Natural Language meets Journalism Workshop at IJCAI*.

[Tsoumakas and Katakis, 2007] Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007:1–13.

[Tuggener et al., 2020] Tuggener, D., von Däniken, P., Peetz, T., and Cieliebak, M. (2020). LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1235–1241. European Language Resources Association.

[Uschold and Gruninger, 1996] Uschold, M. and Gruninger, M. (1996). Ontologies: principles, methods and applications. *Knowl. Eng. Rev.*, 11(2):93–136.

[Uzzaman et al., 2010] Uzzaman, N. et al. (2010). Event and Temporal Expression Extraction from Raw Text: First Step Towards a Temporally Aware System. *International Journal of Semantic Computing*, 04(04):487–508.

[UzZaman et al., 2013] UzZaman, N. et al. (2013). SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the Workshop SemEval 2013*, pages 1–9.

[v. d. L. Gardner, 1983] v. d. L. Gardner, A. (1983). The design of a legal analysis program. In Genesereth, M. R., editor, *Proceedings of the National Conference on Artificial Intelligence, Washington, D.C., USA, August 22-26, 1983*, pages 114–118. AAAI Press.

[van der Vet and Mars, 1998] van der Vet, P. E. and Mars, N. J. I. (1998). Bottom-up construction of ontologies. *IEEE Trans. Knowl. Data Eng.*, 10(4):513–526.

[van Opijnen et al., 2017a] van Opijnen, M., Palmirani, M., Vitali, F., van den Oever, J., and Agnoloni, T. (2017a). Towards ECLI 2.0. In Parycek, P. and Edelmann, N., editors, *2017 Conference for E-Democracy and Open Government, CeDEM 2017, Krems, Austria, May 17-19, 2017*, pages 135–143. IEEE Computer Society.

[van Opijnen et al., 2017b] van Opijnen, M., Peruginelli, G., Kefali, E., and Palmirani, M. (2017b). On-line Publication of Court Decisions in the EU. Report of the policy group of the project 'building on the european case law identifier', BO-ECLI.

[van Opijnen and Santos, 2017] van Opijnen, M. and Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artif. Intell. Law*, 25(1):65–87.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

[Vegetti et al., 2011] Vegetti, M., Leone, H. P., and Henning, G. P. (2011). PRONTO: an ontology for comprehensive and consistent representation of product information. *Eng. Appl. Artif. Intell.*, 24(8):1305–1327.

[Verhagen et al., 2005] Verhagen, M. et al. (2005). Automating Temporal Annotation with TARSQI. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, pages 81–84. ACL.

[Vlek et al., 2013] Vlek, C. S. et al. (2013). Representing and evaluating legal narratives with subscenarios in a bayesian network. In *OASIcs-OpenAccess Series in Informatics*, volume 32. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[Vrandecic and Krötzsch, 2014] Vrandecic, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

[W3C Community Group, 2012] W3C Community Group (2012). JavaScript Object Notation for Linking Data (JSON-LD). `https://www.w3.org/community/json-ld/`. Online; accessed 2020-05-17.

[W3C Community Group, 2021] W3C Community Group (2021). EPUB 3. `https://www.w3.org/publishing/groups/epub3-cg/`. Online; accessed 2021-02-05.

[W3C Ontology-Lexica Community Group, 2016] W3C Ontology-Lexica Community Group (2016). Lexicon Model for Ontologies: Final Community Group Report, 10 May 2016. `https://www.w3.org/2016/05/ontolex/`. Online; accessed 2020-11-27.

[W3C Recommendation, 2012] W3C Recommendation (2012). A Direct Mapping of Relational Data to RDF. `https://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/`. Online; accessed 2020-05-17.

[W3C Working Group, 2014] W3C Working Group (2014). `https://www.w3.org/TR/rdf-schema/`. Online; accessed 2020-09-14.

[Wolf et al., 2019] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv e-prints*, page arXiv:1910.03771.

[World Justice Project, 2019] World Justice Project (2019). Global insights on access to justice. `https://worldjusticeproject.org/sites/default/files/documents/WJP-A2J-2019.pdf`. accessed 2020-09-14.

[Wyner and Peters, 2010] Wyner, A. Z. and Peters, W. (2010). Lexical semantics and expert legal knowledge towards the identification of legal case factors. In *JURIX*, volume 10, pages 127–136.

[Xiang and Wang, 2019] Xiang, W. and Wang, B. (2019). A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

[Zhang et al., 2017] Zhang, W., Wang, L., Yan, J., Wang, X., and Zha, H. (2017). Deep extreme multi-label learning. *CoRR*, abs/1704.03718.

[Zhong et al., 2017] Zhong, X. et al. (2017). Time expression analysis and recognition using syntactic token types and general heuristic rules. In *Proceedings of the 55th Annual Meeting of the ACL*, volume 1, pages 420–429.