UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI INFORMATICA

SCUOLA DI DOTTORATO IN INFORMATICA
XXXIII CICLO

Law and Data Science: Knowledge Modeling and Extraction from Court Decisions

INF/01 INFORMATICA

Mattia Falduti

Supervisor: Prof. Silvana Castano

Assistant supervisor: Prof. Alfio Ferrara

Headmaster of the Ph.D. School: Prof. Paolo Boldi

Academic Year 2019/2020

Die abschließenden Entscheidungen des Gerichts werden nicht veroffentlicht, sie sind nicht einmal den Richtern zuganglich, infolgedessen haben sich uber alte Gerichtsfalle nur Legenden erhalten

[Franz Kafka]

Abstract

In the era of big data, research activity on data science focuses on the collection, processing, and interpretation of large datasets to produce knowledge for decision-making processes in different application domains and contexts. The legal domain is one of these different domains and contexts where data science approaches can be applied. Indeed, thousands of legal documents are constantly produced by institutional bodies, such as Parliaments and Courts, where daily law and court decisions (CDs) are published. Both law and CDs constitute prominent sources of knowledge. Law is general by definition and adopts abstract terminology that creates areas of uncertainty, whereas CDs provide a lot of concrete information about the application of the law. For this reason, legal interpreters, such as judges, public prosecutors and lawyers, are daily involved in the analysis and evaluation of CDs, that currently is mostly run manually by domain experts. Knowledge extraction approaches have been applied to legal texts, but the literature confirms that the application of such techniques on CDs requires unique capabilities, due to the domain language and practice. More recently, many machine learning approaches applied on CDs datasets require expensive tasks of manual annotation, in terms of time, costs, expertise, and agreement, implying also several ethical issues.

In the Thesis, we address these issues by proposing CRIKE (CRIme Knowledge Extraction) a data science approach conceived to support legal knowledge extraction from a corpus of CDs documents, based on a reference legal domain ontology (LATO). First, we introduce the legal knowledge model of LATO, that captures and conceptually formalizes the features and nature of terminology used in law and CDs as entities and relationships, implemented in LATO using SKOS. Then, we present CRIKE that aims to progressively enrich the knowledge specified in a reference LATO ontology by extracting new concrete terminology associated with legal ontology concepts, as it occurs in the corpus. Knowledge extraction in CRIKE is based on multi-label annotation techniques where the corpus of annotated CDs is built by relying on the ontology contents without the need of manual annotation. Information retrieval techniques are applied for discovering new terms to populate term-sets of legal concepts that have been recognized in the CD texts. To evaluate the results obtained through the CRIKE approach, we discuss experimental results of application of CRIKE to a dataset of 180,000 CDs of the State of Illinois taken from the Caselaw Access Project (CAP) that provides public access to U.S. CDs digitized from the collection of the Harvard Law Library. Finally, we discuss the applicability conditions and the ethical issues related to CRIKE access, storage and processing tasks.

Contents

List of Figures

Chapter 1

Introduction

Creation, publication and communication of law had changed through history, with only one exception. Communicating laws requires the use of language, mostly written. In the last centuries law and the related application have been reported in documents. But law evolves with society. Simple society structures have simple law and simple legal disputes, whereas complex social structures have complex law and lots of complex legal disputes[1]. These are resolved with lots of decisions, reported in documents, daily produced by courts. The set of these documents is constantly (and inevitably) growing, creating big volumes of data.

In the era of big data, research activity on data science focuses on the collection, processing, and interpretation of large collections of data to produce knowledge for decision-making processes in different application domains and contexts. This is stimulated, on one side, and made possible on the other side, by the continuous production of data coming from disparate data sources and locations and by the availability of web-based technologies for data storage, integration, analysis and mining, thus enabling behavior and trend prediction as well as descriptive statistics for facts and events. The legal domain is one of these different domains and contexts where data science approaches can be applied. Indeed, many legal documents are constantly produced by institutional bodies, such as Parliaments, Governments, and Courts. Examples are law and case law, where the law is the set of rules which govern human conduct

[1] On comprehension of law, Lee Loevinger wrote in [Loevinger, 1971] that *"one of the greatest anomalies of modern times is that the law, which exists as a public guide to conduct, has become such a recondite mystery that is incomprehensible to the public and scarcely intelligible to its own votaries"*.

and case law is the set of the Court Decisions (CDs), which are the official documents that provide a specific interpretation of law associated with facts featuring a considered single case application.

In civil law countries, the law is the large set of rules, written by the legislative powers, to govern human conduct. But, human conduct is incalculable and unforeseeable. It is impossible to rule on every single possible scenario, just because real-life is too complex and unpredictable to be completely foreseen in legal structures. However, the law needs to be applicable to unpredictable new cases and uncountable future events. To achieve this goal, the solution for legislators has been to include a general and abstract terminology in the legal jargon. But sometimes, such terminology creates areas of uncertainty.

From our perspective, the abstract terminology used by legislators is associated with abstract concepts. These concepts have been already analyzed in the literature and referred using multiple expressions, such as the *indeterminacies in legal rules* as addressed in Endicott [1996] or the *penumbra [of law], where the controversy begins* as recalled in Cardozo [1925]. Moreover, *legal indeterminacy of law* is discussed in Kress [1989] and the *vague expressions* in natural language and their inevitable uses in legal texts are reported in Geert and Poscher [2016]. A main contribution on the topic is presented in Hart [1961], where the author describes *the open texture of law* as the areas of conduct where much must be left to be developed by courts or officials striking a balance, in the light of circumstances, between competing interests which vary in weight from case to case. In other words, interpretation of abstract concepts written in the law is reported in CDs.

Case law is the set of Court Decisions (CDs) written by the jurisdictional powers that provide a concrete application of rules and concepts, by deciding whether the law has been violated in relation to some facts or human conduct. To this end, CDs are written using specific and concrete terminology, in that they provide a contextualized, case-oriented interpretation of law deriving from the way judges/lawyers decide to apply the law statements to the specific circumstances/situation of the case at hand. Furthermore, contrary to law, CDs do not need to be applied to several new cases and future events. Indeed, CDs consider specific and concrete facts. For these reasons, CDs are the documents that make clear the areas of uncertainty in the law, described above. This role is proven by real-world data. The lack of certainty in law is largely

proven by the millions of claims and legal actions daily deposited in courts[2]. The effect is the production of millions of CDs. Some of them are devoted to define legal abstract concepts, where concrete meanings of legal abstract concepts are fully defined by referring to the specific terminology that appears in real CDs. Compared to laws, CDs report much more (legal) information. Therefore, CDs are a core component of the legal system since a clear and exhaustive understanding of them influences everyone everyday life. However, quantity, complexity, and articulation of CDs are constantly growing.

Both law and case law constitute prominent knowledge sources to be considered for the knowledge-based evaluation and judgment of a new case, in that they provide the general legal framework (law) and the specific interpretations (case law) adopted for already processed cases. When a new case is received for judgment, the knowledge-based evaluation process takes into account relevant legal knowledge to support CDs definition, that is, knowledge deriving from i) the law, for understanding the general rules that are relevant/prominent for the current case and ii) the case-law, for detecting possible relevant interpretations of law terminology in history of similar CDs.

Motivation. CDs provide relevant information about the application of the law and legal interpreters, such as judges, public prosecutors and lawyers, are daily involved in analysis and evaluation of them. Also policymakers, public administrators, enterprises, self-employed as well as individuals are involved in searching and possibly exploiting legal knowledge in CDs related to their area of interest. Academic institutes and research centers are addressing research on CDs analysis, classification and learning. For example, the Oxford Reports on International Law in Domestic Courts (ILDC)[3] selects CDs which are relevant to the identification and interpretation of rules of international law, as applied in the domestic courts of around 70 jurisdictions[4]. The Human Rights Centre of Ghent University[5] developed a project for studying the European Court of Human Rights case law with the aim of proposing innovative solutions

[2]Consider for instance that the last European Commission for the Efficiency of Justice (CEPEJ) 2018 report considered the data of 2016, and counted the sum of 22.215.201 new cases in France, Germany Italy, Spain and the UK.

[3]https://opil.ouplaw.com/page/212, last visit 8 May 2020

[4]Their aim is to report case law not only from states but also from certain territorial entities that are not generally classified as states.

[5]https://hrc.ugent.be, last visit 8 May 2020

to strengthen the consistency and persuasiveness of the Court's legal reasoning to improve its accountability and transparency. Other institutional bodies, both national and international, are involved in the analysis and evaluation of CDs. For example, the French National Observatory of Crime and Criminal Justice (ONDRP)[6], founded in 2003, aims to produce, collect, and disseminate data on crime, criminal justice, and safety issues. The European Intellectual Property Office Observatory (EUIPO)[7] performs CDs monitoring on the copyright; the United Nations Office on Drugs and Crime (UNODC) with the *SHERLOC* project[8] is facing the issues of analysing case law and creating a comprehensive case law database.

Nowadays, the massive volume of published CDs, makes retrieval, analysis, processing, and extraction of information from all the international and national court decisions physically impossible. The need for automated processes clearly emerges both from real-world applications and from the literature. In the field of legal big data, more and more researchers start to notice that combining traditional doctrinal legal methods and empirical quantitative methods is a promising approach, as reported in Medvedeva et al. [2019]. The evaluation of CDs is mandatory to discover legal knowledge but it arises many issues nevertheless. CDs document analysis is mostly run manually by domain experts. Indeed, knowledge extraction from CDs needs a strong technical (legal) expertise. However, providing automated knowledge extraction techniques only partially contributes to solve the problem. Other important issues have an impact on the legal knowledge extraction processes. First, access to legal data is not open. Data access policies change from jurisdiction to jurisdiction. CDs metadata are not always and everywhere available neither accessible without special permissions or an *ad hoc* agreements, due to several legal and administrative reasons. Access to CDs documents is often partial and not always entirely obtainable. As pointed out in Agnoloni T. [2014], CDs collections constitute an extremely technical and specialized textual genre. The unique lexicon, the structure of the phrases, the implicit knowledge and, overall, its semantics are difficult to catch even to a human if not supported by a rich background of strong legal knowledge. Even more for machines. Legal documents, as described in Ceci and Gangemi [2016], require special attention when representing their semantics, as they do not typically express factual knowledge, but they rather codify an order of an

[6]https://inhesj.fr/ondrp/the-ondrp, last visit 8 May 2020

[7]https://euipo.europa.eu/ohimportal/it/home, last visit 8 May 2020

[8]https://sherloc.unodc.org/cld/v3/sherloc/index.html?lng=en, last visit 8 May 2020

authority. Unlike a generic text, where the intended meaning of the combination of its signs is either common knowledge or is explained by the author, interpretation of legal documents is a different matter. The manual markup of CDs, however, doesn't seem to be sustainable in the long run. For efficient management of the knowledge acquisition phase, a combination of tools supporting an authored translation of text into semantics should limit the effects of this (still) unavoidable bottleneck. Semantic and rule-based approaches were integrated with text analysis and Natural Language Processing techniques, as presented in Wyner and Peters [2011] and in Winkels and Hoekestra [2012]. But the application of text analysis techniques on CDs requires singular and unique capabilities, due to the nature of the domain language. For instance, as reported in Ashley [2017], text normalization, such as text segmentation, tokenization, stemming and lemmatization, stop-word elimination, could considerably affect the meaning in the legal context, by modifying the information content. Unfortunately, that scenario influences also machine learning approaches, where it is necessary to make the natural language machine-readable, as discussed in Agnoloni T. [2014]. Moreover, machine learning techniques applied to CDs datasets most of the time require expensive manual annotation tasks. Indeed, handling the legal domain implies to solve other issues such as annotation agreement, time, expert human resources, and financial costs, as reported in Ashley [2017]; Breaux et al. [2006]; Kiyavitskaya et al. [2008]; Mochales and Moens [2011]; Grabmair et al. [2015]. Annotation tasks can evenly impose more than one year of involvement of high court judges and domain experts, as described in Mochales and Moens [2011].

To summarize, CDs are produced daily and the legal knowledge changes rapidly. Slow-moving analytical approaches may be not able to keep up with those changes, implying inevitable delay in knowledge extraction. For these reasons, legal knowledge extraction from CDs is becoming urgent. In such a context, techniques and tools for automated extraction of legal knowledge from CDs are strongly demanded, to support annotation, analysis, and understanding of these informative documents. The Ph.D. Thesis addresses these issues and presents CRIKE (CRIme Knowledge Extraction), a data science approach conceived to support legal knowledge extraction processes from a corpus of CDs documents. The main contributions of the thesis work are the following:

- *Design of LATO-KM (Legal Abstract Terms Ontology - Knowledge Model).* The LATO - KM is a three-layer knowledge model where terms featuring legal knowl-

edge, both law and case law, are properly formalized as concepts and relationships and they are implemented in the LATO ontology using SKOS. LATO - KM captures and formalizes the features and nature of terminology used in law and CDs. Our design challenge has been to find a suitable way of modelling the different nature of terms appearing in law and case law as well as their meaning and roles.

- *Definition of the CRIKE approach*. Legal knowledge extraction in CRIKE is based on a multi-label annotation technique that aims to associate CDs with appropriate concepts in the LATO ontology. Information retrieval techniques are used i) to automatically annotate CDs at diverse levels of granularity and ii) to enforce knowledge extraction and enrichment based on a given corpus of CDs, by discovering new terms for a given legal concept that has been recognized in the CD texts.

- *Survey about CDs accessibility and ethical issues involved in CDs storage and processing*. We have completed a survey among the European Union member states in order to define the state of the art about access to institutional CDs repositories considering Data Completeness, Data Access, Data Protection and Data Prediction aspects. To this end, we designed a questionnaire delivered to the Ministry of Justice, Supreme Court, Judiciary Office, and Judge Association of every country, and we analyzed the results. A further contribution of the thesis work regards the data science ethics framework proposed in Floridi and Taddeo [2016] on the CRIKE workflow, to highlight and discuss the ethical issues involved in processing CDs, in the proposed CRIKE approach.

The thesis is organized as follows. Chapter 2 and Chapter 3 are devoted to present and discuss the state of the art. In particular, Chapter 2 describes and classifies Legal Knowledge Modeling approaches, while Chapter 3 describes and classifies Legal Knowledge Extraction approaches. Chapter 4 presents the CRIKE data science approach. In particular, the LATO-KM knowledge model and related LATO ontology are illustrated. Then, techniques for CD documents annotation for knowledge extraction and enrichment based on LATO ontology are presented. Chapter 5 discusses the experimental results of applying the proposed CRIKE approach to a corpus of 180,000 CDs. Finally, Chapter 6 addresses applicability and ethical issues concerning CRIKE.

Chapter 2

State of the Art on Legal Knowledge Modeling Approaches

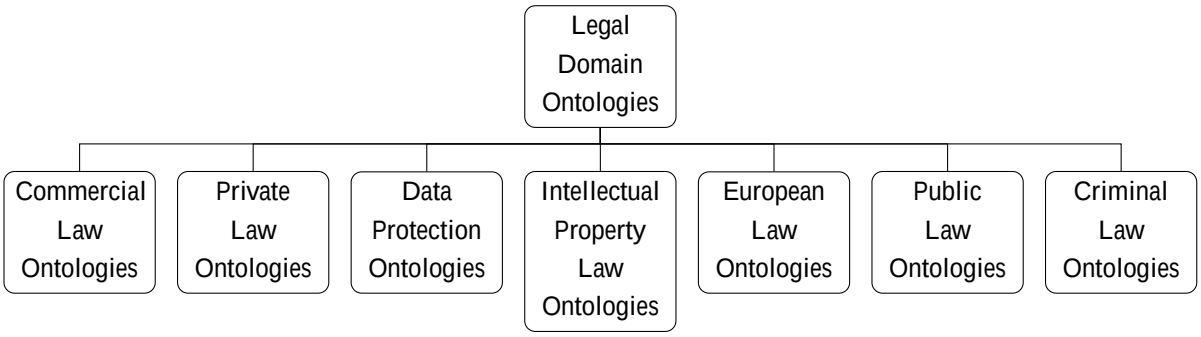
2.1 Introduction

Legal knowledge is usually modeled using ontologies. As reported in Lame [2005], ontologies can be seen as explicit specifications of conceptualisations and as representations of terminological knowledge. By analysing the state-of-the-art, three families of contributions can be recognized, namely *top* ontologies, *core* ontologies and *domain* ontologies. *Top* legal ontologies are devoted to formalizing the very general aspects of the legal knowledge, with the main aim of reasoning and inference on legal concepts. *Core* legal ontologies formalize concepts at a lower level of abstraction, positioned in the middle between the top and domain ontologies. The structure of *core* legal ontologies is influenced by the purpose, that can be: i) to model the knowledge in a structured and formalized way, for defining a reference knowledge framework to be further specialized in specific domain ontologies; ii) semantic indexing and information search, for identifying specific legal knowledge inside legislative provisions or for retrieving relevant legal knowledge from law and case law. Finally, *domain* legal ontologies are devoted to formalizing concepts and relations featuring a single domain such as civil law, commercial law, data protection and copyrights, international law, European law, public law and criminal law. Legal domain ontologies are mainly devoted to i)

understanding the domain, ii) organizing and structuring information for applications/system, iii) semantic indexing and information search. In this chapter, we focus on analysis and classification of the main approaches related to domain legal ontologies, since the thesis concerns modeling legal knowledge in the criminal law domain. The state-of-the-art has been analysed on four main surveys on legal ontologies, that are: Casellas [2011], Griffo et al. [2015], de Oliveira Rodrigues et al. [2019], Leone [2020]. **Casellas [2011]** appears as the first survey on legal ontologies in the literature. The author states that the plurality of theoretical approaches and the diversity of applications that Semantic Web technologies had to offer to the legal domain of 2011, confirmed how Law, Knowledge Management and Artificial Intelligence had already converged in an interdisciplinary area of research. One of the aims of the work is to survey the ontologies as a type of knowledge representation, with a special focus on legal ontologies as a form of representation and formalization of legal knowledge. **Griffo et al. [2015]** describe how over the previous decades, the field of legal ontologies had seen an increase in the number of papers. The literature on legal ontologies now covers a wide variety of topics and research approaches. One of these topics is represented by legal core ontologies, which had received significant attention since the 1990s. In order to provide an up-to-date overview of this research area, this article presents a systematic mapping study of published researches on legal core ontologies. In **de Oliveira Rodrigues et al. [2019]** the authors describe how, over the last 30 years, the field of study known as Artificial Intelligence (AI) & Law has provided breakthroughs in studies involving case-based reasoning, rule-based reasoning, information retrieval and, most recently, conceptual models for knowledge representation and reasoning, known as Legal Ontologies. However, the heterogeneity of the legal domain has led to the construction of content description models of miscellaneous ontological types. For this reason, in this work, a systematic mapping survey of 78 primary studies from the literature review is reported. The main aim of this work is to classify legal ontologies, proposed from the late '90s to 2017, along certain dimensions, such as purpose, level of generality, and also domains, nationalities, legal theories. **Leone [2020]** realises a second recent survey on 14 domain legal ontologies. Differently from the aforementioned de Oliveira Rodrigues et al. [2019], the authors here classified the most recent released legal ontologies by the details of their implementation, considering practical information concerning their actual availability for use and reuse. The survey focuses on resources that model a legal domain referring to some European, or globally ap-

plicable, legal framework. For this reason, ontologies based on national jurisdiction and whose sources files were not available for download have been excluded by the author. Fig 2.1 shows a classification of the main domains for which legal ontologies have been proposed and studied.

Figure 2.1: Overview of the application domains of legal domain ontologies contributions.



Literature contributions related to a given legal domain are discussed with respect to the following properties and criteria:

- **Name** of the ontology or project where the ontology has been developed;
- **Target Purpose** of the ontology, that is the main task the ontology is pursuing, by distinguishing between: i) *Organizing and Structuring Information*, where the knowledge is organized and structured in a logical formalization, for becoming part of an application/systems; ii) *Semantic Indexing and Information Search*, where the ontology main task is focused on retrieving information based on the ontology knowledge representation; iii) *Understanding the Domain*, where the ontology main task is to understand a particular legal domain and the formalization is used for sharing, communicating or unifying the domain knowledge.
- **Creation** of the ontology, that is the methodology used for modeling the knowledge: i) *Manual*, where the ontology has been manually realized by a domain expert; ii) *Semi-automatic*, where the ontology has been realized by relying on other ontologies as support, such as core or top ontologies, or by relying on text analysis techniques or other forms of metadata.

- **Formalization** of the ontology, that is the language used for ontology specification and implementation, by distinguishing between ORM, OWL, SKOS, RDF.

In the following, each legal domain is discussed in more detail.

2.1.1 Commercial Law Ontologies

Jarrar et al. [2003] This work presents the customer complaint ontology (CContology). The proposed model aims at capturing the knowledge elements of the so-called customer complaint domain. Even if the ontology is intended to become the basis for a future core ontology in the domain of customer complaint management, we notice that the model is only devoted to commercial law, and in particular to complains and litigations. The CContology is created with DOGMA, a methodological framework for ontology engineering created by the same authors and discussed in Jarrar and Meersman [2009]. Again on commercial law, an ontology for the U.S. Uniform Commercial Code (UCC) is presented in **Bagby and Mullen [2007]**. In this work, the main focus is on building a composite lawyer's ontology refined with the law from actual practice. Authors developed their ontology using OWL, with the future aim of being part of a framework for e-commerce and dispute resolution. Moreover, in **Stolarski and Tomaszewski [2008]**, a statute-specific legal ontology of the Polish Commercial Companies Code (PCCC) is presented. In this work, the Methontology methodology described in Corcho et al. [2005] has been used for defining concepts, properties and axioms. The authors point out that a problem of all legal knowledge is the fuzziness of reality compared with the restricted language of the law. Finally, **Abrahams et al. [2011]** present an OWL ontology about Australian commercial law, where legal arguments and outcomes of past cases, heard by the Victorian Civil and Administrative Appeals Tribunal (VCAT) are modeled. In this work, a domain expert formalized judicial reasoning for corporate cases in order to capture the discrete areas of decision making and factors used in legal arguments. The aim of the work is to improve the judicial case management before proceeding to litigation.

Table 2.1: Comparative classification of the commercial law domain ontologies.

	Name	Target Purpose	Creation	Formalization
Jarrar et al. [2003]	Costumer Complain Ontology (CContology)	Understanding the Domain	Manual	ORM
Bagby and Mullen [2007]	Uniform Commercial Code (UCC) Article II Ontology	Semantic Indexing and Information Search	Manual	OWL
Stolarski and Tomaszewski [2008]	Legal Ontology of Polish Commercial Companies Code (PCCC)	Understanding the Domain	Manual	OWL
Abrahams et al. [2011]	Best Alternative To a Negotiated Agreement (BATNA) Ontology	Semantic Indexing and Information Search	Manual	OWL

2.1.2 Private Law Ontologies

An OWL ontology, presented by **Salam [2007]** details the conditions and rules for contracts. This work takes on the issue of the incompatibility between contractual clauses (typically buried in legal documents) and the performance measures, used to evaluate and reward (or penalize) supply participants in the extended enterprise. In **Mittal et al. [2016]** a second OWL ontology aimed at describing, managing and reasoning about cloud Service Legal Agreements (SLAs) is described. In this work, the authors propose a process for extracting, managing and monitoring cloud SLAs documents using natural language processing techniques and Semantic Web technologies. Moreover, in **Lovrencic and Tomac [2006]** a family law ontology aimed at an easier understanding of the Croatian Family Legislation is described. The model is formalized in PAL, Protégé Axiom Language, with the direct involvement of a legal expert and family law judge, who helps to cover the issues related to the legal development, realized by practice. Family law issues are also taken on in **Tanapon and Nuanwan [2010]** where an ontology for Thai succession law is described. The ontology is used

inside a system devoted to improving CDs sentences retrieval. Consumer Protection law has been the object of several works. Agnoloni et al. [2007] and Francesconi et al. [2007] discuss an ontology developed within the DALOS (Drafting Legislation with Ontology-based Support) project[1]. In these works, the knowledge resource is expected to support legislative drafting, by providing legal drafters and decision-makers control over the legal language at the European level. These works point out how the project focuses on the consumer protection domain, by providing a multi-lingual lexical layer combined with an ontological layer, where conceptual representations of the domain at a language-independent level are formalized in RDF/OWL. In Francesconi et al. [2007] the Knowledge Organization System (KOS) is introduced as a specification of the DALOS resource. But, it is in **Francesconi et al. [2010]** that DALOS KOS is presented as a middle-out legal ontology and another type of lexical relationship, the so–called *fuzzynym*, is described. In this work, the authors organize DALOS KOS in *two layers*: the Ontological Layer and the Lexical Layer. Classes and properties have been implemented on the basis of the terminological knowledge extracted from the chosen EU Directives on consumer protection law. In DALOS KOS, concepts at the Ontological layer are linked by taxonomical as well as object property relationships (e.g. `has object role`, `has agent role`, `has value`, etc.). On the other hand, the Lexical layer aims at describing the language–dependent lexical expression of the concepts contained in the Ontological layer. At this level, lexical units can be linked through linguistic relationships such as synonymy, hypernymy, hyponymy, meronymy, etc. Another type of lexical relationship, so–called *fuzzynym*, appears to hold between the terms, for example *consumer* and *supplier*. Such a relationship refers to a wider associative relation linking words which may share a number of salient features (in the case at hand, of being involved in a commercial transaction) without being necessarily semantically similar. More in particular, the authors point out that this type of lexical relationship, the so–called *fuzzynym*, is the last resort for all the remaining terms relationships that are not relationships such as synonymy, hyponymy/hypernymy, meronymy, antonymy. Moreover on consumer law, special Air Transport Passenger issues are faced in **Rodríguez-Doncel et al. [2014]**, where an OWL ontology for passenger rights is modeled. This work aims at formalizing air-travel legal information in case of incidents for dispute resolution hypothesis, that can be con-

[1]DALOS (DrAfting Legislation with Ontology-based Support) e-Participation project (01-01-2007/30-04-2008) website: http://www.dalosproject.eu.

nected with institutional resources and airlines. For the validation of the ontology, a legal expert was involved. The work by **Ceci and Gangemi [2016]** instead presents the Judicial Ontology Library (JudO). Here the authors describe a semantic tool to enrich and reason on the XML mark-up of precedents and metadata of CDs. The modeling of the metadata taken from CDs and represented in JudO was carried out manually by a legal expert. In **Laarschot et al. [2005]** an OWL ontology abut Dutch tort law is formalized as follows: i) tort law, where the authors mainly copied in the ontology the structure of tort law as it is described in Dutch law, ii) entities subject to the law (a legal person, a natural person) and iii) objects in tort law (motor vehicles, animals, product). The ontology has multiple aims, such as, analyse the input of a layman in terms of a layman ontology, retrieve relevant CDs, and present the results in a comprehensible way to the layman.

Table 2.2: Comparative classification of private law domain ontologies.

	Name	**Target Purpose**	**Creation**	**Formalization**
Lovrencic and Tomac [2006]	Family Legislation Act Ontology	Organize and Structure Information	Manual	PAL
Salam [2007]	Extended Enterprise Supplier Performance Contract Ontology	Organize and Structure Information	Not specified	OWL
Tanapon and Nuanwan [2010]	SCROI - SCRO II	Semantic Indexing and Information Search	Not specified	Not specified
Mittal et al. [2016]	Ontology for Cloud Service Level Agreement	Semantic Indexing and Information Search	Manual	OWL
Francesconi et al. [2010]	Drafting Legislation with Ontology-based Support (DALOS KOS)	Semantic Indexing and Information Search	Manual and Semi-automatic	RDF - OWL

Continued on next page

Table 2.2 – *Continued from previous page*

	Name	Target Purpose	Creation	Formalization
Rodríguez-Doncel et al. [2014]	Air Transport Passenger Incidents and Rights Ontology	Organize and Structure Information	Manual	OWL
Ceci and Gangemi [2016]	Judicial Ontology (JudO)	Semantic Indexing and Information Search	Manual and Semi-automatic	OWL
Laarschot et al. [2005]	Ontology of Dutch Tort Law (BEST)	Semantic Indexing and Information Search	Manual and Semi-automatic	OWL

2.1.3 Data Protection Ontologies

In **Mitre et al. [2006]**, an OWL ontology concerning the main Spanish data protection code, called LegLOPD (Legal Ontology Domain), is presented. LegLOPD is composed by five top concepts directly extracted from the LRI-Core ontology, a core ontology that covers the main concepts that are common to all legal domains, described in Breuker et al. [2005]. An ontological semantics for Data Privacy Compliance, again about the Spanish law, is presented in **Casellas et al. [2010]**, as part of the NEURONA project. Here, a modular ontology is based on the knowledge acquired and organized by legal experts, where the main aim is to model data protection concepts for a reasoning system. The construction of the ontology was thus focused on the acquisition of conceptual domain knowledge extracted from legal documents. In **Cappelli et al. [2007]**, the OntoPrivacy, and ontology of the (previous) Italian data protection code is presented. A glossary of keywords extracted from the regulations was manually created by a legal expert. Moreover, **Bartolini et al. [2017]** present a bottom-up ontology describing the constituents of the data protection domain and its relationships with respect to the General Data Protection Regulation (GDPR). The ontology was modeled by a legal expert, who confirmed some difficulties in legal ontology design. Another ontology for GDPR is reported in **Pandit et al. [2018]**. In this work, concepts and obligations of the European data protection regulation are formalized using SKOS, with the set of attributes and terms provided by European Legislation Identifier (ELI)

metadata. In **Oltramari et al. [2018]** the PrivOnto ontology, a semantic representation of US company privacy policies and US privacy law and standards, is presented. The main aim is to model and reason over privacy practice statements. GDPR concepts are also modeled in **Palmirani et al. [2018]**, where the PrOnto ontology is presented. In this work, the model of the GDPR main concepts (data types, documents, processing purposes, legal bases, processing operations) aims to support legal reasoning and compliance checking.

Table 2.3: Comparative classification of data protection domain ontologies.

	Name	**Target Purpose**	**Creation**	**Formalization**
Mitre et al. [2006]	The Legal Ontology Domain (LegLOPD)	Organize and Structure Information	Manual and Semi-automatic	OWL
Cappelli et al. [2007]	OntoPrivacy	Semantic Indexing and Information Search	Manual	RDF
Casellas et al. [2010]	NEURONA Modular Ontology	Organize and Structure Information	Manual and Semi-automatic	OWL
Bartolini et al. [2017]	Data Protection Ontology	Understanding the Domain	Manual	OWL
Pandit et al. [2018]	GDPRtEXT	Understanding the Domain	Manual	SKOS
Oltramari et al. [2018]	PrivOnto	Semantic Indexing and Information Search	Manual	OWL
Palmirani et al. [2018]	PrOnto	Semantic Indexing and Information Search	Manual and Semi-automatic	RDF-OWL

2.1.4 Intellectual Property Ontologies

In **Delgado et al. [2003]**, the Intellectual Property Rights Ontology (IPROnto) is presented. Legal concepts are formalized with respect to the World Intellectual Property Organisation (WIPO), a common legal framework for IPR (Intellectual Property Rights), used as a support for the ontology creation, which was formalized with the Semantic Web ontology language DAML+OIL. IPROnto focuses on e-commerce applications, in which disputes on multimedia content rights arise quite regularly. **Adams [2008]** discusses an OWL ontology for representing and communicating knowledge of intellectual property from a trans-disciplinary perspective. This work confirms that, with regards to ontology creation and design, even if the advantage of using automated methods, such as natural language processing (NLP) is clear, a manual analysis of source documents done by an expert in the field can perform better in identifying and organising concepts. An intellectual property rights formalization, the ALIS IP ontology, is discussed also in **Cevenini et al. [2008]**, where the project ALIS (Automated Legal Intelligent System) on Intellectual Property (IP) law is presented. In this work, ALIS-IP ontology is compared with the IPROnto ontology, discussed above. Difficulties with abstract concepts emerged during the ontology design and the selection of the legal concepts. Indeed, the creators, facing the Intellectual Property law issues, pointed out that, even if the law reports the concept of works of the mind, describing this concept as the list of all the works of mind protected by copyright, it does not also define what is a works of the mind, leaving the interpreter without an official definition. Moreover, in **Baumann and Loës [2010]** a formalization of the German copyright law for the internet of services is presented. In this work, the authors highlight that they use the legal code as their primary source in order to achieve compliance, with respect to the German copyrights, formalized in ODRL.

Table 2.4: Comparative classification of intellectual property domain ontologies.

	Name	Target Purpose	Creation	Formalization
Delgado et al. [2003]	Intellectual Property Rights Ontology (IPROnto) Copyright Ontology (CO)	Organize and Structure Information	Manual and Semi-automatic	DAML+OIL
Adams [2008]	Transdisciplinary Ontology of Innovation Governance	Understanding the Domain	Manual and Semi-automatic	OWL
Cevenini et al. [2008]	Intellectual Property Ontology	Organize and Structure Information	Manual and Semi-automatic	OWL
Baumann and Loës [2010]	Copyright Ontology	Organize and Structure Information	Manual and Semi-automatic	ODRL

2.1.5 International, European and Public Law Ontologies

Alexander Boer [2001] This work describes the CLIME (Cooperative Legal Information Management and Explanation) ontology. The CLIME ontology was developed for the purpose of a web-based legal advice system, called MILE (Maritime Information and Legal Explanation) for both extended conceptual information retrieval and normative assessment. Also in this work, the difficulties related to abstract concepts are confirmed. In **Kerremans et al. [2003]** an ontology aimed at representing multilingual information about the European VAT (value-added tax) regulatory is discussed. This work highlights the differences between VAT regulations among countries, that have to account both i) culture-specific (as well as non-culture specific) VAT units of understanding and ii) internal relationships, in order to cover the VAT legislation of the different European member states. LOTED2, an OWL ontology for European public procurement notices, is discussed in **Distinto et al. [2016]**. This ontology is considered as a legal ontology, first because representations of information regarding public tender notices in Europe are modeled, and secondly, because it supports the identification of legal concepts and allows for legal reasoning. **Muñoz-Soro et al. [2016]** describe

the PPROC ontology, an OWL ontology for Transparency in Public Procurement in Spain. The authors highlight the completeness of the model, indicating that PPROC is extensive, covering not only the usual data about the tenders, such as deadlines and awardees, but also the details of the whole process, from the initial contract publication to its termination.

Table 2.5: Comparative classification of international, European and public law domain ontologies.

	Name	Target Purpose	Creation	Formalization
Kerremans et al. [2003]	European VAT Regulatory Ontology	Semantic Indexing and Information Search	Manual and Semi-automatic	DOGMA
Alexander Boer [2001]	Cooperative Legal Information Management and Explanation (CLIME)	Semantic Indexing and Information Search	Manual and Semi-automatic	RDF
Distinto et al. [2016]	Linked Open Tenders Electronic Daily (LOTED2)	Semantic Indexing and Information Search	Semi-automatic	OWL
Muñoz-Soro et al. [2016]	Public Procurement Ontology (PPROC)	Organize and Structure Information	Manual	OWL

2.1.6 Criminal Law Ontologies

Breuker et al. [2005] This work presents the use of various ontologies for the information management of documents produced inside the criminal trial hearings. These ontologies are used in the e-COURT project and cover the Dutch criminal law by following the structure of the LRI-Core ontology described above. The main aim of this work is to support information retrieval by tagging and annotating the hearing documents. **Shankhdhar et al. [2014]** present a semantic web-based recommendation system. More in detail, an OWL ontology helps legal expert users in extracting CDs

from a case law repository, with respect to similar cases. The authors report that the ontology is designed exclusively for that purpose. Indeed, the knowledge represented derives from the Indian Penal Code and an application on murder cases is proposed. In **Rodrigues et al. [2015]** an ontological approach for simulating legal action in the Brazilian Penal Code is discussed, with a focus on drunk-driving law. Two models are presented, OntoCrime and OntoLegalTask. OntoCrime is an ontological representation of the Brazilian criminal code in machine-readable forms. OntoLegalTask is a task ontology created to enable a set of actions such as checking violation of the law, detecting inconsistencies, and legal punishment. Both models have been realized according to the METHONDOLOGY methodology, discussed in **Corcho et al. [2005]**. This approach has been tested on the Brazilian drunk-driving law, which is based on three legislative levels, the traffic code, and two dedicated laws, by formalizing the legal content. Some practical example cases with reasoning results are then reported. Difficulties emerge in terms of norms conflicts, multiple violations and chronological interpretative issues. **Ghosh et al. [2017]** present a criminal ontology defining Lebanese criminal law, formalized in SWRL. Authors modeled this ontology for constructing a legal rule-based decision support system, for the criminal domain, named CORBS. The system supports legal decision-making thanks to a rule-based reasoning approach. The rule-based decision system contains a set of logic rules composed of atoms, that are defined based on the ontology elements and formalized using SWRL. Authors report that SWRL is used because it is better suited to express deductive knowledge by rules composed of atoms. **Fawei et al. [2019]** present an ontology engineering methodology and a semi-automated approach of legal ontology generation, both derived from a collection of legal documents. The generation approach also includes legal rules to provide reasoning support to an automatic legal question answering system. The system uses the Stanford parser to pre-process the input text for producing semantic triples. By applying the Stanford CoreNLP to perform tokenization, part-of-speech tagging (POS) and named entity recognition, the proposed approach aims at identifying and extracting both legal and common-sense keywords, in order to answer the Bar exam questions. **Asaro et al. [2003]** describe the first and only Italian criminal law ontology, to the best of our knowledge. A UML ontology, based on the Italian crime law, used as a support tool for the judges activity in the criminal field is here presented. The ontology formalizes the difference between the criminal behaviours with respect to the offences and to the interests protected by the law, presenting the crime structure

as follows: an offender, a behaviour, a penalty and optionally, an event and a coercions. We conclude this literature review on criminal law ontologies with two recent works. First, in **Mezghanni and Gargouri [2017]**, the CrimAr ontology for Arabic criminal law is presented. CrimAr is based on the top-levels of LRI-Core ontology (Breuker and Hoekstra [2004]) and represents all the relevant knowledge in the Arabic legal domain, especially in the criminal matter. Secondly, in **Soh et al. [2018]** an ontology for the Korean criminal law is described. In particular, this work discusses an ontology-based legal knowledge representation first, and then a logic-based legal rule design methodology, with an application tested on the Korean anti-graft act.

Table 2.6: Comparative classification of criminal law domain ontologies.

	Name	**Target Purpose**	**Creation**	**Formalization**
Breuker et al. [2005]	OCN.NL	Semantic Indexing and Information Search	Manual	OWL
Shankhdhar et al. [2014]	Legal Semantic Web A Recommendation System	Semantic Indexing and Information Search	Manual and Semi-automatic	OWL
Rodrigues et al. [2015]	OntoCrime	Organizing and Structure Information	Manual and Semi-automatic	OWL
Ghosh et al. [2017]	Lebanese Criminal Law Ontology	Understanding the Domain	Manual and Semi-automatic	SWRL
Fawei et al. [2019]	Criminal Law Ontologies and Rules (CLOR)	Understanding the Domain	Manual and Semi-automatic	OWL + SWRL
Asaro et al. [2003]	Italian Crime Ontology	Semantic Indexing and Information Search	Manual and Semi-automatic	UML

Continued on next page

Table 2.6 – *Continued from previous page*

	Name	Target Purpose	Creation	Formalization
Mezghanni and Gargouri [2017]	CrimAr	Semantic Indexing and Information Search	Manual and Semi-automatic	Not specified
Soh et al. [2018]	Anti-Graft Act Domain Ontology	Organizing and Structure Information	Manual	SWRL
Castano et al. [2019]	Legal Abstract Terms Ontology (**LATO**)	Semantic Indexing and Information Search	Manual	SKOS

Considerations. With regard to *formalization*, OWL is largely preferred and most popular. Moreover, OWL is combined with RDF and SWRL. Single usage appears for other formalisms, such as ODRL, SKOS, DOGMA, PAL, ORM. Concerning the knowledge base, the law is undoubtedly the main base of knowledge, whereas only single works modeled the legal knowledge of polices, (Oltramari et al. [2018]), or of contracts (Mittal et al. [2016]), or of both law and case law (Ceci and Gangemi [2016]). Regarding the *target purpose*, we have detected fragmented results. Most of the compared works show domain ontologies as part of a complex system, or as part of a structured framework. About the domain, we have noticed that the largest part of the effort is devoted to the domains of Data Protection and Criminal law. Private law and Commercial law come immediately after, and remaining domains follow.

2.2 Thesis Contribution

From the literature comparison, two important issues emerged. First, the problem of modeling abstract concepts and secondly, the problem of capturing concrete interpretations for abstract concepts. Indeed, many works report difficulties in modeling legal knowledge caused by abstract concepts and their interpretation. In particular, in **Bartolini et al. [2017]** it is reported that in the proposed ontology some concepts still required a judicial decision, whereas some other concepts appeared as generic or eval-

uative, just because they are expressed as such in the law. In **Cevenini et al. [2008]**, difficulties with abstract concepts emerged during the ontology design and the selection of legal concepts. In fact, the creators of this ontology, facing Intellectual Property law issues pointed out that, even if the law reports the concept of works of mind, it fails to define what work of the mind is. Again **Ghosh et al. [2017]** confirmed the problem of how to model vague or open-textured concepts, precisely what we call abstract concepts. For instance, uncertain and fuzzy legal concepts such as *reasonable* and *intentional* cannot be modeled in any way analogous to human thinking. Furthermore, in **Alexander Boer [2001]** the authors confirmed how the CLIME ontology does not capture all the legal knowledge or uses of the contents of legal documents, due to the fact that the terminological knowledge base is never definitive. In fact, legal concepts are by necessity open-texture at a certain degree, and these concepts become more rigorously defined only as time progresses and new interpretations are established. The need for capturing interpretations to resolve ambiguities is confirmed in **Ajani et al. [2017]**, where the authors pointed out that the legal domain is constantly subject to evolving conceptualizations and neologisms. Legal sublanguages, despite their conservative style, show remarkable flexibility and dynamism in redefining new words and coining new ones when the need arises: capturing this sublanguage is a challenge in the legal domain ontology design. Indeed, **Fawei et al. [2019]** point out that legal knowledge is usually expressed with domain-specific terminology and conveyed in textual form, so that its expression and presentation do not provide a standard structure for a machine to use and reason with. Representing human-created semantic information from the text for machine processing is not a linear process. The manual method of extracting and classifying legal text according to classes and object properties involves reading textual documents from different sources that deal with a given legal domain. On the same issue, **Kiyavitskaya et al. [2008]** observe how a number of challenges complicate the automated annotation of regulatory texts. For example, U.S. federal regulations are highly structured and written in legalese. Despite this structure, the conventions of legalese are not always used consistently, with intentional and unintentional ambiguities, and individual requirements are described across multiple sentences and paragraphs using cross-references. The problem remains for every legal knowledge system and is connected with the fuzziness of reality compared with the restricted language of the law, as expressed in **Stolarski and Tomaszewski [2008]**. But, as confirmed by **Lovrencic and Tomac [2006]** the law is built upon practice and

individual interpretations, which can be very different from case to case. Finally, also in **Francesconi et al. [2010]** the authors recognize the presence of generic situations in the law, and indeed, they present a the two-layer ontology, the DALOS KOS, that is expected to support legislative drafting, by providing legal drafters and decision-makers control over the legal language at the European level. But, in the DALOS KOS domain ontology classes and properties have been implemented on the basis of the terminological knowledge based on chosen EU Directives on consumer law, a subcategory of the main topic civil law, without presenting a dedicated kind/role for representing the legislative generic situations at an ontological layer. Neither at the lexical layer, where the so–called *fuzzynym* is intended as a vague relation ascribed to a generic syntagmatic relatedness of the terms reported in the law. In fact, the so–called *fuzzynym* has not the aim of capturing and formalizing the features and nature of the terminology used in both law and CDs, nor the aim of representing the unique conceptual relations that abstract concepts have in CDs.

Contribution of the thesis. For modeling abstract concepts and extracting the related interpretation in CDs, we propose the three-layer LATO-KM knowledge model, defined from the criminal law perspective, whose goal is to capture and formalize the features and nature of the terminology used in both law and CDs. Our design challenge has been to find a suitable way of modeling the different nature and relation of terms appearing in law and case law as well as their meaning and roles, by capturing also abstract concepts. This modeling approach permits LATO-KM to present legal concepts, and their relations, according to two different views, a law-oriented view and a case law-oriented view.

Chapter 3

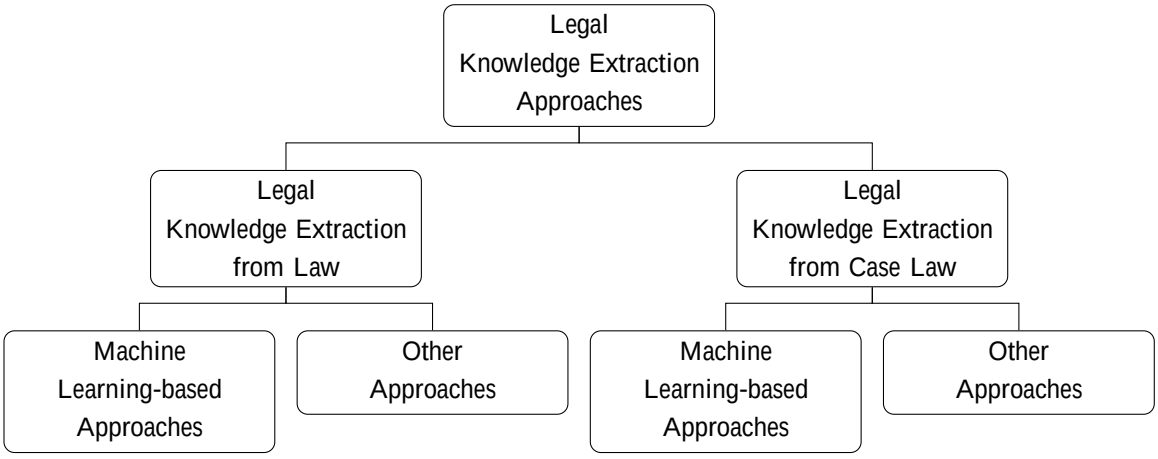
State of the Art on Legal Knowledge Extraction Approaches

3.1 Introduction

As literally reported in Wyner and Peters [2011], one of the long term goals in data science and law is to identify, extract, and formalise conditional or normative rules from legal source materials. This activity is particularly challenging because knowledge extraction from legal text has to face the peculiarities of the legal domain that radically influence every mining approach. Legal documents are a technical and specialized textual genre, where both semantic-based approaches and machine learning techniques find peculiar difficulties. First, legal texts have their own unique characteristics. The lexicon, the structure of the periods, the implicit knowledge, and the overall semantics are difficult to understand if not by exploiting a rich background of technical (legal) knowledge (Agnoloni T. [2014]). In particular, in the thesis we deal with CDs which are characterised by a peculiar lexicon (e.g., the legal jargon), a logical phrase structure (e.g., the legal argumentation) and official document layouts (e.g., the legal document structure). For all these reasons, legal text pre-processing activities, such as tokenization, normalization (lower-case, stemming, stop-words elimination), annotation, and other text segmentation strategies are not always useful, in that simple raw text transformation may affect the legal meaning. Indeed, as addressed in Savelka and

Ashley [2016], even a single word may be crucial for the understanding of the provision as applied in a particular context. Machine learning approaches have to face the same issues when text is represented as feature vectors. Additionally, as it will clearly emerges from the literature comparison, manual annotation of legal texts for training set creation is one of the most relevant issues in many research works. Considering the goals of the thesis, we focus on related work on knowledge extraction from legal text, by recognizing two main families, *Knowledge extraction approaches applied to law*, and *Knowledge extraction approaches applied to case law*. In both cases we distinguish *Machine learning-based approaches* and *other approaches* (see Fig 3.1).

Figure 3.1: A classification of the main legal knowledge extraction approaches.



Various approaches are analysed and classified according to the following properties and criteria:

- **Name** of the system/approach;
- **Technique** featuring the approach;
- **Dataset** on which knowledge extraction techniques are applied;
- **Main Task** for which the knowledge extraction approach has been conceived, by distinguishing between: i) *Text Classification* where techniques are combined with text classification tasks; ii) *Knowledge Extraction* where techniques are mainly devoted to obtaining knowledge from the legal texts; iii) *Information Retrieval* where techniques are combined with information retrieval tasks.

- **Training-Set Creation** that is, how the training set has been realized for machine learning tasks;

The literature comparison is based on **Ashley [2017]**, a recent survey, where an entire chapter is devoted to legal text analytics topics such as machine learning with legal text, extracting information from statutory and regulatory text, and extracting argument-related information from legal case text.

3.2 Knowledge Extraction from Law

In this section, we introduce an overview of the approaches dedicated to knowledge extraction from law. This classification includes the works applied to law (datasets), intended as the set of rules, usually made by a government, formally recognized as binding or enforced by a controlling authority.

3.2.1 Machine Learning-based Approaches

In this subsection, the knowledge extraction approaches based on machine learning techniques and applied on the law are described. An overview of the main contributions is presented in Table 3.1. In **Savelka et al. [2015]** the authors relied on an interactive machine learning application that has gained successful results in classification tasks in many domains, such as web image search, data analysis, and electronic discovery (e-discovery). The reported approach has been tested to check if it could be a useful extension to traditional legal Information Retrieval (IR) systems for statutory analysis, by assessing the ability of a classification model to gradually learn from the feedback provided by a human expert and the ability to improve the suggestions as additional feedback becomes available. The results show that interactively trained machine learning classifiers provide reasonable suggestions about the relevance of statutory provisions, with increasing accuracy as more of the provisions are being processed and it becomes possible to re-use the classifiers in future analyses. **Francesconi and Passerini [2007]** describe the *xmLegesClassifier*, a module able to automatically detect the type of provisions contained in legislative texts. As part of the NIR project, where a standardized description of legislative documents is developed, the module relays on a standard for legislative document representation, realized using XML techniques,

named XML-NIR. Later, in **Francesconi [2009]** a tool called *xmLegesExtractor* is presented. The tool is realized as a suite of NLP tools for the automatic analysis of Italian texts, designed to cope with legal jargon. The *xmLegesExtractor*, combines MSVM with NLP techniques, in order to support a bottom-up knowledge acquisition from the legal text, and presents a twofold aim: i) supporting taxonomy implementation or suggesting concepts for handcraft ontologies and ii) extracting rules directly from the legal text. SVM for legislative classification has been used in two more works. First, in **Opsomer et al. [2009]**, the authors take on the issue of organizing legislative texts into a hierarchy of legal topics. This work confirms that manually placing every part of new legislative texts in the correct place of the hierarchy turns out to be an expensive and slow process. Consequently, this work deals with the ability of machine learning methods to develop a model that automatically classifies legislative texts in a legal topic hierarchy, applied to the Belgian law. Secondly, in **de Maat et al. [2010]** SVM was tested for classifying sentences in Dutch legislation against a knowledge engineering approach for text classification. The reported results show that the SVM classifier scores worse when it comes to generalising laws outside its training set.

Table 3.1: Legal knowledge extraction from law - machine learning-based approaches overview.

Approach	Name	Technique	Dataset	Training Set Creation	Main Task
Francesconi and Passerini [2007]	xmLeges Classifier	SVM and Naïve Bayes	Italian Law	Pre-annotated	Text Classification
Francesconi [2009]	xmLeges Extractor	SVM and NLP	Italian Law	Manual	Information Extraction
Opsomer et al. [2009]	EMIS Navigator	SVM	Belgian Law	Manual	Text Classification
de Maat et al. [2010]	Classification of Legal Texts	SVM	Dutch Law	Pre-annotated	Text Classification
Savelka et al. [2015]	Interactive Tool for Statutory Analysis	SVM with Relevance Feedback (RF)	US Law	Manual	Text Classification

3.2.2 Other Approaches

In this subsection, the knowledge extraction approaches, different from those based on machine learning, and applied on the law are described. An overview of the main contributions is presented in Table 3.2. **Breaux et al. [2006]** present a methodology for extracting and prioritizing rights and obligations from regulations. The methodology shows how semantic models can be used to clarify ambiguities and to balance rights with obligations through focused elicitation. The proposed process is applied to the Privacy Rule from the U.S. Health Insurance Portability and Accountability Act (HIPAA). The authors adopted a process called Semantic Parameterization, in which rights and obligations from regulation texts are restated into Restricted Natural Language Statements (RNLS), to describe discrete activities. The relaxed form of Semantic Parameterization uses only two RNLS patterns: i) activities that distinguish subjects and objects, and ii) activities following condition keywords such as, if, unless, and except. The same topic has been explored in **Kiyavitskaya et al. [2008]**, where the authors employed a semiautomatic annotation tool called Cerno with an extension for regulatory text called Gaius T, implemented using the Cerno framework, aimed at extracting rights, obligations and exceptions again from the U.S. Health Insurance Portability and Accountability Act (HIPAA). A similar aim is pursued in **Takano et al. [2010]**, where text processing and NLP techniques have been applied to extract logical structures from statutory paragraphs. In this work, a system for logical formulation of law sentences called WILDCATS is presented. The system is applied on the Japanese law and it translates a law sentence into a logical formula, assigning content words in the sentence to a logical predicate. Regulation phrases analysis is faced also in **Wyner and Governatori [2013]**. This work focuses on the translation of statements in regulations into formal semantic representations, which can then be used for several tasks, such as, i) drawing inferences given ground facts, and ii) providing users with meaningful explanations. In this paper, the authors discuss a pilot study to translate the regulatory statements of the Australia Telecommunications Consumer Protections Code into semantic representations. Moreover, on the issue of extracting rules for business compliance, **Zhang and El-Gohary [2015]** propose an automatic approach for regulatory requirements extraction, applied on a construction law dataset. The authors use rule-based approaches, semantic representations and NLP tools to extract tags, later used for automatic extraction of information from law. However, even if the

results appeared promising, the proposed approach was only partially tested, due to the large amount of manual effort required in developing a proper gold standard. The same problem of human annotation of regulation is faced in **Yoshida et al. [2013]**. In this work, pre-defined templates with the expression of functional requirements to identify legal texts are described. This approach responds to the need for the development of systems, that are compliant with laws in public administration, especially with new laws that are introduced or with existing laws that are amended. A different linguistically-oriented, rule-based approach for identifying and extracting rules in regulations is proposed in **Wyner and Peters [2011]**. In this work, the authors identify and extract high-level components of rules from regulations in English, such as the US Code of Federal Regulations, US Food and Drug Administration, Department of Health and Human Services regulation, always applying and extending widely available NLP tools. In particular GATE, an open-source framework for language engineering applications was tested. This framework enables linguists and text-engineers to develop and apply a variety of natural language processing tools to a corpus. Automatic extraction of concept and definition from law, using NLP and standard semantic web technology is presented in **Winkels and Hoekestra [2012]**. In this work, the authors describe a system applied on Dutch Tax law, focused on automatic recognition of concepts in legislative texts. Language processing issues are reported, for instance, the fact that Dutch is a difficult language and far fewer parsing tools are available than for e.g. English. The extraction of definitions appears not successful even though the same method and patterns were tested on different legal domains. This may confirm that legal domains present common aspects, but also multiple peculiarities. **Jinhyung et al. [2012]** describe how the Ministry of Justice of Korea and the Korean Institute of Science and Technology Information have developed an intelligent legislation support system, namely, the iLaw system. Semantic Web and text mining techniques were tested, aimed at targeting various information resources including domestic and foreign legal information. Similarly, in **Steinberger et al. [2012]** the main focus is mining law for automatic classification of EU documents. In this work, JEX, a JRC-developed multi-label classification software[1] is described. This system learns from manually labelled data how to automatically assign EuroVoc[2] descriptors to new documents in a

[1] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer.

[2] A highly multilingual thesaurus consisting of over 6,700 hierarchically organised subject domains used by European Institutions and many authorities in Member States of the European Union (EU) for

profile-based category-ranking task. Results are diverse with respect to the language itself and the language families.

Table 3.2: Legal knowledge extraction from law - other approaches overview.

Approach	Name	Technique	Dataset	Main Task
Takano et al. [2010]	WILDCATS	Logic Rule-based, Text Processing and NLP	Japan Law	Text Classification
Wyner and Governatori [2013]	C/Boxer and Defeasible Logic (DL)	Logic Rule-based, Text Processing and NLP	Australian Law	Knowledge Extraction
Breaux et al. [2006]	Semantic Parameterization	Semantic Model and Text Processing	US Law	Knowledge Extraction
Kiyavitskaya et al. [2008]	Gaius T.	Semantic Model and Text Processing	US and Italian Law	Knowledge Extraction
Wyner and Peters [2011]	On Rule Extraction from Regulations	Rule-based, Text Processing and NLP	US Law	Knowledge Extraction
Steinberger et al. [2012]	EuroVoc Indexer JEX	Ontology-based and Text Processing	EU Law	Text Classification
Zhang and El-Gohary [2015]	Automated Compliance Checking	Ontology-based, Text Processing and NLP	US Law	Knowledge Extraction
Winkels and Hoekestra [2012]	MetaLex Annotator	NLP and Semantic Model	Dutch Law	Knowledge Extraction
Yoshida et al. [2013]	Pre-defined Templates	Rule-based	Japan Law	Knowledge Extraction
Jinhyung et al. [2012]	Ontology Model for Legal Information (iLaw)	Semantic Model	Korean Law	Information Retrieval

the classification and retrieval of official documents.

3.3 Knowledge Extraction from Case Law

Similar issues and approaches emerged from the literature comparison of the works that present knowledge extraction techniques applied on case law datasets. For the development of the thesis, we consider now the literature comparison approaches different from those based on machine learning.

3.3.1 Other Approaches

In this subsection, the knowledge extraction approaches not based on machine learning techniques and applied on case law datasets are described. An overview of the main contributions is presented in Table 3.3. In **Dick [1991]** an approach for concepts legal information retrieval on Vermont case law is presented. First, legal knowledge is represented in an argument schema designed with concepts such as intention, contract and claim. This representation is used together with information retrieval techniques to retrieve information from CDs. An approach for automatic categorization of CDs of civil court of the first instance is presented in **Agnoloni T. [2014]**. In this work, NLP approaches and domain features extraction are applied on a CDs dataset composed of approximately 7000 documents, namely all the available deposited CDs pronounced by an Italian court of first instance on civil matters, over a time span of 5 years, from 2008 to 2013. Various text processing techniques were applied, such as sentence splitter, tokenization, POS tagger, lemmatizer, and term identification. In the experimental activity on the given corpus, a task concerning the categorization of CDs was included. Legal experts were involved in identifying classification classes, different from just the civil legal topic, so as to consider also the procedural phase of the CDs, and not only civil law core content. After that, the legal experts have performed a manual semantic characterization of each category. Decisions were assigned to one or more categories and an accuracy test was run against a set composed of 328 decisions labelled by a legal expert. The test showed promising results. A different ontology-based approach aimed at extracting information from Brazilian CDs is presented in **Araujo et al. [2017]**. The system is based on a domain ontology of legal events and a set of linguistic rules, integrated through an inference mechanism. To identify these legal events in CDs, the domain ontology called ODomJurBR was used. CDs used for the experiment were selected from documents returned by a query performed on the official CDs research tool, available at the Brazilian State Superior Court (TJRS) website. CDs were mined

also in **Savelka et al. [2019]**. This work is the development of the approach proposed in Savelka and Ashley [2016], where sentence retrieval techniques are applied on CDs to extract the meaning of statutory terms. However, this development was tested on a dataset of 4,635 sentences, that were provided as responses to three statutory queries. CDs were assembled and labelled in terms of their usefulness for interpretation. The dataset was obtained from the Caselaw Access Project (CAP) provided by Harvard Law Library. This time, IR approaches were applied. These included i) measuring the similarity between the sentence and the query, ii) using the context of a sentence, iii) expanding queries, or iv) assessing the novelty of a sentence with respect to a statutory provision from which the interpreted term comes. Results confirmed that i) retrieving the sentences directly by measuring similarity between the query and a sentence yields mediocre results and that ii) taking into account sentence context turns out to be the crucial step in improving the ranking performance.

Table 3.3: Legal knowledge extraction from case law - other approaches overview.

Approach	Name	Approach	Dataset	Main Task
Dick [1991]	Representation of Legal Text for Conceptual Retrieval	Ontology-based and Text Processing	US Case Law	Information Retrieval
Agnoloni T. [2014]	Legal Keyword Extraction and Decision Categorization	NLP	Italian Case Law	Knowledge Extraction
Araujo et al. [2017]	Automatically Classifying Case Texts and Predicting Outcomes	Ontology-based	Brazilian Case Law	Text Classification
Savelka et al. [2019]	Improving Sentence Retrieval from Case Law for Statutory Interpretation	Vector Similarity Measure	US Case Law	Information Retrieval
	CRIme Knowledge Extraction (CRIKE)	Ontology-based and Text Analysis	US Case Law	Knowledge Extraction

3.3.2 Machine Learning-based Approaches

In this subsection, the knowledge extraction approaches based on machine learning techniques and applied on case law datasets are described. An overview of the main contributions is presented in Table 3.4.

In **Ashley and Brüninghaus [2009]** the authors present SMILE + IBP (SMart Index LEarner + Issue-Based Prediction), a program that bridges case-based reasoning and extracting information from texts. From a legal perspective, the program aims at extracting information from textual descriptions of facts inside decided cases and at applying that information to predict the outcomes of new cases. In other words, the program attempts to automatically classify textual descriptions of the facts of legal problems in terms of Factors (e.g. a set of classification concepts that capture stereotypical fact patterns that affect the strength of a legal claim). This work reports an experiment on trade secret misappropriation. SMILE employs a training set of manually classified case texts by hired law students. More precisely, the entire opinion was not annotated but only the so-called *scuibs*, e.g. brief summaries of the main facts of the cases. Sentences were represented as feature vectors following three kinds of representations: Bag fo Word (BOW), Roles-Replaced (RR) and Propositional Patterns (ProP). These three representations were then compared to check their vector similarity. Three ML algorithms were tested: a decision tree algorithm, Naïve Bayes classifier and k-nearest neighbor (k-NN). From the discussion of the very last results it appears that, even if the accuracy of SMILE's classifications is too poor to enable Issue-Based Prediction, (to help the user pose and assess hypotheses about how the problem should be decided using the cases retrieved) SMILE + IBP should be seen as a kind of existence proof: it demonstrates the feasibility of a program reasoning on legal cases, taking texts as input. Mining CDs to extract sentences that deal with the meaning of vague terms is proposed in **Savelka and Ashley [2016]**. This work aims at automatically retrieving the set of useful sentences for terms interpretation. To this end, the authors first collected the interpretation data, e.g. the sentences where the interpretation of the vague term is reported. Then, a small set of sentences mentioning the term was extracted from the top 20 CDs retrieved from the Court Listener database and then labelled. Issues regarding the agreement between the annotators emerged, also because the annotators did not have an opportunity to practice before they started with the annotation. A list of other features was added, including, i) Semantic Re-

lationship between the vague term of interest as used in the statutory provision and the same term as used in the retrieved sentences; ii) Syntactic Importance, e.g. how dominant it is the term in the retrieved sentence when using a syntactic parsing technique; iii) Structural Placement, e.g. the place of the retrieved sentence and the term of interest in the structure of the document; iv) Rhetorical Role, e.g. the rhetorical role that the retrieved sentence has in the document; v) Attribution, e.g. who has uttered the retrieved sentence; vi) Assignment/Contrast, e.g. whether or not the vague term of interest, in the retrieved sentence, is assigned (or not) to some other terms (in order to model this category, the authors used pattern matching on the verb phrase of which the term of interest is part); vii) Feature Assignment e.g. if the vague term of interest in the retrieved sentence is said to be a feature of another term (or vice versa). The greatest limitation of this approach was that acquiring the labels was very expensive. **Jackson et al. [2003]** This work presents an information extraction and retrieval system, called History Assistant. The system aims at extracting rulings (e.g. judge orders) from CDs and retrieve relevant prior cases from a citation database (e.g. that is a database where citations of prior CDs are stored in common law systems). To this end, the authors discuss a combination of information retrieval and machine learning techniques to link each new case to its related documents. The authors confirmed that modeling argument structure is a difficult procedure that requires sophisticated and highly customized knowledge of representation techniques. **Saravanan and Ravindran [2010]** puts forward a system for labelling sentences with their rhetorical roles and extracting structured headnotes (brief summaries of the judgments made by legal experts) automatically from CDs. The authors present a Conditional Random Field to perform document segmentation aimed at identifying the rhetorical roles in CDs. An annotated dataset was created with the help of legal experts and is used as training data. The experts then annotated resources for the three domains of the Indian civil law. Based on the manually annotated CDs, the annotation scheme assigns a label indicating the rhetorical status of each sentence in a specific portion of a document, for identifying: case facts, case history, arguments, ratio decidendi and final decisions. This work confirms once again the need for expert manual annotation. In **Moens et al. [2007]** authors discuss the results of an experiment for the detection of arguments in legal texts. Detection of arguments refers to the identification of a viable argument to claim. The detection of arguments in CDs was carried out as a text classification task. For this reason, a Multinomial Naive Bayes classifier was trained on a set of manu-

ally annotated sentences representing arguments. Different feature sets were evaluated involving lexical, syntactic, semantic and discourse properties of the texts. The experiment is the first step in the context of automatically classifying arguments in legal texts according to their rhetorical type. A conceptual legal document retrieval experiment is presented in **Grabmair et al. [2015]**. In this work, the conceptual mark-up of documents is done automatically using LUIMA, a law specific semantic extraction toolbox based on the UIMA framework. This framework is an open-source Apache framework that has been deployed in several large-scale government-sponsored and commercial text processing applications. The proposed system consists of modules for automatic sub-sentence level annotation, machine learning-based sentence annotation, basic retrieval using Apache Lucene and a machine learning-based re-ranking of retrieved documents. Results demonstrate the feasibility of implementing a conceptual legal document retrieval system going from natural language legal documents to retrieval results for a restricted set of documents in the domain of vaccine injury claims. In this work, the authors aim at supporting this process with automated means, i.e., the automated recognition of an argumentation structure and its arguments in a legal text and the classification of an argument, or set of arguments, according to its argument type (e.g. counter argument or rebuttal). The simple features tested already yield promising results while attaining accuracy of the classification of almost 74% averaged over a variety of text types. This work addressed the first approaches to the innovative area of research of argument detection and classification in legal texts. **Mochales and Moens [2011]** present the results of an experiment on argumentation mining in CDs. Again, unavoidably expert manual annotations imposed an expensive contribution from three annotators, who spent more than a year annotating arguments in the European Court of Human Rights (ECHR) corpus, under the supervision of a judge. After this, sentences were represented as feature vectors. In order to learn how to identify sentences as argument propositions, features extraction techniques were applied, in order to extract: words (themselves, pairs or triples); part of speech (POS), including verbs and adverbs modal auxiliaries such as: may, must, shall, should; punctuation; keywords, such as: but, consequently, because of; depth of phrase tree; and certain text statistics such as sentence length, average words length, and a number of punctuation marks. The feature values were represented as binary features. Once the sentences had been represented as feature vectors, the authors applied three machine learning algorithms, namely Naïve Bayes classifier, Multinomial Logistic Regression

and SVM. The results have indicated that authors achieved an accuracy of 74% and 80% on the Araucaria and on the ECHR, respectively.

Table 3.4: Legal knowledge extraction from case law - machine learning-based approaches overview.

Approach	Name	Approach	Dataset	Training set Creation	Main Task
Thompson [2001]	Automatic Categorization of Case Law	k-NN, Decision Tree, Context Sensitive Learning	US Case Law	Manual	Text Classification
Jackson et al. [2003]	History Assistant - Prior Cases Retrieval (PCR)	Semantic Model and SVM	US Case Law	Pre-annotated	Information Retrieval
Gonçalves and Quaresma [2005]	Classification of Legal Texts	SVM	Portuguese Case Law	Manual	Text Classification
Moens et al. [2007]	Automatic Detection of Arguments in Legal Texts	Multinomial Naïve Bayes	Araucaria Corpus	Manual	Argumentation Mining
Ashley and Brüninghaus [2009]	SMILE+ IBP (SMart Index Learner Plus Issue-Based Prediction)	Decision Tree, Naïve Bayes, k-NN	US Case Law	Manual	Text Classification
Mochales and Moens [2011]	Argumentation Mining	Naïve Bayes, Multinomial Logistic Regression and SVM	Araucaria Corpus	Manual	Argumentation Mining
Saravanan and Ravindran [2010]	Segmentation and Summarization of a Legal Judgment	Random Fields	Indian Case Law	Manual	Information Extraction

Continued on next page

Table 3.4 – *Continued from previous page*

Grabmair et al. [2015]	Legal UIMA (Unstructured Information Management Architecture)	Decision Tree, Logistic Regression and Naïve Bayes	US Case Law	Manual and Semi-automatic	Knowledge Extraction
Savelka and Ashley [2016]	Extracting Case Law Sentences for Argumentation	Naïve Bayes, SVM and Random Forest	US Case Law	Manual	Information Retrieval

3.4 Thesis Contribution

Most of the machine learning-based approaches report experiments where a classifier is trained on a manually annotated dataset of CDs (or of CDs sentences). From the literature comparison it emerged that manual annotation of these documents implies several peculiar difficulties, such as:

- *agreement*: agreement on legal topics means high-level technical discussions and decision-making processes that may slow down manual annotation, as reported in Savelka and Ashley [2016].
- *time*: high-level technical manual annotation needs time, sometimes as long as years, as reported in Mochales and Moens [2011].
- *technical background*: legal manual annotation requires strong domain knowledge and high education, as reported in Saravanan and Ravindran [2010] and in Francesconi et al. [2010], where legal texts, before been processed, are topic-dependent, pre-selected and analysed by a group of legal experts.
- *costs*: involvement of experts with domain knowledge and high education may require a budget, as reported in Ashley and Brüninghaus [2009].

In the thesis, we propose the CRIKE (CRIme Knowledge Extraction) approach. Legal Knowledge Extraction in CRIKE is based on multi-label classification techniques that aim to associate CDs with appropriate concepts in the LATO ontology. By using information retrieval techniques the aim is to automatically annotate CDs, at diverse

levels of granularity, and to enforce knowledge extraction from a given corpus of CDs, by discovering new terms that have been recognized in the CD texts with which to enrich the terminology associated with legal concepts in the ontology. In particular, the contribution of CRIKE is twofold:

1. An approach for knowledge extraction based on an explicit legal knowledge model.
2. A CDs annotation technique based on a limited legal terminology representation, at first realised by a domain expert and then automatically enriched by CRIKE performing a cyclic process.

Chapter 4

The CRIKE (CRIme Knowledge Extraction) Data Science Approach

4.1 Introduction

The CRIKE approach enforces knowledge extraction and enrichment based on a given corpus of CDs using a reference legal domain ontology called LATO (see Figure 4.1). The goal of CRIKE is to progressively enrich the knowledge specified in the LATO

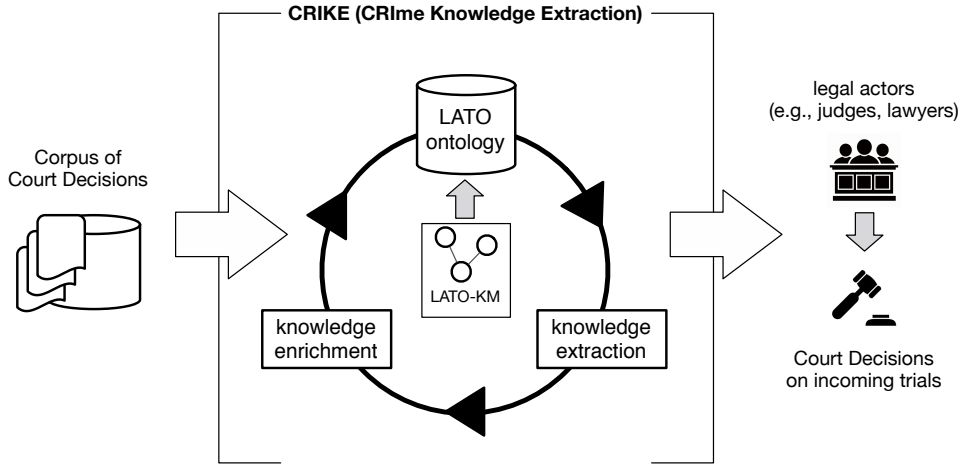


Figure 4.1: The CRIKE approach to knowledge extraction and enrichment.

ontology by extracting concrete terminology associated with ontology concept occurring in the considered document corpus. At the beginning, CRIKE relies on an initial version of the LATO ontology where domain experts manually define a starting set of legal concepts of interest with associated an initial set of terms (term-set). CRIKE is enforced as a cyclic incremental approach where the execution of knowledge extraction and knowledge enrichment tasks produces a new enriched version of the LATO ontology. The enrichment task consists in discovering new terms that have recognized in the CD texts, to populate term-sets of legal concepts. This new ontology version is then exploited to trigger the execution of a new CRIKE cycle to further enrich the LATO ontology. The enforcement of CRIKE cycles is stopped when it is not possible to detect/extract new terms to insert in the LATO ontology. The knowledge currently-available in the LATO ontology, both concepts and related terminology, can be exploited to support legal actors such as judges and lawyers in managing new incoming legal trials and taking appropriate CDs. In the following, we describe the knowledge model of the LATO ontology, the process and the related techniques of knowledge extraction and enrichment in CRIKE.

4.2 LATO-KM and LATO ontology

The legal knowledge model of LATO captures and formalizes the features and nature of terminology used in law and CDs. A design challenge is to find a suitable way of modeling the different nature of terms appearing in law and CDs as well as their meaning and roles. To model legal knowledge and capture these requirements, we define LATO-KM, a three-layer knowledge model based on the following entities and relationships (see Fig. 4.2):

- *Legal concept*: a legal concept C_i denotes a general rule/fact/element defined in the law (e.g., *Act*, *Illinois Controlled Substances Act*) and it is labelled with the terminology that appears in law texts. Legal concepts constitute the intermediate layer of LATO-KM.

- *Term-set*: a term-set T_i represents the concrete interpretation of a legal concept C_i in form of a set of terms occurrences that can be found in CDs. A term in a term-set is a string of characters of the language of the CDs; also multi-term expressions are considered as terms in LATO (e.g. both *Illinois Contr. Sub.*

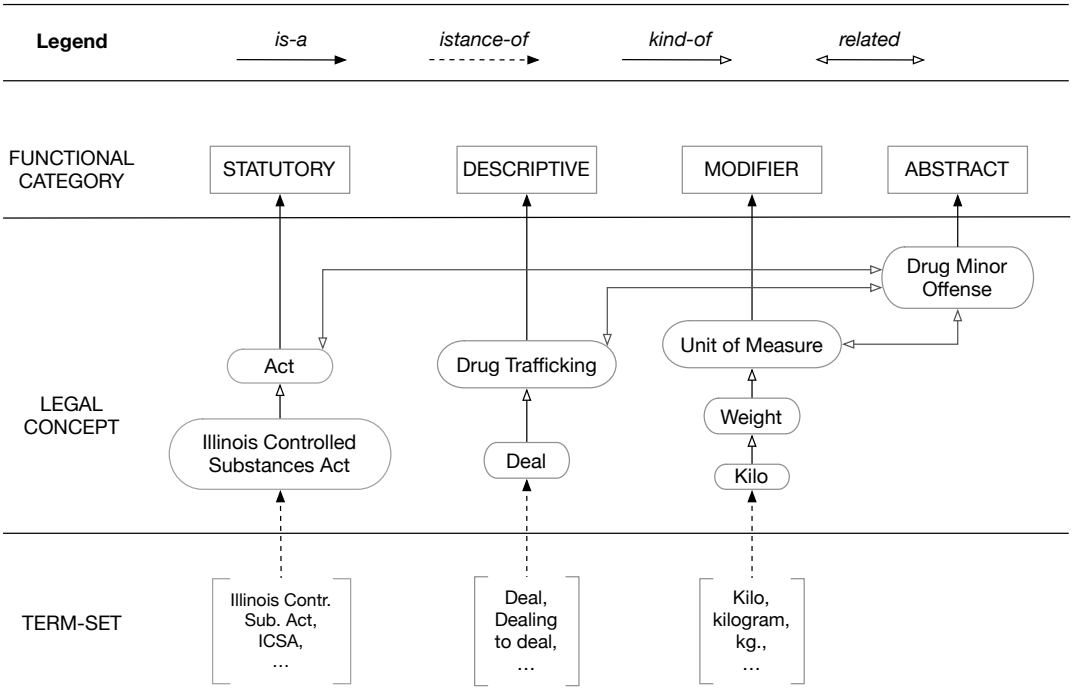


Figure 4.2: Overview of LATO-KM with an example of legal knowledge modeling about drug crimes.

Act and the acronym *ICSA* are terms). Term-sets constitute the bottom layer of LATO-KM.

- *Functional category*: a functional category represents the different kinds/roles of legal concepts in the law formulation, namely descriptive, statutory, modifier, and abstract, respectively. A **statutory category** describes a legal concept featuring something that is directly or indirectly defined in the law specification itself (e.g., *Act*). A **descriptive category** describes a legal concept featuring actions, human activities, and any real-life object in the law specification (e.g., *Drug Trafficking*). A **modifier category** describes a legal concept featuring quantitative/qualitative aspects of things/actions in the law specification (e.g., *Weight*). An **abstract category** describes a legal concept featuring something indeterminate that requires a concrete application for being really defined (e.g., *Drug Minor Offense*) Castano et al. [2019]. Functional categories constitute the top layer of LATO-KM.

According to LATO-KM, the concrete meaning of legal concepts is fully defined by referring to the specific terminology (i.e., term-set) that appears in real CDs. Moreover, legal concepts are classified with respect to the role they play in the law formulation using functional categories. Formally, a legal concept C_i is defined as 3-uple of the form:

$$C_i = hn(C_i), \mathrm{C}(C_i), T_i \ i$$

where:

- $n(C_i)$ is the label of the legal concept;
- $\mathrm{C}(C_i)$ *2* fSC, DC, MC, ACg is the functional category of C_i, either statutory (*SC*), descriptive (*DC*), modifier (*MC*), or abstract (*AC*).
- $T_i \ = ft_1, \ldots, t_n g$ is the term-set of the concept C_i, namely the language terms concretely used in legal document corpora (i.e., CDs) to refer to C_i. The asterisk symbol ("*") denotes optionality, in that we may have some legal concepts not yet associated with a corresponding term-set. For instance, abstract concepts are not directly associated with a specific term-set, but rather they are indirectly expressed through the term-sets associated with the legal concepts to which the abstract concept is related.

Intra- and inter-layer relationships are defined in LATO-KM to capture the semantic relationships that hold between pairs of entities. The following intra-layer binary relationships are defined in LATO-KM:

- **Term-to-Term**: it is a binary relationship between a pair of terms t and t^0 in a term-set T_i at the bottom layer, that holds due to either a morphological or a linguistic relationship between terms. Examples of morphological relationships are:
  - paradigm (e.g. *to deal - dealt - dealt*)
  - conjugation – for verb (e.g. *dealt - deals - dealing*)
  - declension – for nouns (e.g. *drug - drugs - drug's*)
  - abbreviation (e.g. *Illinois Contr. Sub. Act - ICSA*)

– string similarity (e.g. *Substances Act - substances act - Subs. Act*).

An example of linguistic relationship is synonymy (e.g., *Paragraph - Section*).

- **Concept-to-Concept**: it is a binary relationship between two legal concepts C_i and C_j at the intermediate layer, capturing semantic relationships holding between them in the law formulation. In particular, we introduce the kind-of relationship between two concepts to represent a generalization/specialization relationship between them. For example, *Cocaine* kind-of *Drug* is defined to express the fact that the former is a more specific crime than the latter in the law. Moreover, we introduce the related relationship between two concepts to represent a generic positive relationship between them. For example, *Drug Trafficking* related *Drug* is defined to express the fact that the crime of drug minor offence involves detention of drug in some quantity.

The following inter-layer binary relationships are defined in LATO-KM:

- **Term-to-Concept**: it is a binary relationship between a term $t\ 2\ T_i$ and a legal concept C_i denoting that C_i can be "lexicalized" by t in a CD text. A Term-to-Concept relationship is defined through the instance-of relationship for each term $t\ 2\ T_i$ and the corresponding legal concept C_i at the intermediate layer of LATO-KM. For example, *ICSA* instance-of *Illinois Controlled Substances Act* is defined to express the term *ICSA* belongs to the term-set of the concept *Illinois Controlled Substances Act*.

- **Concept-to-Category**: it is a binary relationship between a legal concept C_i and a functional category $C(C_i)$ expressing the nature of the concept in the law formulation. A Concept-to-Category relationship is defined through the is-a relationship. *Act* is-a *Statutory* is defined to express that the notion of *Act* is directly defined in the law.

4.2.1 The LATO Ontology Structure

The LATO-KM is implemented in a LATO ontology by using the Simple Knowledge Organization System (SKOS) which is defined in Isaac and Summers [2009]. SKOS is an RDF vocabulary for representing semi-formal knowledge organization systems

(KOSs), such as thesauri, taxonomies, classification schemes and subject heading lists. SKOS provides a lightweight, intuitive conceptual modeling language for developing and sharing new KOSs and for this reason we decided to choose SKOS for implementation of the LATO ontology. The legal concepts of the intermediate layer are

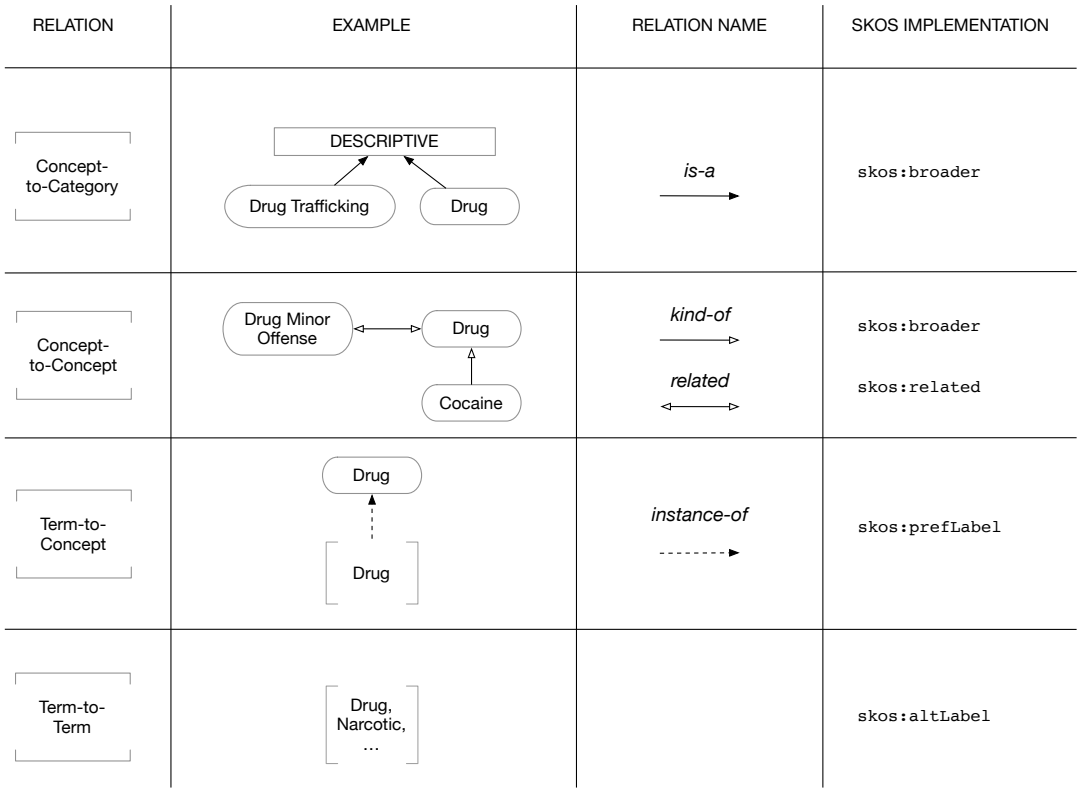


Figure 4.3: SKOS concepts and relations of the LATO ontology.

implemented as SKOS concepts in LATO (see Fig. 4.2). Concept-to-Concept relationships are specified through a corresponding SKOS relation. In particular, the kind-of relationship of LATO-KM is specified through the skos:broader relation. For instance, a skos:broader relation is defined between the concept Cocaine and the concept Drug. The Related relationship of LATO-KM is specified through the skos:related relation. For instance, a skos:related relation is defined between the concept Drug Trafficking and the concept Drug.

The term-sets of the bottom layer are implemented using labels of SKOS concepts. In particular, for each SKOS concept i) a skos:prefLabel is defined to implement the

instance-of relationship, and ii) a number of skos:altLabel are defined to implement the various Term-to-Term relationships denoting possible alternative terms for the considered SKOS concept. For instance, a skos:prefLabel relation is defined between the Drug LATO concept and the Drug term, while a skos:altLabel relation is defined between the Drug term and the Narcotics term.

Finally, functional categories of the top layer are implemented as SKOS concepts, too. Concept-to-Category is-a relationships are expressed through the skos:broader relation. For instance, a skos:broader relation is defined between the Drug LATO concept and the Descriptive category concept.

4.2.2 Views in LATO

Legal concepts in the LATO ontology can be presented according to different views focusing on different intra-layer and inter-layer relations to provide a different perspective of analysis of the legal knowledge related to a target concept. Given a target legal concept C_i, the following ontology views are defined.

- *Law-oriented view*: this view works on concepts and relations at the intermediate and top layers of the LATO-KM, by returning C_i together with its neighborhood, $N(C_i) = fC_j \; j \; 9R(C_i,C_j)orR(C_j,C_i)g$, that is, all the legal concepts having a kind-of or a related-to relation with C_i. Moreover, for each legal concept in $N(C_i)$, its functional category is also provided, by exploiting inter-layer is-a relations Concept-to-Category.

- *Case Law-oriented view*: this view works on concepts and relations at the intermediate and bottom layers of the LATO-KM, by returning C_i together with all terms t in the term-set T_i, by exploiting Term-to-Term and Term-to-Concept relations, respectively.

The law-oriented view for a target legal concept C_i provides the information actually available in the ontology about legal concepts semantically related to the target due to law formulation. The law-oriented view has been conceived to support **legal analytics** tasks where the goal is to exploit the concept knowledge with associated constraints defined in the ontology (i.e., rules/fact/actions) by also providing details about the function played by each concept (i.e., statutory, descriptive, modifier, or abstract category).

The case law-oriented view for a target legal concept C_i provides the information actually available in the ontology about legal concepts and associated terminology extracted from CD documents.

4.3 Knowledge Extraction and Enrichment in CRIKE

The CRIKE workflow is defined as follows:

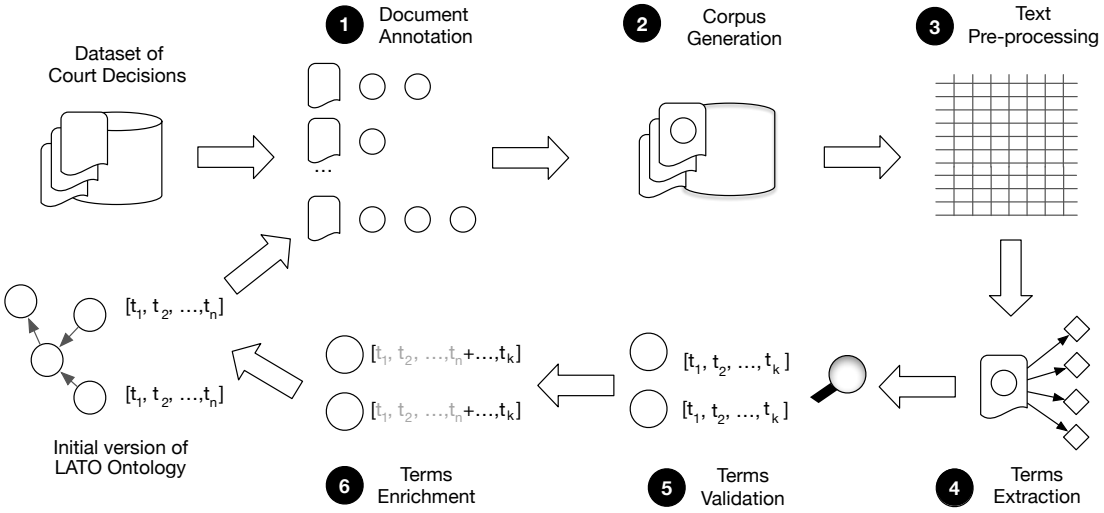


Figure 4.4: The CRIKE workflow.

Knowledge extraction in CRIKE aims at finding new and relevant terminology from CDs, that contributes to the definition of the legal concepts formalized in LATO. Legal knowledge extraction in CRIKE is articulated in six steps.

4.3.1 Document Annotation

CRIKE performs a multi-label annotation of CDs through term retrieval. More in detail, for each CD document d, the goal of the first task is to determine the set C_d of concepts associated with d:

$$C_d = \left\{ C_i : \sum_{t2T_i} w(t,d) \quad th \right\}$$

Table 4.1: Example of CD document sentences associated with the legal concept Drug.

d_1: [...] *Paragraph 14 of section 1 of the same act provides: "***Narcotic Drugs** *means leaves,* **opium**, **cannabis**, *and each substance neither chemically nor physically distinguishable from them."* [...]
d_2: [...] *Defendant, who was charged by indictment with violation of § 402 of the Illinois* **Controlled Substances** *Act"* [...]

where $w(t,d)$ is the weight that a term t, present in the term set T_i of the concept C_i, has in the document d, while th is a threshold used to set the minimum cumulative weight of all the terms t *2* T_i that is required for associating a corresponding concept C_i with the document d. The weight $w(t,d)$ is a coefficient which adjusts the score for matches relative to the length of the document, formalized as follows:

$$w(t,d) = f(0,5 \quad \mathrm{M}/\mathrm{T}) + 0,5g$$

where M is the cardinality of the set of terms matching with T, that is the cardinality of the set of terms of the documents, with stop-word removed. In Table 4.1, we report an example of two court decision fragments, d_1 and d_2 that are associated with the legal concept Drug. CRIKE annotates the documents attributing a set of labels, corresponding to the designed LATO concepts. An example is reported in Table 4.2

4.3.2 Corpus Generation and Text Pre-processing

For each concept C in C_d, CRIKE creates a pseudo-document D_C, that is the set of all documents $fd_1, d_2, ..., d_ng$ associated with the concept C in C_d as follows:

$$\mathrm{D}_C = fd_i : C \; 2 \; \mathrm{C}_dg$$

A corpus G is then generated by CRIKE with all the pseudo-documents $\mathrm{D}_{C_l}, ..., \mathrm{D}_{C_p}$ as follows:

$$\mathrm{G} = \mathrm{D}_{C_l} \; [\; ... \; [\; \mathrm{D}_{C_p}$$

Text pre-processing. On raw text, only tokenisation, lower case transformation, punctuation and stop-words removal are performed by CRIKE. An example is reported in Table 4.4.

Table 4.2: Example of the CRIKE multi-label document annotation.

Documents	Labels
d_1: "The **crack** [...] **plastic bag** [...] 20 **gr.**"	[drug, evidence, unit of measure]
d_2: "**cocaine** was found [...] **720 ILCS 570** and [...] **seizure** [...] controlled"	[drug, criminal procedure, illinois legislation]
d_3: "The apartment [...] **paraphernalia** [...] "	[evidence]
d_4: "Inside the car [...] **drug sale** [...] **arrest** during [...] "	[drug trafficking verbs, criminal procedure]
d_5: "The day before [...] **grams** [...] inspected"	[unit of measure]
d_6: "[...] **Illinois Controlled Substances Act** [...] orders [...] tested positive [...] **cocaine**"	[drug, illinois legislation]
...	
d_n: "... "	[...]

Table 4.3: Example of the pseudo-document creation.

Annotated Documents	Pseudo-document Creation (Drug)
d_1: "The crack [...] plastic bag [...] 20 gr." ! *labels:* {drug, evidence, unit of measure}	"The crack [...] plastic bag [...] 20 gr., [...] cocaine was found [...] 720 ILCS 570 and [...], Inside the car [...] drug sale [...] street [...]
d_2: "[...] cocaine was found [...] 720 ILCS 570 and [...]" ! *labels:* {drug, illinois legislation}	
d_3: "The apartment [...] paraphernalia [...]" ! *labels:* {evidence}	
d_4: "Inside the car [...] drug sale [...] street [...]" ! *labels:* {drug trafficking verbs}	...
...	"[...] Illinois Controlled Substances Act [...] tested positive [...] cocaine"
d_n: "[...] Illinois Controlled Substances Act [...] tested positive [...] cocaine" ! *labels:* {drug, illinois legislation}	

Table 4.4: Example of CRIKE text pre-processing.

Pseudo-Document (Drug)	
Before Text Pre-processing	**After Text Pre-processing**
"The crack [...] plastic bag [...] 20 gr., [...] cocaine was found [...] 720 ILCS 570 and [...], Inside the car [...] drug sale [...] street [...] Illinois Controlled Substances Act [...] tested positive [...] cocaine"	[crack, plastic, bag, 20, gr., cocaine, found, 720, ilcs, 570, inside, car, ..., drug, sale, street, illinois, controlled, substances, act, tested, positive, cocaine]

4.3.3 Terms Extraction

CRIKE aims at extracting the best terms, from each pseudo-document, that could be potentially used for enriching the terminology of the LATO concepts. To achieve this goal, CRIKE calculates the weights of the terms inside the pseudo-documents of the corpus G, using TF-IDF and PMI. TF-IDF is a statistical measure that permits us to evaluate how relevant a term t is to a document in our corpus G, by multiplying two metrics: how many times a term appears in a document tf, and the inverse document frequency of the terms across a set of documents idf. The terms-documents frequency matrix permits us to have the raw count of the term frequency tf. The calculation of the inverse document-frequency permitted us to compute the TF-IDF weights of the terms w.r.t. the pseudo-document D, as follows:

$$TF \quad IDF(t, \mathrm{D}, \mathrm{G}) = tf(t, \mathrm{D}) idf(t, \mathrm{G})$$

where

$$tf(t, \mathrm{D}) = log(1 + freq(t, \mathrm{D})) \quad \text{and} \quad idf(t, \mathrm{G}) = log\frac{1+n}{1+df(t)} + 1$$

where n is the total number of documents in the document set, and $df(t)$ is the number of documents, in the corpus G, that contain the term t. PMI is a statistical measurement that permits to calculate the probability of observing x and y together with the probabilities of observing x and y independently. Formally, PMI is presented as follows:

$$PMI(X, Y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

With this approach, it is possible to compare the probability of observing the term (t) and the pseudo-document (D) together with the probability of observing the term (t) and the pseudo-document (D) independently. With the terms-documents frequency matrix we calculated first $P(t, \mathrm{D})$ as follows:

$$P(t, \mathrm{D_C}) = \frac{freq_{(t,\mathrm{D})}}{N}$$

where N is the sum of all frequencies, formalized as follows:

$$N = \overset{t}{\mathring{a}} \overset{\mathrm{D}}{\mathring{a}} freq_{(t,\mathrm{D})}$$

Moreover, calculated N, it is possible to calculate the $P(t)$ and $P(\mathrm{D})$ as follows:

$$P(t) = \frac{\overset{\mathrm{D}}{\mathring{a}} freq_{(t,\mathrm{D})}}{N} \quad \text{and} \quad P(\mathrm{D}) = \frac{\overset{t}{\mathring{a}} freq_{(t,\mathrm{D})}}{N}$$

Finally, the PMI value between the term (t) and pseudo-document (D) is calculated as follows:

$$PMI_{(t,\mathrm{D})} = log_2 \frac{P(t, \mathrm{D})}{P(t)P(\mathrm{D})}$$

Finally, CRIKE ranks the terms by their weights, and considering the first top-k terms, it creates a set T_k for each pseudo-document, for both approaches (TF-IDF and PMI), as follows:

$$\mathrm{D}_{C_j} \; ! \; T_{k_j}^{(TF \quad IDF)}; T_{k_j}^{(PMI)} ..., \mathrm{D}_{C_n} \; ! \; T_{k_n}^{(TF \quad IDF)}; T_{k_n}^{(PMI)}$$

Knowledge enrichment in CRIKE aims at selecting the best extracted terms, that can be used now to enrich the terminology of the concept. It is important to recall that, as described in Section 4.2, LATO ontology provides an initial terms-set T_i for each concept, before the start of the CRIKE cycle. Thus, the enrichment phase is articulated in the last two CRIKE steps.

4.3.4 Terms Validation

CRIKE removes from the created set T_k the terms already present in the set T_i, creating the set of the new extracted terms T_{new}, as follows:

$$T_{new} = (T_k^{(TF \quad IDF)} \; [\; T_k^{(PMI)}) \quad T_i$$

The new extracted terms need to be validated. To this end, a legal expert is involved. The terms in the set T_{new} are manually validated by the legal expert in order to define the set $R_i \quad T_{new}$ containing the terms that are new and relevant for the concept C_i. In the validation step, the degree of relevance $h_{C_i}(t)$ is exploited by the expert i) to filter out terms whose association with the concept C_i is poor (i.e., low values of $h_{C_i}(t)$), and ii) to select terms whose association with the concept C_i is strong (i.e., high values of $h_{C_i}(t)$).

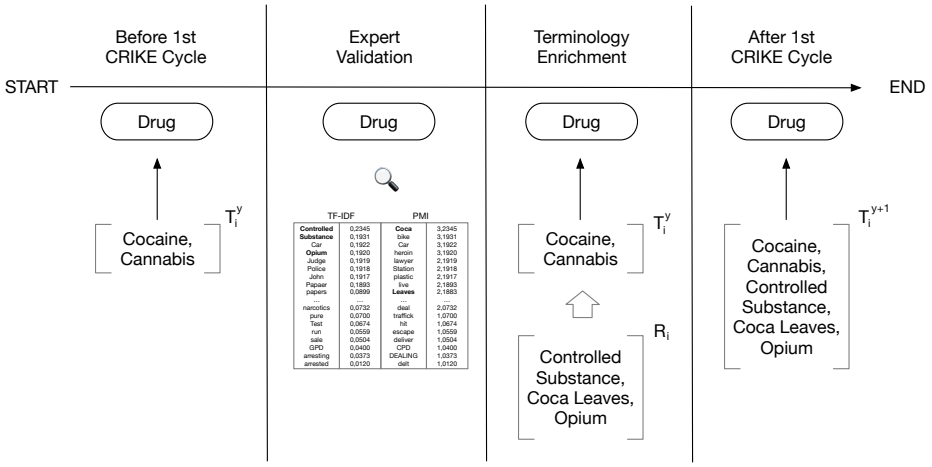


Figure 4.5: Example of the LATO concept *Drug* terms-set enrichment

4.3.5 Terms Enrichment

According to the expert validation, the term-set T_i associated with the concept C_i is enriched. Being y the current CRIKE cycle, enrichment is enforced as follows:

$$T_i^{y+1} \quad T_i^y \; [\; R_i$$

where T_i^y is the term-set initially associated with the concept C_i and T_i^{y+1} is the term-set associated with C_i after enrichment. An example is presented in Figure 4.5.

4.3.6 CRIKE Endpoint

On the basis of the results of the extraction step of the current cycle, a further CRIKE cycle could be activated to extract additional new terminology. Decision to run a new cycle is taken by the user on the basis of the new terminology extracted at current step. The user can decide to stop running CRIKE after the current cycle if extracted terminology is considered satisfactory or if poor new terminology has been extracted with respect to previous cycle. There is however a formal condition for CRIKE bootstrapping termination, that is, when no new terminology is extracted in current cycle, that is $T_{new} = \emptyset$. In this case, the user is notified that a new cycle is not activated (i.e., CRIKE reaches its endpoint).

Chapter 5

Experimental Results

In order to test the CRIKE approach, we perform an evaluation on real-world data. At first, we find a consistent dataset of CDs in the Caselaw Access Project (CAP) that provides public access to U.S. law (`https://case.law/bulk/download`) digitized from the collection of the Harvard Law Library. For the current evaluation, we run CRIKE against 180,000 decisions of courts of the State of Illinois. The dataset has been selected due to its data quality and data completeness. To run CRIKE, we design a LATO version compatible with both, the Illinois Criminal Law concerning the drug related offences, namely the Illinois Controlled Substances Act, and the related case law.

5.1 Goal

The goal of experimentation is to assess the capability of our approach to discover new and useful terms for enriching the term-sets of the legal concepts designed in LATO. CRIKE is evaluated considering three levels of effectiveness in terms extraction and enrichment.

- How many extracted terms are **correct**
  (that is, they are pertinent for a given LATO concept)
- How many extracted terms are **useful**
  (that is, they are new w.r.t. the actual term-set of a given LATO concept)

- How much the first version of LATO has been enriched?
  (that is, the comparison of the first version of LATO against the second version of LATO)

We investigated also the efficiency of CRIKE in terms of quality of the extracted terminology, by providing a detailed qualitative analysis, term by term, for each enriched LATO concept. These measurements are used to evaluate CRIKE after the first cycle and to answer the following explicit hypothesis: *it is possible to extract new and useful legal knowledge from a CDs dataset, without the need of expert-human reading and expensive manual-annotation of legal text?*

5.2 Methodology

The methodology of the current experimentation replicates the CRIKE cycle.

1. *Concepts selection.* First, we select a set of concepts and an initial set of terminology associated, creating the first version of LATO.

2. *Document annotation*. This first version of LATO guides the CDs annotation through term retrieval.

3. *Corpus generation*. The annotated CDs are labeled. We compose a pseudo-document for each label with the annotated CDs. Each pseudo-document is included in the corpus.

4. *Terms extraction.* Each pseudo-document is pre-processed and the weights of the terms are calculated with two approaches, namely TF-IDF and PMI.

5. *Terms evaluation*. The weighted terms are ranked and evaluated by a qualified criminal lawyer specialised in criminal law and in drug-related crimes, selected for his academic background and working experience.

6. *Terms enrichment*. The criminal lawyer evaluation permits us to find the best terms that enrich the LATO concepts and, finally, to calculate the CRIKE performance, in terms of correctness and usefulness.

1. Concepts selection. For the experiment, we select six concepts from our legal ontology all related to the abstract term Drug Trafficking, namely Drug, Drug Trafficking Verbs, Unit of Measure, Illinois Legislation, Criminal Procedure, and Evidence. A portion of LATO is presented in Figure 5.1.

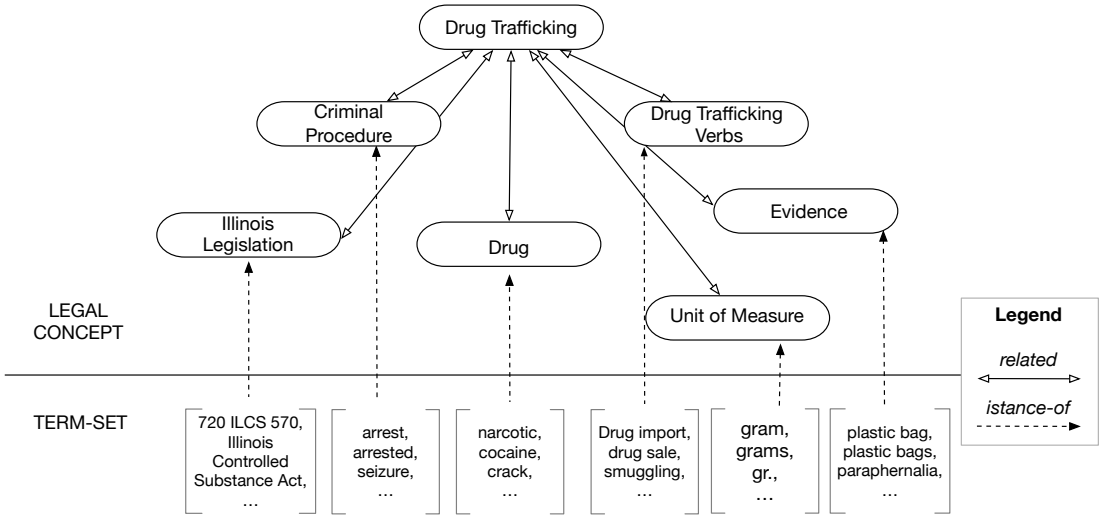


Figure 5.1: Example of the law-oriented view and the case law-oriented view in LATO.

The initial term-sets associated with the selected concepts have been manually defined by a legal expert and they are shown in Table 5.1.

Table 5.1: Term-sets Used for the Evaluation

Legal Concept	Term-set	Size
Drug	narcotic, cocaine, crack, [...]	15
Drug Trafficking Verbs	Drug import, drug sale, smuggling, [...]	18
Unit of Measure	gram, grams, gr., [...]	10
Illinois Legislation	720 ILCS 570, Illinois Controlled Substances Act [...]	6
Criminal Procedure	arrest, arrested, seizure, [...]	7
Evidence	plastic bag, plastic bags, paraphernalia	3

2. Document Annotation. We decide to work at the granularity level of sentences. The dataset is composed of 14,000,000 sentences derived from about 180,000 decisions of courts of the State of Illinois taken from the Caselaw Access Project (CAP).

We annotate the sentences with CRIKE, setting *th* to 1. In order to have a balanced corpus of documents for each LATO concept, we set an annotation limit to 40.000 sentences for each concept. Sentence annotation details are reported in Table 5.2.

Table 5.2: Details of the Documents Annotation Task

Concepts (*C*)	**Annotated Sentences for Concept (*d*)**
Drug	25.138
Drug Trafficking Verbs	40.000
Unit of Measure	7.241
Illinois Legislation	40.000
Criminal Procedure	40.000
Evidence	33.417
Unique Sentences: 158.398	

3. Corpus generation. With the annotated documents we create a pseudo-document D for each concept *C*, including all the sentences associated with the given concept. For example, the pseudo-document of the concept drug (D_{Drug}) contains all the sentences associated with the label `drug` by CRIKE. In total, six pseudo-documents are present in the corpus G.

4. Terms extraction. Each pseudo-document is pre-processed. Tokenisation, lower case transformation, punctuation and stop-words removal are performed, using *Spacy*, a free open-source library for Natural Language Processing in Python accessible at `https://spacy.io` and reported in Honnibal and Montani [2017]. That process permitted us to create a terms-documents frequency matrix on which we test two terms-weight calculations, namely, TF-IDF and Pointwise Mutual Information (PMI). Due to PMI weights are very high for unfrequent terms, we considered only terms with a frequency greater or equal to 20.

5. Terms evaluation. We rank of the top-40 terms by their weights. These top-40 terms are evaluated by a qualified criminal lawyer specialised in criminal law and in drug-related crimes, selected for his academic background and working experience. Considering both approaches and the six concepts formalised in LATO, we send for

the evaluation 480 terms in total. We inform the involved criminal lawyer that the terms have been computed by CRIKE, and we ask him to evaluate these terms by also considering intra- and inter-layer relationships as defined in LATO- KM. Moreover, we ask to express a judgment and a degree for each extracted term according to the grid reported in Table 5.3.

Table 5.3: The Grid Used for the Terms Evaluation

Judgment	Degree	Resulting Set
Relevant and new	2	T^2
Relevant but present	1	T^1
Pertinent but not relevant	0	T^0
Non - pertinent (error)	-1	T^{-1}

We decide to associate each judgement with a degree:

- 2 (two) to the terms new and relevant, i.e., the terms that are considered useful to enrich the terms-set of the concept
- 1 (one) to the terms that are already defined in terms-set of the concept
- 0 (zero) to the pertinent terms that belong to the domain, but that are not associated with the specific concept
- -1 (negative one) to the terms that are not pertinent, and are considered as errors

The terms with the same judgment/degree are inserted in a dedicated resulting set. In particular, in T^2 there are all the terms evaluated with 2 (two), in T^1 there are all the terms evaluated with 1 (one), in T^0 there are the terms evaluated with 0 (zero), and finally, in T^{-1} there are the terms evaluated with -1 (negative one). As an example, we report in Table 5.4 the terminology extracted from the pseudo-document drug (D_{drug}) and the related evaluation details.

5.3 Evaluation

We calculate the correctness and the usefulness of CRIKE, answering the goal of the evaluation defined before. We define the correctness (*CO*) and the usefulness (*US*) of

Table 5.4: Example of the Terms Evaluation

PSEUDO-DOCUMENT: Drug						
	TF-IDF			**PMI**		
	term	*weight*	*degree*	*term*	*weight*	*degree*
1	traffickers	0,2306	0	arres	3,2921	-1
2	unprescribed	0,1931	2	microscopic	3,1824	0
3	addicts	0,1709	2	smoked	3,1129	0
...	...	...	...	...	...	...
38	hershey	0,0596	2	cracked	2,5430	2
39	psilocyn	0,0596	2	lsd	2,5415	1
40	propagate	0,0569	-1	marijuana	2,5385	1

CRIKE as follows:

$$CO = \frac{jT^1 \,[\, T^2 j}{jT^0 \,[\, T^{-1} j} \quad \text{and} \quad US = \frac{jT^2 j}{jT^1 \,[\, T^0 \,[\, T^{-1} j}$$

where CO is the fraction of the relevant terms on the non-relevant terms, and US is the fraction of the useful terms on the non-useful terms.

Quantitative Results. First, the quantitative results represent the CRIKE performance for each concept, by comparing, i) the cardinality of the initial term-set T_i (*old*) with ii) the cardinality of each resulting set (T^2, T^1, T^0, T^{-1}) of the first CRIKE cycle (*1st* CRIKE *cycle*) and with iii) the enriched term-set T_i (*new*), after the first CRIKE cycle. Secondly, the performance of CRIKE is calculated in terms of correctness (CO) and usefulness (US). The quantitative results after the 1st CRIKE cycle are reported in Table 5.5.

We observe that TF-IDF performs better in terms of correctness, revealing a stronger capability in extracting terms already defined in the first version of LATO. Considering our evaluation goal, the current experiment confirms that, from a quantitative point of view, in terms of correctness, the TF-IDF performs better, but in terms of the usefulness PMI reaches better performance. To observe the general performance of both approaches, we calculate the harmonic mean for both TF-IDF and PMI as follows:

$$\mathrm{H} = \frac{2(CO)(US)}{(CO)+(US)}$$

Table 5.5: The Quantitative Results after the 1st CRIKE cycle (y=0)

		Drug	Drug Trafficking Verbs	Unit of Measure	Illinois Legislation	Criminal Procedure	Evidence	**Tot.**
TF-IDF								
old	$jT_i^y j$	15	18	10	6	7	3	59
	$jT_k j$	40	40	40	40	40	40	240
1st	$jT^2 j$	18	7	10	10	24	15	84
CRIKE	$jT^1 j$	0	8	3	6	4	3	24
cycle	$jT^0 j$	7	20	16	14	10	16	83
	$jT^{-1} j$	15	5	11	10	2	6	49
new	$jT_i^{y+1} j$	33	25	20	16	31	18	143
CO	$\frac{jT^1[T^2 j}{jT^0[T^{-1} j}$	0,82	0,60	0,48	0,67	2,33	0,82	**0,82**
US	$\frac{jT^2 j}{jT^1[T^0[T^{-1} j}$	0,82	0,15	0,34	0,38	1,14	0,68	0,51
H	$\frac{2(CO)(US)}{(CO)+(US)}$	0,82	0,24	0,40	0,49	1,53	0,74	0,63
PMI								
old	$jT_i^y j$	15	18	10	6	7	3	59
	$jT_k j$	40	40	40	40	40	40	240
1st	$jT^2 j$	18	13	19	9	19	14	92
CRIKE	$jT^1 j$	4	0	4	1	2	0	11
cycle	$jT^0 j$	9	8	0	13	7	4	41
	$jT^{-1} j$	9	19	17	17	12	22	96
new	$jT_i^{y+1} j$	33	31	29	15	26	17	151
CO	$\frac{jT^1[T^2 j}{jT^0[T^{-1} j}$	1,22	0,48	1,35	0,33	1,11	0,54	0,75
US	$\frac{jT^2 j}{jT^1[T^0[T^{-1} j}$	0,82	0,48	0,90	0,29	0,90	0,54	**0,62**
H	$\frac{2(CO)(US)}{(CO)+(US)}$	0,98	0,48	0,1,08	0,30	0,99	0,54	**0,68**

and we calculate, for both approaches, the following harmonic mean on the totals:

$$H^{(TF\ \ IDF)} = 0.63 \quad \text{and} \quad H^{(PMI)} = 0.68$$

The harmonic mean indicates the better performance of PMI. However, considering the extracted terminology, and the lawyer evaluation, the concepts terms-sets are considerably enriched with both approaches. We calculate the rounded enrichment percentage by comparing T_i before and after the first CRIKE cycle and we report an overview in Figure 5.2.

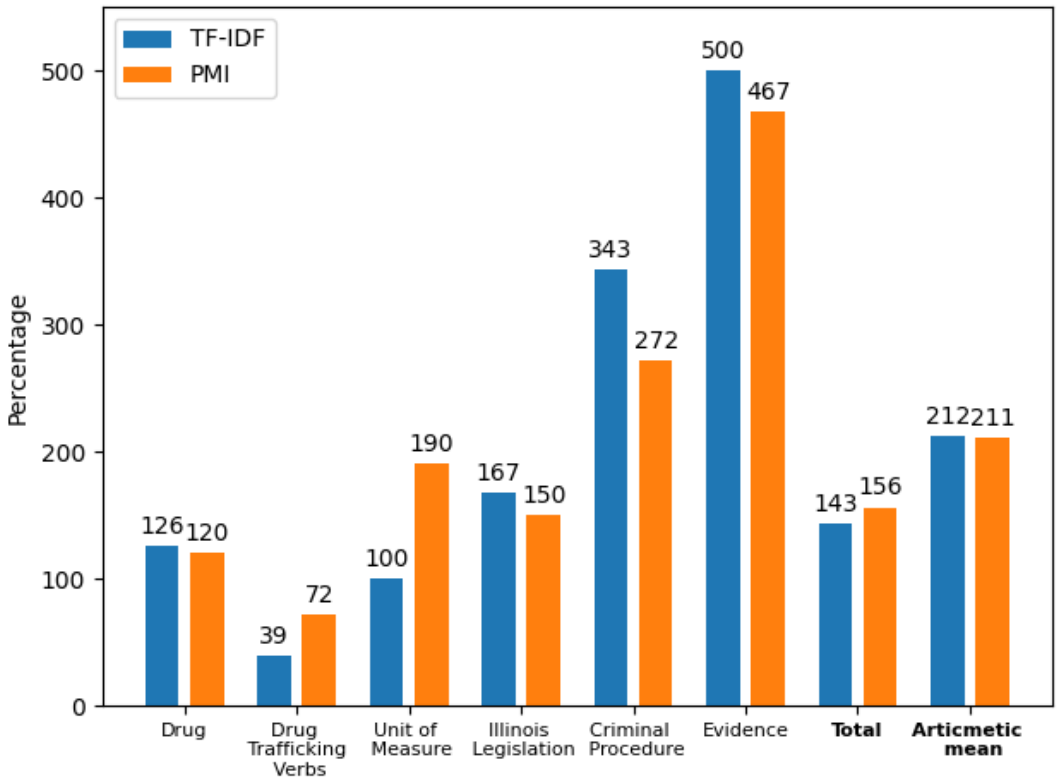


Figure 5.2: The rounded percentages of the terms-set enrichment for each concept.

More in detail, the concept *Drug* reports an incrementation of its terms-set of 126% with TF-IDF and of 120% with PMI. *Drug Trafficking Verbs* is enriched by the percentage of 39 % with TF-IDF, and by the percentage of 72% with PMI. The terminology of the concept *Unit of Measure* presents an enrichment of 100% with TF-IDF and of 190% with PMI. *Illinois Legislation* reports an incrementation of its terms-set of 167% by using TF-IDF and of 150% by using PMI. *Criminal Procedure* reports a consistent enrichment: 343% using TF-IDF and 272% using PMI. The concept *Evidence* is enriched the most, 500% and 467% with TF-IDF and PMI, respectively. On the total, the

enrichment percentage with TF-IDF is 143%, and with PMI is 156%, but by considering the arithmetic means of the enrichment performances, both approaches performed similarly.

Qualitative Results. From the ranked top-40 terms, the lawyer highlights a set of new terminology for each concept, that includes verbs conjugations, paradigms, declinations, abbreviations and very short acronyms, unknown and unpredictable during the LATO modelling. Moreover, CRIKE makes possible the extraction of particular categories of terms such as, the idiomatic expressions, the slang or jargon, and terms that belong to domain contexts different from the legal one, that are, for instance, the biological context, the medical context and the social-service context. This terminology enriches the general knowledge, completing one of our research goals. A second important result is the discovery of terminology used by legal professionals, which is the legal knowledge reported in CDs. Furthermore, CRIKE extracts also terms that are already formalized in LATO and terms that are related to the general domain of the criminal law.

Drug. In Table 5.6 we present the top-40 terms extracted by CRIKE for the concept Drug with the associated weights, calculated with TF-IDF and PMI, and the judgement degree expressed by the criminal lawyer.

With TF-IDF we notice at first new and relevant terms from the legal context, such as, *unprescribed, noncontrolled*, usually used with substances or drugs. Secondly, we find acronyms such as, i) *cr* for Controlled Release, ii) *gcms* for Gas Chromatography Mass Spectrometry, that is a forensic drug examiner, and iii) *thc* for Tetrahydrocannabinol, all related to the biological domain. Another term from the same domain is *psilocyn*, a hallucinogenic alkaloid isolated in trace amounts from Psilocybe mushrooms[1]. CRIKE finds also idiomatic terms considered as slang. These are *hershey* a famous chocolate company that produces a chocolate bar, used to represent cannabis, *blues*, a term used for Fentanyl and Fentanyl Derivatives and Oxycodone, also combined with other terms, such as *Flat Blues* or *Royal Blues* for LSD (Lysergic Acid Diethylamide) and *French Blues* for Amphetamine as reported in the DEA Intelligence Report[2], and

[1] https://pubchem.ncbi.nlm.nih.gov/compound/4980.

[2] https://www.campusdrugprevention.gov/sites/default/files/ Also, *blues* could refer to Fentanyl and Fentanyl Derivatives, as described in the sameSlangTermsandCodeWords.pdf.

Table 5.6: Evaluation of the Terms Extracted from the Pseudo-document: Drug

Drug	TF-IDF			PMI		
	term	*weight*	*degree*	*term*	*weight*	*degree*
1	traffickers	0,2306	0	arres	3,2921	-1
2	unprescribed	0,1931	2	microscopic	3,1824	2
3	addicts	0,1709	2	smoked	3,1129	0
4	stouder	0,1546	-1	cracking	3,0791	2
5	forfaitable	0,1488	0	crack	3,0145	1
6	noncontrolled	0,1361	2	laced	2,9915	2
7	cr	0,1193	2	snorting	2,9442	0
8	analog	0,1166	-1	gcms	2,9442	2
9	toxic	0,1112	2	pgp	2,8738	2
10	sponsor	0,1108	0	stalks	2,8674	-1
11	profiteers	0,1076	-1	sativa	2,8558	2
12	cracks	0.1004	2	cake	2.8357	2
13	capability	0.0923	-1	thc	2.8332	2
14	practitioner	0.0895	0	resin	2.8201	2
15	cracking	0.0887	2	raney	2.8068	-1
16	microscopic	0.0863	2	dumontelle	2.7954	-1
17	duquenois	0.0823	2	wigginton	2.7743	-1
18	bruchert	0.0792	-1	hallucinations	2.7536	2
19	arres	0.0788	-1	blues	2.7313	2
20	ingest	0.0760	0	clem	2.7154	-1
21	racketeering	0.0760	0	reacted	2.6919	0
22	mclaren	0.0760	-1	extractions	2.6674	0
23	opelt	0.0760	-1	addictive	2.6530	2
24	arpaio	0.0732	-1	chromatography	2.6223	2
25	lapses	0.0728	-1	capsule	2.6223	2
26	purveyors	0.0728	0	ingested	2.6223	0
27	dumontelle	0.0728	-1	marihuana	2.6146	1
28	blues	0.0697	2	bruchert	2.6108	-1
29	thc	0.0678	2	aide	2.6003	0
30	unverzagt	0.0678	-1	levine	2.5888	2
31	gcms	0.0665	2	smoking	2.5826	0
32	synthesis	0.0602	2	extracted	2.5692	0
33	addictive	0.0602	2	glutethimide	2.5519	2
34	carrico	0.0600	2	madrid	2.5519	-1
35	preludin	0.0597	2	rhonda	2.5519	-1
36	raess	0.0597	-1	cracks	2.5464	2
37	newberry	0.0597	2	baggie	2.5433	0
38	hershey	0,0596	2	cracked	2,5430	2
39	psilocyn	0,0596	2	lsd	2,5415	1
40	propagate	0,0569	-1	marijuana	2,5385	1

to last, *carrico*, the Spanish translation of load. Again with TF-IDF, we notice a name, that a deeper analysis confirmed as relevant. This name is *dunquenois*, which is part of the name *Rapid Modified Duquenois–Levine test* (also known as the simple Rapid Duquenois Test). This test is an established screening test for the presence of marijuana. Finally, synonyms of drug, such as *toxic, addictive* are also extracted by CRIKE together with two conjugations of the verb to crack, precisely *cracks* and *cracking*.

With PMI, consequently, we notice again verb conjugation of the verb to crack.

These are *cracking*, *cracks* and *cracked*. Another conjugation, *laced* is evaluated as new and relevant because it is intended as the act of adding one or more substances to another. This activity is a typical action referred to drug manipulation. *snorting*, on the contrary, refers to the act of taking drug with the nose. Usually, the most popular expression is 'snorting cocaine'. Moreover, also with PMI acronym emerged. These are *gcms* already described, and *pgp* P-glycoprotein (Pgp), a plasma membrane protein which acts as a localized drug transport mechanism, actively exporting drugs out of the cell[3]. Finally, PMI approach extracts again *thc* for Tetrahydrocannabinol and the term *sativa*, which refers clearly to cannabis. The extracted terms *blues* and *cake* are both considered as slang. In particular, *blues* is described above. In addition, *cake* refers to a kilo of cocaine or to a homemade synthetic drug, as reported in unofficial slang dictionaries. However, the same term (*cake*) could be considered as one term of a slang compound term, such as *Layer Cake*, used for Synthetic Cannabinoids, or *Yellow Cake*, used for Methamphetamine. The name *levine* refers again to the same test *Rapid Modified Duquenois–Levine test*, but with PMI the second name has emerged. To last, the involved lawyer considered *addictive*, already extracted with TF-IDF, and *capsule*, both terms that, combined with the ones already formalised in LATO, are synonyms or specification of drug. Examples are *addictive substance* and *metamphetamine capsule*. Biological domain terms have emerged also by using PMI, in particular *chromatography*, an analytical technique utilised widely in the pharmaceutical industry[4] and *glutethimide* a hypnotic and sedative, today superseded by other drugs. That substance is a DEA Schedule II controlled substance that has a high potential for abuse which may lead to severe psychological or physical dependence[5].

Drug Trafficking Verbs. In Table 5.7 we present the top-40 terms extracted by CRIKE for the concept Drug Trafficking Verbs with the associated weights, calculated with TF-IDF and PMI, and the judgement degree expressed by the criminal lawyer.

With TF-IDF we notice that the terms are mostly verbs or verbs nominalisation. Indeed, at first emerged conjugations or nominalisation of verbs, such as *delivered*, *delivery* and *deliver*, and terms identifying exchange of goods, such as *contract* and

[3]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3762612/.

[4]https://www.europeanpharmaceuticalreview.com/article/116263/how-is-liquid-chromatography-used-in-the-pharmaceutical-industry/.

[5]https://pubchem.ncbi.nlm.nih.gov/compound/Glutethimide.

Table 5.7: Evaluation of the Terms Extracted from the Pseudo-document: Drug Trafficking Verbs (D.T.V.)

D.T.V.	TF-IDF			PMI		
	term	*weight*	*degree*	*term*	*weight*	*degree*
1	said	0.3845	-1	vandolah	1.8995	0
2	possession	0.3411	1	deepfreeze	1.8995	2
3	defendant	0.3075	0	clusman	1.8995	-1
4	sale	0.2392	1	seizin	1.8995	0
5	use	0.2087	1	lurie	1.8995	-1
6	court	0.1950	0	harpster	1.8995	-1
7	substance	0.1933	1	sorg	1.8995	-1
8	controlled	0.1749	1	symbolical	1.8995	0
9	property	0.1703	0	greenlaw	1.8995	-1
10	delivered	0.1324	2	hpdi	1.8995	0
11	delivery	0.1322	2	salomon	1.8995	-1
12	control	0.1195	0	pedicle	1.8995	-1
13	plaintiff	0.1191	0	loven	1.8995	-1
14	time	0.1067	-1	klor	1.8995	0
15	state	0.0972	0	cusack	1.8748	0
16	deliver	0.0955	2	mince	1.8450	-1
17	act	0.0929	0	pabst	1.8353	-1
18	shall	0.0920	-1	grandchild	1.8324	-1
19	company	0.0904	-1	cape	1.8291	-1
20	unlawful	0.0753	0	fixation	1.8291	2
21	drug	0.0752	1	vendee	1.8117	2
22	deed	0.0743	2	lyman	1.8064	2
23	intent	0.0714	0	cestui	1.7999	2
24	premises	0.0700	0	donee	1.7967	-1
25	evidence	0.0699	0	geer	1.7964	-1
26	section	0.0679	0	importer	1.7951	2
27	case	0.0655	0	que	1.7926	2
28	illinois	0.0632	0	legacies	1.7905	2
29	person	0.0617	0	henrietta	1.7840	2
30	ill	0.0596	0	leigh	1.7785	-1
31	contract	0.0591	2	josef	1.7739	-1
32	opinion	0.0577	0	doubling	1.7739	2
33	transfer	0.0569	1	tubular	1.7682	2
34	appellant	0.0568	0	smyth	1.7682	-1
35	sales	0.0565	2	kilbourne	1.7682	-1
36	drugs	0.0563	1	grantor	1.7658	2
37	right	0.0556	0	tik	1.7620	0
38	estate	0.0554	-1	reversion	1.7620	0
39	trial	0.0553	0	marr	1.7620	-1
40	business	0.0533	2	bailment	1.7575	2

deed. The terms *sale* and *business* close the list of the new and relevant terms extracted with TF-IDF for the concept Drug Trafficking Verbs.

With PMI we notice first the terms *deepfreeze*, which is considered as a necessary action in drug storage and conservation. The lawyer explained us that, drug deep-freezing could express the purpose of manufacture or delivery that substance, or the intent to manufacture or deliver the same substance itself, or combined with others.

Secondly, we notice the term *doubling* that is considered as the act of multiply, or the action of fold or bend something. *Tubular* is also evaluated as new and relevant because, as an adjective, it refers to the chemical/biological context. In particular, it indicates the act of involving tubules or other tube-shaped structures used in home-made drug preparation. Moreover, PMI extracts terms again related to the economical exchange of goods, or goods transfer, such as *vendee*, *grantor*, *cestui*. The latter is an abbreviation of *Cestui que*, a shortened version of *cestui a que* that literally means the person who benefits the contract. To last, *importer*, *legacies* and *bailment*, are all referring to the action of transferring physical possession of personal property between subjects. Finally, only one slang term is emerged using PMI, that is *henrietta*, a term used apparently to name the substance mostly know as heroin.

Unit of Measure. In Table 5.8 we present the top-40 terms extracted by CRIKE for the concept Unit of Measure with the associated weights, calculated with TF-IDF and PMI, and the judgement degree expressed by the criminal lawyer. This concept is more general, and not referring to drug offences only. Thus, more general terms are emerged, with both approaches. However, even if new units of measure emerged, the terms are mostly related to the unit of measure of the weight.

With TF-IDF we notice paradigms, conjugations and declensions of the verb to weight, such as *weight*, *weighed*, *weighing* and *weigh*. Secondly, are considered new and relevant units of measure, the extracted terms *inches* and *feet* for distance, *years* for time, and *tons* for weight. Lastly, the involved lawyer evaluates positively also the adjectives *tall* and *old*.

With PMI approach we notice first paradigms, conjugations and declensions of the weight, such as *weighs*, *weighed*, *weighing*, *weights* and *weigh*. Similar to these terms, some other terminologies are rewarded positively. These are *overweight* and *weight-master*, both compound words of weight. Differently from TF-IDF, the new units of measure extracted using PMI are *milligram*, with the related plural *milligrams*, and *milliliters*, together with *ton*, and *tons*, and to last, *inches* for measuring distance. PMI extracts also adjectives, such as *tall*, already compared with TF-IDF, now together with *taller*. To last, *loaf* and *loaves* are appeared as slang terms, both used for indicating cocaine and marijuana quantity, in particular one ounce, as pointed out in the aforementioned DEA report, in the list of the slang terms and code words for drug measurement. The terms *tare* and *ranged* belong both to the LATO-KM concept *unit*

Table 5.8: Evaluation of the Terms Extracted from the Pseudo-document: Unit of Measure (U.o.M.)

U.o.M	TF-IDF			PMI		
	term	*weight*	*degree*	*term*	*weight*	*degree*
1	pounds	0.5612	1	liters	5.1708	1
2	weight	0.3398	2	liter	5.1384	1
3	weighed	0.2979	2	moustache	5.1250	-1
4	feet	0.2174	2	tall	5.1141	2
5	defendant	0.2101	0	taller	5.0968	2
6	weighing	0.2032	2	pounds	5.0924	1
7	evidence	0.1841	0	overweight	5.0677	2
8	grams	0.1674	1	axles	5.0526	-1
9	inches	0.1518	2	mustache	5.0314	-1
10	tall	0.1489	2	tandem	5.0037	-1
11	said	0.1146	-1	milligram	4.9864	2
12	court	0.1089	0	weighed	4.9746	2
13	testified	0.0898	-1	weighs	4.9736	2
14	weigh	0.0890	2	weighing	4.9599	2
15	approximately	0.0862	-1	loaves	4.9395	2
16	time	0.0847	0	complected	4.9363	-1
17	years	0.0818	2	tons	4.9341	2
18	plaintiff	0.0804	0	complexion	4.9310	-1
19	0	0.0795	0	axle	4.9057	-1
20	trial	0.0767	0	loaf	4.8780	2
21	cocaine	0.0634	0	haish	4.8576	-1
22	substance	0.0608	0	milligrams	4.8511	2
23	testimony	0.0599	0	weighmaster	4.8489	2
24	tons	0.0551	2	tare	4.8338	2
25	old	0.0533	2	lifting	4.8209	-1
26	shall	0.0528	-1	weights	4.8199	2
27	car	0.0528	-1	whistled	4.8193	-1
28	described	0.0513	-1	weigh	4.8113	2
29	man	0.0493	-1	muscular	4.7988	-1
30	10	0.0490	-1	pound	4.7964	1
31	jury	0.0445	0	goatee	4.7558	-1
32	state	0.0409	0	whistling	4.7420	-1
33	pound	0.0407	1	inches	4.7392	2
34	30	0.0406	0	lift	4.6877	-1
35	case	0.0390	0	afro	4.6733	-1
36	stated	0.0376	-1	ton	4.6694	2
37	witnesses	0.0375	-1	milliliters	4.6347	2
38	coal	0.0374	-1	truss	4.6183	-1
39	manifest	0.0370	0	ranged	4.5858	2
40	police	0.0370	0	twenties	4.5858	-1

of measure by themselves, the first because is an essential component of the action of weight anything, and the second because it is used for indicating a quantifier.

Illinois Legislation. In Table 5.9 we present the top-40 terms extracted by CRIKE for the concept Illinois Legislation with the associated weights, calculated with TF-IDF and PMI, and the judgement degree expressed by the criminal lawyer. This concept is

also not strictly related to drug offences only. Thus, more general terms emerged and are evaluated positively for the enrichment phase by the involved lawyer.

Table 5.9: Evaluation of the Terms Extracted from the Pseudo-document: Illinois Legislation (I.L.)

I.L.	TF-IDF			PMI		
	term	*weight*	*degree*	*term*	*weight*	*degree*
1	act	0.5833	1	schlieper	2.0044	-1
2	illinois	0.3384	1	simulation	2.0044	-1
3	ilcs	0.2456	2	vand	2.0044	-1
4	west	0.2134	-1	nonproperty	2.0044	0
5	section	0.2044	0	cgl	2.0044	0
6	court	0.2006	0	ihda	2.0044	2
7	defendant	0.1787	0	arbitrability	2.0044	-1
8	state	0.1619	2	foia	2.0044	2
9	said	0.1487	0	parentage	1.9888	-1
10	controlled	0.1132	1	ems	1.9679	0
11	substance	0.1093	1	universities	1.9659	-1
12	720	0.1066	1	telecommunications	1.9654	-1
13	policy	0.1059	2	contaminants	1.9586	-1
14	unlawful	0.0890	0	retaliatory	1.9532	-1
15	ill	0.0884	2	architecture	1.9519	-1
16	shall	0.0853	-1	predatory	1.9478	0
17	control	0.0799	2	kills	1.9461	0
18	public	0.0791	0	tortfeasor	1.9393	0
19	possession	0.0712	0	snowmobile	1.9373	-1
20	person	0.0681	0	tortfeasors	1.9317	0
21	law	0.0622	2	prejudgment	1.9304	0
22	plaintiff	0.0619	0	deceptive	1.9292	0
23	trial	0.0530	0	promulgate	1.9242	2
24	acts	0.0529	2	arbitration	1.9218	2
25	drug	0.0517	1	antitrust	1.9170	2
26	rev	0.0497	-1	seq	1.9151	2
27	county	0.0474	-1	arousal	1.9142	0
28	company	0.0466	-1	patterned	1.9113	-1
29	code	0.0460	2	underinsured	1.9101	-1
30	opinion	0.0450	0	riverboat	1.9049	-1
31	case	0.0447	0	repealing	1.9025	0
32	provides	0.0446	0	wage	1.9001	0
33	insurance	0.0440	-1	revived	1.8975	2
34	property	0.0440	-1	unemployment	1.8952	0
35	criminal	0.0434	2	legislation	1.8916	1
36	provisions	0.0433	2	finality	1.8889	-1
37	violation	0.0433	0	philosophy	1.8841	-1
38	use	0.0421	-1	eden	1.8841	-1
39	intent	0.0420	-1	arbitrate	1.8831	2
40	12	0.0420	-1	et	1.8827	2

With TF-IDF we noticed, at first, abbreviations and acronyms, such as *ilcs* for Illinois Controlled Substance and *ill* for Illinoins. Direct synonyms of *legislation* are also evaluated positively by the involved lawyer. These are *policy*, *law*, *control*, *acts*, *code* and *provisions*. The last new and relevant terms are *state* and *criminal*, referred

in our domain to *Illinois* and to *legislation*, respectively.

With PMI we notice first acronyms, such as *ihda* for Illinois Housing Development Authority and *foia* Freedom of Information Act. Other terms refer directly to the law, such as *promulgate*, *revived* and *antitrust*, or refer to litigation alternatives, such as *arbitration* and *arbitrate*, that are dispute resolution alternatives able to create precedents. To last, *seq* and *et* are usually used to express legislative details such as articles or paragraphs.

Criminal Procedure. In Table 5.10 we present the top-40 terms extracted by CRIKE for the concept Criminal Procedure with the associated weights, calculated with TF-IDF and PMI, and the judgement degree expressed by the criminal lawyer.

With TF-IDF we notice at first many terms are referred to the police procedures. These are *police*, *officer*, and *officers* which are referred directly to the police administration. Secondly, the terms *defendant*, *case*, *evidence*, *motion* are referred to the advanced investigation phases, where also the court is usually involved. Moreover, on the same topic, we notice *quash*, that refers to the act of asking the judge for an order setting aside or nullifying an action, and *warrant*, that is an official document, signed by a judge or other authority, that gives the police permission to search someones home, arrest a person, or take some other action. Terms referring to details of police procedures are also emerged using TF-IDF. These are *person*, *circumstances*, *car* and *apartment*. Indeed, in real-life scenarios police investigations reports describe goods carried on the *person*, or hidden in a *car* or in an *apartment*, in particular *circumstances*. Moreover, verbs that are referred to police-investigations appear in the set of the new and relevant terms. These are *search*, *arrest*, *suppress*, *find*, and the related paradigms and conjugations, *testified*, *searched*, and *finding*. Adjectives such as *probable* and *reasonable* belong again to the first phases of the investigations, where the police officers have to assume some probabilistic conclusions before the scientific and official tests are done. To last, the terms *state* and *time* are referring to the phases of preserving the forensic evidence. With PMI the terms considered new and relevant are referring again to standard police procedures and investigations phases, such as *flagrancy*, *warrant*, *fingerprinting*, *attenuation*, *identifications*, *arrestee* and *bust*. But, also *nonconsensual*, *intrusion* and *warrantless* belong to the topic. Also with PMI are emerged verbs used for describing police procedures, such as *quash*, *search*, and *suppress*, with the related paradigms and conjugations, such as *quashing*, *quashed*, *searches*, *suppressing* and

Table 5.10: Evaluation of the Terms Extracted from the Pseudo-document: Criminal Procedure (C.P.)

C.P.	TF-IDF			PMI		
	term	*weight*	*degree*	*term*	*weight*	*degree*
1	defendant	0.4480	2	paulus	2.5187	-1
2	police	0.3830	2	klatt	2.5187	-1
3	search	0.3557	2	halmon	2.5187	-1
4	arrest	0.2863	2	flagrancy	2.5187	2
5	warrant	0.2722	2	shatley	2.5187	-1
6	court	0.2446	0	nonconsensual	2.5187	2
7	found	0.1930	1	caretaking	2.4822	-1
8	evidence	0.1845	2	krull	2.4643	-1
9	officer	0.1312	2	creach	2.4598	0
10	trial	0.1311	0	chimel	2.4586	2
11	apartment	0.1029	2	quashing	2.4539	2
12	officers	0.1009	2	attenuation	2.4516	2
13	cause	0.0982	0	dunaway	2.4516	-1
14	motion	0.0958	2	wead	2.4488	-1
15	suppress	0.0956	2	exigent	2.4364	0
16	probable	0.0843	2	belton	2.4281	-1
17	state	0.0808	2	warrantless	2.4193	2
18	finding	0.0788	2	exigency	2.4055	0
19	arrested	0.0788	1	bust	2.4032	2
20	seized	0.0680	1	coolidge	2.3838	-1
21	time	0.0599	2	quash	2.3769	2
22	seizure	0.0592	1	warrant	2.3758	2
23	person	0.0586	2	suppressing	2.3730	2
24	case	0.0564	2	search	2.3670	2
25	find	0.0564	2	dykema	2.3667	-1
26	said	0.0548	-1	taint	2.3656	0
27	reasonable	0.0478	2	arrest	2.3654	1
28	testified	0.0468	2	suppress	2.3602	2
29	searched	0.0467	2	intrusion	2.3582	2
30	car	0.0447	2	suppressed	2.3567	2
31	2d	0.0421	-1	lawfulness	2.3552	0
32	facts	0.0390	0	parolee	2.3552	0
33	plaintiff	0.0385	0	fingerprinting	2.3488	2
34	hearing	0.0375	0	seizure	2.3449	1
35	station	0.0359	0	arrestee	2.3399	2
36	testimony	0.0358	0	identifications	2.3381	2
37	quash	0.0353	2	consensual	2.3350	0
38	ill	0.0349	0	quashed	2.3270	2
39	circumstances	0.0337	2	searches	2.3190	2
40	jury	0.0334	0	gant	2.3180	-1

suppressed. Finally, one of the best examples of CRIKE quality performance in knowledge extraction from CDs is the name *chimel*. From a deeper analysis of that name is emerged that the name *chimel* is known in the legal context due to the so-called *Chimel Rule*, that is a U.S. legal principle that allows police to perform a warrantless search of an arrested person, and the area within the arrestee's immediate control, in the interest

of officer safety, the prevention of escape, and the preservation of evidence[6].

Evidence. In Table 5.11 we present the top-40 terms extracted by CRIKE for the concept Evidence with the associated weights, calculated with TF-IDF and PMI, and the judgement degree expressed by the criminal lawyer.

Table 5.11: Evaluation of the Terms Extracted from the Pseudo-document: Evidence

Evidence	TF-IDF			PMI		
	term	*weight*	*degree*	*term*	*weight*	*degree*
1	evidence	0.7202	2	gob	2.4499	-1
2	defendant	0.2773	0	pregler	2.4499	-1
3	place	0.2625	2	polyethylene	2.4499	2
4	said	0.1815	0	diatomaceous	2.3759	2
5	court	0.1795	0	wigmore	2.3759	-1
6	plaintiff	0.1557	2	inscribe	2.3759	2
7	placed	0.1290	2	chalk	2.3597	2
8	time	0.1152	0	cinder	2.3344	2
9	jury	0.1084	0	nominating	2.3187	0
10	case	0.0989	2	gehr	2.3187	-1
11	bag	0.0964	1	vacancy	2.3182	-1
12	paper	0.0907	2	genuineness	2.3124	2
13	trial	0.0827	0	orth	2.3055	0
14	state	0.0624	0	nationality	2.3055	-1
15	plastic	0.0611	1	reweigh	2.3055	2
16	appellant	0.0608	0	culvert	2.3018	-1
17	appellee	0.0562	0	competently	2.2997	0
18	question	0.0561	0	bagdonas	2.2979	-1
19	papers	0.0547	2	pralle	2.2979	-1
20	shall	0.0507	-1	fill	2.2931	2
21	found	0.0503	2	opponent	2.2895	0
22	car	0.0490	2	alight	2.2872	-1
23	fact	0.0457	0	overwhelmingly	2.2825	-1
24	record	0.0457	2	filling	2.2716	2
25	testimony	0.0454	2	caboose	2.2694	-1
26	took	0.0420	0	paper	2.2495	2
27	property	0.0400	2	electoral	2.2413	-1
28	company	0.0397	-1	closeness	2.2384	-1
29	error	0.0384	-1	filled	2.2355	2
30	law	0.0375	0	polling	2.2311	-1
31	believe	0.0371	-1	originals	2.2306	-1
32	instruction	0.0366	-1	grefco	2.2275	-1
33	defendants	0.0360	0	perilous	2.2275	-1
34	testified	0.0357	2	surroundings	2.2232	-1
35	find	0.0344	2	regurgitate	2.2155	2
36	person	0.0335	0	fills	2.2118	2
37	facts	0.0332	0	duffle	2.2089	-1
38	bags	0.0329	1	scintilla	2.1999	2
39	places	0.0326	2	preponderates	2.1999	-1
40	cause	0.0325	-1	tfie	2.1962	-1

With TF-IDF the term *evidence* is ranked in the first position. This term was not in-

[6]https://supreme.justia.com/cases/federal/us/395/752/.

cluded in terms-set of the concept during the design of the initial version of LATO. For this reason, it appears a promising result for the correctness of the TF-IDF approach. Verbs referring to the action of finding and testing evidences are also evaluated positively. These are, for instance, *place* and *find* with the related paradigms *placed* and *found*, or conjugation *places*, together with *testified* and *record*. The terms *case*, *paper* and *papers* refer all to the materiality of the evidence, together with *car* and *property*. To last, the terms *testimony* and *plaintiff* refer, on the contrary, to the personality of the evidence.

With PMI the terms appear more heterogeneous. First, we notice terms that are referring to the act of packaging and 'cutting' narcotic substances, such as *polyethylene*, *paper*, *chalk* and *diatomaceous*, also know as siliceous sedimentary rock, that is easily crumbled into a fine white to off-white powder. With respect to this action the term *genuineness* is related to evidence, together with the verbs *reweigh* and *inscribe*, that are referring to the action of verifying and preserve the materiality of the evidence. The other verbs that are rewarded by the lawyer are all related to the act of preparing and hiding drug-doses. These are *fill* with the related paradigms and conjugations *filling*, *filled*, *fills* and *regurgitate* that is intended here as the act of expelling the narcotic hidden in the stomach, as usually done by international drug traffickers. To last, the term *cinder* refers to the combustion residuals, that usually are left after narcotic substance assumptions and are collected as evidence, and the term *scintilla*, which is used as metaphorical expression *scintilla of evidence*. This expression is used for describing a very insignificant or trifling item of evidence[7].

5.4 Goal Achievement

After the evaluation of the first CRIKE cycle, we can address our experimentation goal. The experimental results demonstrate that it is possible to extract, semi-automatically, new and useful terminology from a CDs dataset, without the need of expert-human reading and expensive manual-annotation of CDs. With respect to the three levels of efficiency in terms extraction and enrichment, after this experimentation we can state that:

[7] https://repository.law.umich.edu/cgi/viewcontent.cgi?article=1875context=articles.

- The experimental results reveal that CRIKE appears more correct than useful at this stage, presenting the best total correctness of 0,82 with TF/IDF.
- The experimental results show that each concept has been enriched with useful terminology, presenting the best total usefulness of 0,62 with PMI.
- The experimental results show that CRIKE can enrich LATO initial terms-set consistently, even doubling the initial terminology by an average of 212% with TF-IDF and 211% with PMI.

5.5 Discussion

From the evaluated terminology, we notice that our approach has been successful. In fact, not only paradigms, conjugations, declinations, abbreviation, acronyms and string similarities emerged, but also synonyms and idiomatic language appeared in the list of the extracted terms. Moreover, CRIKE is able to extract important new and relevant legal knowledge from CDs, as proven by the extraction of the so called *Chimel Rule*, with respect to criminal procedure. However, comparing the two approaches, we notice that some terms appear both in the TF-IDF and in the PMI top-40 terms ranking. For this reason, we decide to look at the intersection between the set of terms extracted using TF-IDF and the set of terms extracted with PMI, keeping the record of their evaluation. To this end, we consider the set T_{C_p} as the intersection of $T_{C_i} \setminus T_{C_j}$. In Table 5.12 we present the terms extracted with both approaches.

We notice that the results obtained from the combination of IR techniques are promising and encourage future applications, where the involvement of the expert validation could be reduced. Indeed, from a deeper analysis, we notice the names evaluated with -1 could have a connection with our domain context. For instance, the name *arress* at first has been considered by the lawyer as a misspelling of the verb *to arrest*, due to the fact that, the CAP webpage literally reports that data inevitably includes countless errors as part of the digitization process, because case text and general head matter has been generated by machine OCR and has not received a human review. However, from a deeper analysis, emerged that this name *Arress* occurred in the Naperville (Illinois) Police Department webpage, and it turns out the *Jason Arres* is the Deputy Chief of

Table 5.12: The Intersection Between the Set of Terms Extracted with TF-IDF and the Set of Terms Extracted with PMI.

	Drug		Drug Trafficking Verbs		Unit of Measure		Illinois Legislation		Criminal Procedure		Evidence	
	term	*points*	*term*	*points*	*term*	*points*	*term*	*points*	*term*	*points*	*term*	*points*
1	blues	2			inches	2			search	2	paper	2
2	addictive	2			pounds	2			quash	2	/	/
3	thc	2			weighed	2			warrant	2		
4	arress	-1			pound	1			suppress	2		
5	cracking	2			weighing	2			seizure	2		
6	buchert	-1			tall	2			/	/		
7	dumontelle	-1			tons	2						
8	cracks	2			weigh	2						
9	gcms	2			/	/						

Patrol of Naperville[8]. Moreover, following the same approach, the name *Dumontelle* turns out to occur in the Illinois Public Salaries Database, where *Tyler L Dumontelle* is indicated as an Investigator for Illinois Secretary Of State[9]. On the contrary, from a deeper analysis of the name *Buchert* nothing relevant emerged.

5.5.1 Potential Uses

The CRIKE approach could be useful in real applications and we identify three main possible employments.

1. Legal Text Annotation. CRIKE could serve as a semi-automatic and multi-label annotator, useful for annotating legal documents according to a set of concepts, at different levels of granularity. CRIKE would reduce the human-expert effort required in that task.

2. Legal Information Expansion. CRIKE could be used as a legal-professional assistant tool, to expand the criminal legal knowledge of the subjects involved in the trial, namely lawyers, investigators, law enforcement authorities, public prosecutors,

[8]https://www.naperville.il.us/services/naperville-police-department/office-of-the-chief-of-police/.
[9]https://salary.bettergov.org/person/tyler-l-dumontelle-35066904/.

and judges. These professionals need to expand their knowledge, extracting relevant terminology from CDs, possibly without reading them all entirely.

3. Criminal Evidence Processing. CRIKE could be useful as a support for criminal text analytics, in particular in all the cases where law enforcement agencies need to process a large dataset of investigation reports, phone-tapping and emails, social network interactions and messages, or any sort of textual data forensics. This task is usually performed by searching a set of given keyword, and CRIKE could be useful in discovering new and relevant terminology, initially unknown by the investigators, such as specific cross-domains terminology and peculiar idiomatic language.

5.5.2 Limitations

First, we notice that in CRIKE the human interaction, even if reduced considerably, is yet inevitable. The results are promising and encourage future work in terms of reducing even more the human-intervention, but at this stage is not yet possible. Moreover, at the moment CRIKE has not been developed with an ontology learning scope. Indeed, the terminology extraction performed by CRIKE does not influence the ontology design at this stage, and CRIKE has not yet the goal of discovering new concepts. However, the terms in T^0, the set of pertinent but not relevant terms, could represent the starting point for future work in this direction. Finally, we expect that CRIKE, before reaching its endpoint, will start to collect some noise after several cycles. On one side, this case is mitigated now by the fact that the expert evaluator has the power to stop CRIKE before it reaches its natural endpoint ($CRIKE_{end}$). On the other side, this full human-control over CRIKE would be seen positively in any legal application. Indeed, at the moment, the topic of explainability and governability of CDs processing systems is particularly relevant, especially for the criminal matter.
The analysis of both the extracted names and the other terms made us consider the aspect related to the applicability of CRIKE and its ethical challenges. Indeed, the extraction of legal knowledge from CDs has been possible thanks to the availability of an informative dataset as the one provided from CAP Project. But, at the same time, from the terminology extracted we noticed potential personal data, such as names, and other potentially critical terminology, such as, for instance, *afro*, which has been extracted from the pseudo-document of the concept *Unit of Measure*. These results imposed us

to face, at first, the issues related to real the applicability of knowledge extraction from CDs, and secondly, the inevitable ethical issues connected with analysis of criminal data.

Chapter 6

Applicability and Ethical Issues

6.1 Accessibility of Court Decisions in Europe

Knowledge extraction approaches, such as CRIKE, can be tested and developed on CDs, clearly only if these are available and accessible. But access to CDs is fragmented and variable among European countries. For this reason, we decided to run a survey on the topic of the access to CDs in the European Union. However, this access depends on a number of issues, such as judicial transparency, data governance, right of the defense and access to public data. The lack of a complete European regulation on the topic implies that every EU Member State rules access to CDs independently. In order to draft a European context on the topic, we sent a questionnaire to each institution involved from every single EU country, such as Ministries of Justice, Judiciary Offices, Judge Associations and Supreme Courts. We have done so, in order to inquire how access to CDs is managed. In this section, we present the results of the survey.

6.1.1 Court Decisions in the European Union

Considering some real-world data, Table 6.1 represents the details of the incoming cases in the courts of the most populated countries in the European Union[1]. The data

[1]Constitutional courts, Military courts and Court of Auditors, and the other European or International courts were excluded.

has been extracted from the European Justice Scoreboard 2019, a report realized by the European Commission for the Efficiency of Justice (CEPEJ), that presents an evaluation of the European judicial systems (2016-2018 cycle)[2] with respect to the efficiency and quality of justice in the European Union.

First, we can consider that incoming cases inevitably will have to be resolved through some sort of court's decisional act, i.e. a CD. We noticed that all jurisdictions start from millions of cases in courts of first instance, followed by a dramatic reduction of cases in courts of third instance. However, even if the trend follows the traditional jurisdictional direction, which implies a natural reduction of cases through instances, some clear differences emerged among the countries we analysed. For instance, the United Kingdom[3] reported a number of first instance cases similar to that of other countries, but for the second and third instance, the number of cases is considerably smaller by comparison. More in detail, considering the total amount of cases in the Supreme Court, the UK reported only 60 cases, whereas Italy for instance has 92.177. The causes of this may be twofold. Firstly, the justice system in the UK is structurally different from that of Italy, where instead of a common-law system, based on caselaw precedents, the Roman tradition has established a civil-law system. Secondly, the UK Supreme Court is the final court of appeal for all civil and criminal cases and hears appeals on arguable points of law of general public importance. On the contrary, a country like Italy has many more cases in the third instance, specially 2,4 times of the cases of France, that reports 37.667 cases. With regards to this, there are two opposing points of view. On the one hand, widespread access to the court of third instance can be seen as a synonymous of guaranteed and complete access to justice. On the other hand, large numbers of incoming cases can be seen as a lack of justice efficiency and certainty. The UK is also the only country that reports such a threshold between civil and criminal cases in courts of second instance. Indeed, contrary to the trend followed by the other countries, the UK has 13 times more criminal than civil cases at the court of second instance[4].

[2] Data available at https://www.coe.int/en/web/cepej/2016-2018-evaluation-cycle-reply-bey-country, last visit 8 May 2020.

[3] In UK the survey was completed during the Brexit procedure.

[4] In the CEPEJ details, the UK experts have addressed that the reduction in incoming cases in the courts of second instance is explained by a procedural and terminological difference between the Royal Court of Justice and the Court of Appeal courts. Indeed, not all Royal Courts of Justice courts are second instance courts. Some of them are courts where a case is initially heard, but only the Court of Appeal

Table 6.1: Incoming Cases for each Jurisdiction

TOTAL CASES				
Country	*First Instance*	*Second Instance*	*Third Instance*	TOTAL
France	3.253.649	329.688	37.667	3.621.004
Germany	5.699.447	205.634	18.635	5.923.716
Italy	5.102.805	259.774	92.177	5.454.756
Spain	2.756.317	244.200	23.987	3.024.504
UK	4.176.393	14.768	60	4.191.221
TOTAL	20.988.611	1.054.064	172.526	**22.215.201**
CIVIL CASES				
Country	*First Instance*	*Second Instance*	*Third Instance*	TOTAL
France	2.253.976	282.835	30.018	2.566.829
Germany	4.686.504	142.619	15.591	4.844.714
Italy	3.657.690	135.081	39.793	3.832.564
Spain	1.972.326	184.339	19.956	2.176.621
UK	2.540.573	1.012	52	2.541.637
TOTAL	15.111.069	745.886	105.410	**15.962.365**
CRIMINAL CASES				
Country	*First Instance*	*Second Instance*	*Third Instance*	TOTAL
France	999.673	46.853	7.649	1.054.175
Germany	1.012.943	63.015	3.044	1.079.002
Italy	1.445.115	124.693	52.384	1.622.192
Spain	783.991	59.861	4.031	847.883
UK	1.635.820	13.756	8	1.649.584
TOTAL	5.877.542	308.178	67.116	**6.252.836**

Once we calculated the percentage of the cases in the three jurisdiction on the total amount of cases, the percentage of instances and matters reported in Table 6.2 confirmed the analysis. First instance cases compose around 90% of the total, and among these, civil cases are the large majority. Indeed, i) commercial, familiar, contractual, and other sub-domains of private law claims and legal actions, and also ii) the administrative cases, are all included in the civil cases set. The European Justice Scoreboard 2019 aforementioned names this set as *other than criminal cases*. Table 6.2 confirms a common trend among the most populated European countries, with the only exception once again being the UK. The percentages confirm the dominance of criminal cases in the courts of second instance in the UK (93,14%) that is dramatically higher compared to Italy, the second country for number of criminal cases in the same jurisdiction (48,00%). Table 6.3 shows the rounded fraction between the total amount of cases and the population, in order to count the magnitude of justice on society in 2016. Italy counts one case every 11 people, and France one case every 18 people. Germany appears as the first country for civil cases, with one case every 17 people, but the same country counts only one criminal case every 76 people. Contrary to Germany, Italy reports one criminal case every 37 people. Considering the total amount of cases and the total sum of the population of the countries in question, for 2016 we found a total of one case every 15 people.

Access to CDs has also been discussed within European institutions, as addressed in EU Parliament and Council of European Union [2018]. For instance, the European Council and the representatives of the Governments of the member states meeting within the Council for discussing the best practices regarding the Online Publication of Court Decisions (2018/C 362/02). This work has emphasised that in modern democracies, the Rule of Law requires a transparent judiciary, where citizens have adequate access to the sources of law. The publication of CDs provides insight in how the law is applied by the judge and to this end, continues the Council, modern technologies have revolutionized the way information can be disseminated to the public and retrieved by citizens. However, the online publication of CDs needs to balance a variety of interests. The reported best practices suggest to activate the publication of CDs on the internet, specifying that, in the case where only a subset of CDs is published on the internet, publishing the selection criteria is required for transparency reasons.

should be defined as a second instance court. These differences have influenced the data collection of the incoming cases in the courts of second instance in the UK.

Table 6.2: Percentage of Jurisdiction and Matter

	First Instance (%)			*Second Instance* (%)			*Third Instance* (%)		
Country	**Impact**	Civil	Crime	**Impact**	Civil	Crime	**Impact**	Civil	Crime
France	**89,86**	69,28	30,72	**9,10**	85,79	14,21	**1,04**	79,70	20,30
Germany	**96,21**	82,22	17,78	**3,47**	69,36	30,64	**0,31**	83,67	16,33
Italy	**93,54**	71,68	28,32	**4,76**	51,99	48,01	**1,68**	43,18	56,82
Spain	**91,13**	71,56	28,44	**8,07**	75,49	24,51	**0,79**	83,92	16,08
UK	**99,64**	60,83	39,17	**0,35**	6,86	93,14	**0,001**	86,67	13,33
TOTAL	**94,47**	72,00	28,00	**4,74**	70,77	29,23	**0,77**	61,10	38,90

Table 6.3: Population Index

Country	*Population*	Cases	**Index**	Civil Cases	**Index**	Crime Cases	**Index**
France	66.809.816	3.621.004	**1/18**	2.566.829	**1/26**	1.054.175	**1/63**
Germany	82.5216.53	5.923.716	**1/14**	4.844.714	**1/17**	1.079.002	**1/76**
Italy	60.483.973	5.454.756	**1/11**	3.832.564	**1/16**	1.622.192	**1/37**
Spain	46.658.447	3.024.504	**1/15**	2.176.621	**1/21**	847.883	**1/55**
UK	66.273.576	4.191.221	**1/16**	2.541.637	**1/26**	1.649.584	**1/40**
TOTAL	322.747.465	22.215.201	**1/15**	15.962.365	**1/20**	6.252.836	**1/51**

Furthermore, the Council indicates that regarding the protection of personal data in CDs published on the internet, each member state has to take into account the implications of data protection, considering, for instance, obscuring data while preserving readability and comprehensibility. Moreover, in the European Justice Scoreboard 2019 aforementioned, a comparison focused on litigious civil, commercial and administrative cases is presented, in order to assist member states in their efforts to generate more investment and businesses as well as creating a more citizen-friendly environment. In this report, access to CDs is listed as one of the four categories selected for measuring the quality of justice. As considered by the Commission, online access to CDs increases the transparency of the justice system and helps citizens and businesses in understanding their rights. The EU Commission confirmed that the publication of CDs requires balancing a variety of interests, but at the same time, the Commission supports open data initiatives from the public sector, including the judicial system.

Access to CDs is directly and indirectly influenced by several factors, such as type of jurisdiction, technological resources and legislative environment. Online accessibility to published CDs and the related arrangements for publication of CDs are reported again in the European Justice Scoreboard 2019. However, details and differences regarding publication criteria have yet to be reported. This also goes for issues related to access to CDs with respect to the General Data Protection Regulation (GDPR) and justice outcomes prediction. For these reasons, we wanted to understand how access to CDs is managed among all the European Member States.

6.1.2 Methodology

We have asked about CDs access to every administration involved, such as Ministry of Justice, Supreme Court, Judiciary Office, and Judge Association. To complete this analysis, we sent a detailed questionnaire to the institutional offices involved, in order to receive an official answers regarding four main research topics: i) Data Completeness, ii) Data Access, iii) Data Protection, iv) Data Prediction. The obtained data was first transferred to an answer sheet, and then, aggregated by topic. The results are discussed below with the help of pie charts and maps.

6.1.2.1 Survey Research Topics

Usually, access to CDs is evaluated only considering if the general public has a free and direct access to CDs. But, form an analytical point of view, this information may not be sufficient. For this reason, we considered also the details regarding completeness of the data, actual access modalities and the data protection strategies provided by the administrations. Finally, we also wanted to understand if EU member states have introduced some regulations concerning legal outcome prediction based on past CDs analysis.

Data Completeness (DC) Data completeness is rated by quality level and low-quality of data is a serious problem for any mining approaches, as reported in Liu, Yong-Nan and Li, Jian-Zhong and Zou, Zhao-Nian [2016]. Incomplete data is a subclass of low-quality data and reduces the usability as well as compromising the analysis. For this reason, data completeness is the first aspect we considered for evaluating CDs datasets, given how access to incomplete data may lead to incomplete analysis and partial results. With respect to access to CDs, we first considered the level of Data Completeness of the published CDs in Europe, and then researched what the publication criteria are in the countries where only a subset of CDs is published. By Data Completeness we obviously mean the grade of completeness of the available CDs provided by data producers (e.g. Courts) for data consumers (e.g. professionals, institutions, or individuals).

Data Access (DA) Data access can be considered as the actual possibility of consulting or acquiring detailed information, electronically and for free, without barriers and without being drowned into the maelstrom of locating data in a myriad of complex systems, as drafted in Kierkegaard [2009]. Accessing a large data collection of CDs implies two main aspects: *where* the data producer provides access to CDs and *who* can access them. For this reason, we wanted to investigate if the judiciary allows access CDs on dedicated public databases and if the general public, or conversely only some categories of people, have access to CDs. In other words, by Data Access we mean the access modalities designed and implemented by data producers in order to allow access CDs by data consumers.

Data Protection (DPRO) The European General Data Protection Regulation (GDPR) touches issues related to CDs publication. Data needs to be protected when it is processed, because according to GDPR, almost anything done with data counts as processing, including, for instance, collecting, recording, storing, using, analysing, combining, disclosing or deleting it, as reported in Wilms [2019]. CDs are data that is constantly processed and provides a great deal of information. Data in CDs may report personal data, such as: identification data of the involved people, criminal records, language and nationality, race and ethnic origin, religious or philosophical beliefs, political opinions, trade union memberships, biometric data used to identify individuals, genetic data, health data, and data related to sexual preferences, as well as sexual habits, and/or sexual orientation. GPDR impacts on court and justice administration differently. CDs are data that needs to be handled according to the GPDR indications, even if every single state has the power to develop particular strategies that respond to different needs. For this reason, in this survey we consider the issues related to CDs anonymization and to the role of the Data Protection Officers (DPO) inside court administration in the EU countries. In other words, by Data Protection we mean the strategies designed and implemented by data producers to protect the personal data contained in CDs.

Data Prediction (DPRE) Data can be analysed both with descriptive and predictive purposes. Description provides information about something that presently exists, existed in the past or may exist in the future. Prediction is the task of estimating the value of a target attribute for a given instance based on the values of other attributes for that instance, as reported in Kelleher and Tierney [2018]. Thus, predictive modeling and machine learning based on big datasets are currently being applied to almost any domain of research and justice predictive approaches have also been tested. With respect to CDs, it is now possible to predict court outcomes, based on the previous CDs, as proposed in Katz et al. [2017]. Possible approaches have already been presented and are currently under discussion, as reported in Medvedeva et al. [2019]. But justice prediction approaches rise two fundamental issues: legal authorization and structural feasibility. For this reason, we first investigated whether countries faced these issues with a dedicated regulation and then, we considered all the topics presented above (data completeness, access and protection), in order to establish if justice prediction

approaches can be tested with the current state of affairs. By Data Prediction we mean the permission and the feasibility given to data consumers to test predictive models of court outcomes, based on collections of CDs.

6.1.2.2 Survey Questionnaire

Data Completeness (DC): with respect to this research topic we wanted to understand i) if courts publish all their CDs or only a subset of them and ii) if only a subset of CDs is published, we inquired which are the publication criteria.
Data Access (DA): with respect to this research topic we wanted to understand the facilities and the authorization for accessing CDs, considering which user category has access and where CDs are stored and able to be accessed.
Data Protection (DPRO) with respect to this research topic we wanted to understand the strategies adopted for protecting the personal data reported in CDs. First, we asked if the published CDs are completely anonymized. Secondly, considering all the rights and the administrative issues involved in CDs publication and data protection, we asked whether or not courts have nominated a Data Protection Officer (DPO).
Data Prediction (DPRE) with respect to this research topic we wanted to understand the state of the art for regulations concerning justice outcome prediction approaches.

Table 6.4: The List of the Submitted Questions

Research Topics	Questions
DC 1	*Do Courts publish all their decisions?*
DC 2	*What are the publication criteria?*
DA 1	*Do courts publish CDs on dedicated and public databases?*
DA 2	*What user category can access the published CDs?*
DPRO 1	*Are the published CDs completely anonymized?*
DPRO 2	*Have Courts nominated a Data Protection Officer (DPO)?*
DPRE 1	*Do any regulations or restrictions exist regarding the automatic prediction of case outcomes?*

These questions composed a questionnaire that was designed as closed-ended questions, with only one exception (DC2). More in detail, **DC1** reports the following list of courts, *Court of first instance, Court of second instance* and *Court of third instance* with three possible answers choices for each court: i) *yes*, ii) *no*, iii) *only a subset*. **DC2** is as an open-ended question. **DA1** is composed as a *yes* or *no* question and **DA2** presents a closed set of several options, such as *everybody / only lawyers / only judges or prosecutors / researches / other (please specify)*. **DPRO1**, **DPRO2** and **DPRE1** are all *yes* or *no* questions. To understand how access to CDs is provided in all the 28 European countries we decided to contact the involved administrations directly. We identified the followings administrations as the ones that could answer to our research topics:

Ministry of Justice because represents the governments department for the administration of the entire justice system.

Council of the Judiciary because, as the body that allocates jurisdiction and guarantees the autonomy and independence of magistrates, it intervenes in aspects linked to the organisation and good functioning of the justice-related services.

Supreme Court because, as the highest court within the hierarchy of courts in many jurisdictions, it has an overview on the court of lower instances and may have central administrative and statistical offices.

Judge Association such as the *Deutscher Richterbund* in Germany or the *Union Syndicale des Magistrats* in France, because as representative bodies of the magistrates, they are able to provide the point of view of the category with regard to access to CDs.

All the offices have been contacted through the official website of the institutions. Some of them have only one general contact or a contact form to fill, whereas others present a list of internal sub-offices mail addresses. In these cases, we searched for the dedicated internal sub-offices such as,
i) *general secretary* because, being a general office, it could forwards our questions to the appropriate office, otherwise not reported or not accessible from the official website;

ii) *international offices*, because our questionnaire was part of a non-national request of information;
iii) *statistical offices* because often such offices, reporting statistics, have an overview on collection and accessibility of CDs;
iv) *public relation offices* because as a general office they can forward our questions to the appropriate office, not reported or not accessible from the official website;
v) *data protection officer (DPO) offices* for the institutions listed above, because knowing how data is processed and accessible in the justice system, we assumed that they would be facing the majority of our research topics.
To this end, we sent a few questions to the official contact details reported on the institutional web-page of the 28 European Countries.

6.1.3 Analysis of the Survey Results

We received informative answers from twenty-two countries[5], four countries responded by refusing the questionnaire[6] and two countries did not responded at all[7]. For the only open-ended question **DC2** we have received various answers, that we have aggregated in three categories.

1. `Public interest`, for all the cases where respondents have reported that the courts select the decisions that they account to be of interest for the public or that are important for the interpretation of a law.

2. `Jurisdiction policies`, for all the cases where respondents have reported that i) some CDs are not published due statutory restrictions, ii) only a subset of CDs is published because each court has its internal policies.

3. `Leave to appeal`, that occurred only once, for the case where only a subset of decisions is published because parties have their leave to appeal and the trail is not concluded yet.

[5] Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech, Denmark, Estonia, Finland, Germany, Ireland, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Sweden, United Kingdom

[6] France, Hungary, Italy, Malta

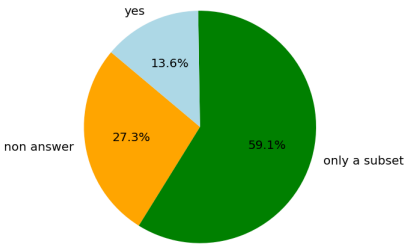
[7] Spain, Greece

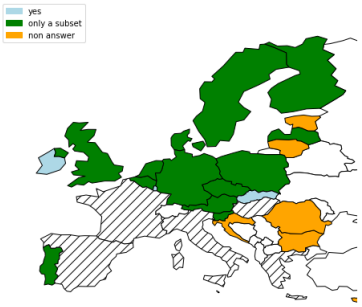
Moreover, even if DPRO1, DPRO2 were composed as closed-ended (yes or no) questions, we received varied answers. In particular, for **DPRO1** some countries reported: `only a subset`, where only a subset of CDs has been anonymized, and `pseudonymized` where CDs have been pseudonymized instead of anonymized. For **DPRO2**, some countries indicated the office where the DPO was nominated, such as: `all federal courts` in Germany and `only the Supreme Court` in Estonia.

6.1.3.1 Data Completeness

We have decided to group the answers received for **DC 1** in one pie chart and one map, in order to summarise the results. The aggregation method has been the following. Based to the answers provided (`yes`, `no`, `only a subset`) we have finally indicated: `yes` to the countries where all CDs are published in each instance, `non answer` to the countries that have not answered the question with respect to at least one court (because without a part of the information it is not possible to assume if all CDs are published or not) and to last `only a subset` to the countries that, in at least one court, have responded that only a subset of CDs is published.

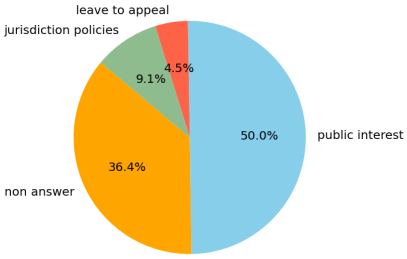
DC 1 Do Courts publish all their decisions?

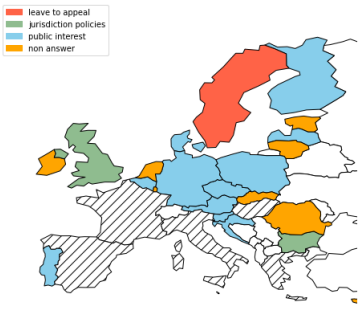




More in detail, CDs from courts of third instance are mostly published (72,7% of countries). Differently, the percentage drops to 63,9% for CDs of courts of second instance, but for CDs of first instance the percentage falls dramatically to 27,3%. From this overview, it emerged that only the 13,6% of the countries publish all their CDs with respect to every instance, whereas 59,1% publish only a subset of CDs. For this reason,

DC 2 What are the publication criteria?





we have decided to investigate what are the publication criteria, i.e. how courts choose the CDs to publish.

We have here reported directly the received answers for **DC2**. In the majority of the cases, the involved administrations have responded that the main criterion for choosing which CDs to publish is the public interest, intended as the importance of CDs with respect to two main aspects: i) the interpretations of law and ii) the legal issues involved. We have reported `public interest` in all the cases where the countries have indicated that the published CDs are selected according to their relevance for the justice system. Only Sweden has reported briefly `leave to appeal`. We understand this as the fact that in this case the trail is not yet concluded. Finally, we have reported `jurisdiction policies` for all the countries that reported internal jurisdiction policies for deciding which CDs to publish. These could be either legislative, such as direct regulations on the topic, or administrative, such as courts discretion. For example, in Bulgaria, the criminal CDs of the Court of Cassation are all published except those containing classified information, and in Germany, each federal state decides by itself for lower instances CDs, but public interest is the main selection criterion. The UK confirmed that each Court/Tribunal has its own policies on CDs publication.

6.1.3.2 Data Access

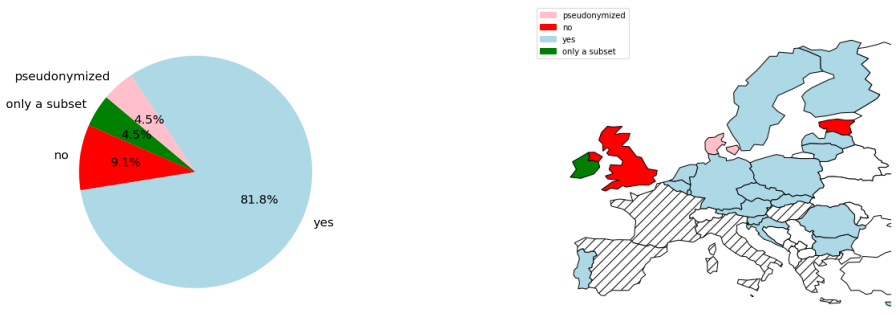
From the obtained answers about **DA 1**, it emerged that all courts have a public dedicated database where CDs are published and stored. Denmark confirmed that the Danish Court Administration was developing at the time a united database for CDs.

Concerning user access to (courts) public CDs databases, the answers about **DA 2** confirmed that the general public has access to the published CDs, even if many private actors, such as legal publishers, provide access to CDs through pricing plans. Many countries confirmed the fact that private publishers have access to court decisions too and are populated as same as public databases.

6.1.3.3 Data Protection

CDs publication guarantees transparency of the judicial action, but at the same time, demands considerations about data governance and privacy in the public administration. The regulation defines that processing data such as CDs *shall be carried out only under the control of official authority or when the processing is authorised by Union or Member State law providing for appropriate safeguards for the rights and freedoms of data subjects. Any comprehensive register of criminal convictions shall be kept only under the control of the official authority*. From the received answer about **D PRO 1**, it emerged that most of the countries anonymized the published CDs (81,8%) confirming the relevance of the involved issues and the practical solution adopted. Only Denmark has reported that personal data in CDs are protected through pseudonymization strategies. The UK and Estonia adopted a full transparency policy, whereas Ireland has decided to anonymize only a subset of CDs. The received answers present a fragmented approach to data protection.

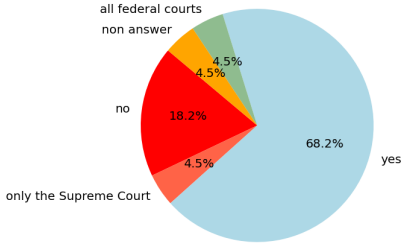
D PRO 1 Are the published CDs completely anonymized?

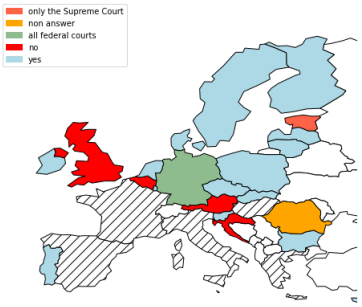


Furthermore, GDPR imposes for every public administration the adoption of a Data Protection Officer (DPO). Considering the European inclusive perspective of the def-

inition of public administration, courts can be considered not only as the expression of the judiciary power, but also as a public administration, in particular as justice administrations, considering all the public services and administrative powers that courts offices have, and that are not strictly jurisdictional activity. For example, courts sometimes express also typical administrative powers, especially when they i) provide access to criminal records or to other documentation, ii) are proposing public competitions, iii) provide public recognition of qualifications, or finally iv) handle the issues related to political elections. Taking into account the large definition of 'public administration' drafted by the European Courts, we have accepted the fact that, for identifying an institution as a public administration, it is necessary to consider the real power expressed by this institution. It is clear that courts, for their typical judicial action, are not included in the recipients of the General Data Protection Regulation (GDPR), but for the additional administrative powers they have, courts could be included with respect to the DPO nomination. Form the received answer we noticed that the large majority of the courts nominated a DPO, with some specification. In Germany, all federal courts have a DPO whereas in Estonia only the Supreme Court.

D PRO 2 Have Courts nominated a Data Protection Officer (DPO)?





6.1.3.4 Data Prediction

The received answers about **D PRE 1** show how the large majority of the countries have not regulated the topic of automatic legal case prediction. For the moment, we can exclude only France from the set, considering the last regulation introduced in the

country regarding prediction systems based on the analysis of the judges behaviours[8]. From the received answers only Romania has not responded to the question. The reported deregulation on the topic would potentially permit the implementation of models able to predict judges behaviours and legal case outcomes. Such approaches, as discussed above, have been already tested. However, the development of justice prediction approaches requires two practical conditions: i) full access to all CDs and ii) a compatible regulatory environment. If both these conditions are realized, it will be possible to move forward on that field of research[9].

6.1.4 Considerations

In this paragraph, we point out first few details about thr countries that refused to answer to our survey.

1. France
The recent French justice reform introduced Article 33 that states literally, *the identity data of magistrates and members of the judiciary cannot be reused with the purpose or effect of evaluating, analysing, comparing or predicting their actual or alleged professional practices*. That means that French CDs will be published anyway, but without reporting the judges names.

2. Greece
On the web page of the Greek Supreme Court it is possible to access a CDs search engine, that permits to choose between number, department, and year of CDs. The database is updated to the CDs of 2018. The web-site permits to choose between penal and civil cases, but it is not mentioned if all CDs are published or only a subset of them.

[8] LOI n 2019-222 du 23 mars 2019 de programmation 2018-2022 et de réforme pour la justice (1) - Article 33

[9] Automatic decisions have been already faced by GDPR, where the right not to be subject to a decision based solely on automated processing, including profiling, is described which produces legal effects concerning him or her or similarly significantly affects him or her. Moreover, the same paragraph recognizes the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision EU Parliament and Council of European Union [2016].

3. Hungary
The Hungarian Judiciary Council has launched the E-File Digital Court Project Court Collection (BHGY) for improving the publication and anonymization of court decisions in a dedicated access system. This system includes the decision of the Supreme Court and lower courts, but it is not mentioned if all CDs are published or only a subset of them.

4. Italy
In 2017 the Italian Judiciary Council approved a project for reconstituting a CDs database of lower courts, dismissed years ago. The project aims at permitting access to CDs of the court of the first instance and court of appeal. However, in October 2020 the Council stated that judges have to select a bunch of CDs that worth to the published in the database, in terms of legal public interest. This approach favours quality over quantity and will not encourage quantitative data analysis.

5. Malta
The Court Services Agency developed eCourts, a system that permits access to Civil Case and Act information. eCourts provides mainly two access modalities, one for citizens and one for legal professionals. Access to civil cases appears for both profiles, but it is not mentioned if all CDs are published or only a subset of them.

6. Spain
In Spain the web-site of the CENTRO DE DOCUMENTACIÓN JUDICIAL (CENDOJ) permits free access to anonymized CDs of both Spanish higher and lowers courts through a dedicated web application. However, it is specified the only a selection of lower courts CDs is published.

Finally, we can claim that access to CDs across the European countries is not completely guaranteed. Only some countries provide full access to all CDs. Considering the results for Data Completeness and Data Access, we can state that every court (of every country that responded) has a public database that is accessible by the general public, but full access to CDs is guaranteed only where courts publish all their decisions (13,6%). Generally, citizens, and in the same way lawyers, prosecutors, judges and administrators, have access to public databases provided by the administrations,

but do not have access to all CDs of every court level. Considering the impact that cases in courts of first and second instances have on the total amount of cases, access to Supreme CDs permits access to less than 2% of all CDs, and even access to the most relevant CDs of second instance courts is not enough to fill the gap. What most clearly emerges form this survey is the lack of a common European administrative strategy on CDs open data. Additionally, simply publishing CDs does not imply the availability of a ready-to-use dataset which can be used for data mining approaches. Usually public resources provided by the administrations, do not allow the direct download of bulk data, metadata or, more in general, a dedicated Application Programming Interface (API). Indeed, other platforms such as the Italian Supreme Court database user interface, does not allow automatic and massive downloads, literally, as a term of use. Moreover, from the received answers we noticed some conflicts on the same questions, where different institutions of the same country reported contradictory responses. For example, on Data Protection, three countries reported conflicting responses on the question regarding the DPO nomination, where one administration confirmed the nomination and the other denied it. In these cases, we considered the nomination as confirmed, assuming the reported negative as a result of misunderstanding or ineffective communications between administrations. It appears likely that one administration ignores the nomination done by another. Regarding access to CDs, it is useful to mention the long answer received from Estonia about legal knowledge diffusion. The Estonian Circuit Court clarified that the Estonian Official Gazette is connected to the court's information system. Data is available and systematical analysis of CDs is provided *in order to make it possible to determine judicial practices, which in the past could be objectified via hypotheses.* According to their report, the mentioned analysis gives litigants greater predictability as regards CDs. Viewed from a more collective perspective, the same analysis gives judges, either individually or as part of the judiciary, an awareness of failings or assumptions that they have not identified in their practice, thereby making progress possible. After all, the main aim of this research is to draft the European state of the art concerning access to CDs, believing that the access to CDs is crucial, not only for the development of quantitative legal research, or for legal data analytics for the justice system, but mostly for the dissemination of the legal knowledge throughout the society and the consequential development of the democracy that can derive from this.

6.2 Ethics in Court Decisions Processing

To highlight and discuss ethical issues in processing CDs, we map the data science ethics framework proposed by Floridi in Floridi and Taddeo [2016], on the CRIKE activity workflow resulting in the three-layer framework shown in Fig.6.1): i) *ethics of data*, involving ethical issues related to collection and analysis of large CDs dataset; ii) *ethics of algorithms*, involving ethical issues related to complexity and autonomy of CRIKE algorithms, and iii) *ethics of practices*, more strictly related to ethics in target oriented classification and prediction activities of CRIKE.

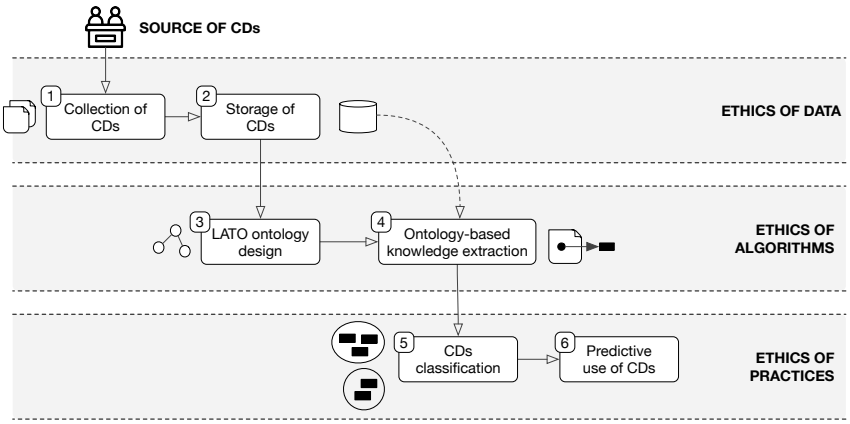


Figure 6.1: Three-layer framework of CRIKE ethical issues

6.2.1 Ethics of Data

Ethics of data primarily refers to the source providing data as well as to the procedures used for data acquisition and storage. In terms of data acquisition, working in the legal domain imposes us to acquire data from a specific, secure and certified source. Both laws and CDs have an institutional creator which should be accessed by directly interacting with the public administration offices in order to acquire genuine data in terms of data format and completeness. But, as reported above in Section 6.1, this result in Europe has not yet been reached. In CRIKE, we process CDs obtained directly from the Caselaw Access Project (CAP) that provides public access to U.S. law (`https://case.law/bulk/download`) digitized from the collection of the Harvard Law Library. The direct access to the these databases guarantees the institu-

tional provenance of data as well as their integrity. A second relevant issue concerning ethics of data involves personal data. In particular, criminal CDs may contain three different categories of personal data, namely (i) identification data, (ii) special categories of personal data and (iii) criminal records. Identification data are defined by the General Data Protection Regulation (GDPR) as those data describing an identifiable person EU Parliament and Council of European Union [2016]. Special categories of personal data are described at paragraph 9 of the GDPR as "those data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, as well as trade union membership, genetic and biometric data, data concerning health or data concerning sex life or sexual orientation". Criminal records are the records concerning a person's criminal history. This last category of personal data is protected at paragraph 10, where GDPR specifies that "access to those data is permitted only under the control of an official authority or when the processing is authorized by European Union or Member State law". The aim of the regulation is to protect personal data against illicit handlings. In particular, main ethical issues related to CDs acquisition and storage concern the risk associated both to the privacy of groups of people and to re-identification of individuals. Specifically, the risk associated with groups regards the possibilities to combine data and groups of individuals, for example, by committed crime, by race or nationality, by spoken language or dialect, by age or gender. These activities could violate groups privacy and could permit re-identification through inference Floridi [2014]. Concerning re-identification of individuals, the main risk is to violate the right of being forgotten, as drafted in Bennett [2012]. These issues are faced in different research fields. For example, Benitez and Malin [2010] presents an estimation of re-identification risk for data sharing policies of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, as well as an evaluation of the risk of a specific re-identification attack using voter registration lists. In general, uncontrolled re-identification risks can conduct to a dangerous information control loss and privacy violation, due to the fact that information privacy concerns specifically the capacity of an individual to maintain control of his or her information Van Wel and Royakkers [2004]. Since privacy regulation is based on the notion of meaningful consent, having trust in data acquisition and processing is a crucial issue Schermer [2011]. In particular, the topic of privacy in accessing individual criminal history information is addressed in Karst [1966], where the authors define policies for providing public access to individual criminal records in Spain and the USA, considering access to court

records, protection of honor, privacy and personal data, free speech and rehabilitation. In this context, CRIKE is compliant with the privacy regulation in that it is conceived to detect exclusively legal concepts inside the CDs and to group the CDs by legal concepts and their application. In other terms, we consider only legal terminology and crime argumentation. Our goals are not related to personal data, directly or indirectly. Knowledge extraction and text mining activities are only devoted to find legal concepts application and how they are expressed by judges inside various CDs. We note that our dataset avoids the group privacy issues in that we obtained a whole set of CDs, rather than only selected CDs targeted to a specific topic/objective to be analysed, like for instance all CDs related to a specific crime or to a specific group of crimes.

6.2.2 Ethics of Algorithms

The ethical issues related to design and implementation of algorithms that elaborate criminal data are transparency, accountability and discrimination. First, in terms of transparency, the risk is to use or implement processes and algorithms that are unclear, incomprehensible and unrepeatable Turilli and Floridi [2009]. Transparency is related to the concepts of accessibility and comprehensibility of information, as reported in Turilli and Floridi [2009]; Rubel and Jones [2016]. Real-world algorithmic decision-making processes designed to maximize fairness and transparency are described in the Open Algorithm (OPAL) project Lepri et al. [2018]. Transparency itself is insufficient, on one side, because companies would not reveal and disseminate proprietary algorithms not to lose their competitive edge, and on the other side, because of the so-called transparency paradox Nissenbaum [2011]. This refers to the fact that, it is clear what machine learning algorithms do in taking decisions about, for example, credit, medical diagnose, personalized recommendations, advertising or job opportunities, but it is still less clear how these decisions are taken Spice [2016]. This issue is directly related to accountability, which is the problem of associating the blame for problems and errors of very complex systems to specific individuals Kraemer et al. [2011]; Matthias [2004].

A further issue to be addressed is how and to whom to enforce accountability for discriminatory outcomes of data analysis. Handling criminal data means in fact to face the risk of associating a criminal behavior with groups of individuals on the basis of their race, religion, cultural background, language, age or gender. An example of

data mining discriminatory outcome in ranking job candidates is described in Barocas and Selbst [2016]. Authors demand caution in the use of data mining techniques and they advocate that this should be part of a comprehensive set of strategies for contrasting discrimination in the workplace and for promoting fair treatment and equality. Other interesting contributions on this issue are the idea of Classification with No Discrimination (CND) Kamiran and Calders [2010] and the proposal of a guideline for researchers and anti-discrimination data analysts on concepts, problems, application areas, datasets, methods, and approaches from a multidisciplinary perspective, as presented in Romei and Ruggieri [2014]. A discussion of algorithm fairness issues on criminal data analysis and racial disparities, in particular focusing on the problem of designing an algorithm for pretrial release decisions, is given in Corbett-Davies et al. [2017]. Since CRIKE knowledge extraction enforces an ontology-based approach with LATO, we comply with the need of transparency in terms of comprehensibility and human intervention. In particular, we decided to base the process of knowledge extraction mainly on quite simple functionalities for searching LATO terminology within the CDs documents in order to guarantee a transparent and easily repeatable process. We handle CRIKE accountability issues by arguing that LATO can be changed and modified directly by the designer, to influence CRIKE results. Our goal is to preserve human intervention and direct control over the system behavior and over the achieved results. Furthermore, in order to avoid the reported discriminatory risks, we base knowledge extraction and classification processes only on general legal concepts and application, by considering for instance crime paragraph, article, verdict and the related terminology.

6.2.3 Ethics of Practices

The issues concerning the ethics of practices are related to the use of the outcomes of data analysis. In particular, we need to face risks concerning anonymity and informed consent, secondary use, and data protection. Informed consent appears insufficient to solve ethical problems related to individuals privacy as discussed in Barocas and Nissenbaum [2014], where authors point out how privacy and big data are simply incompatible without a definition of new approaches having anonymity has one of their primary goals since the design. In particular, they point out how anonymity is different from nameless and reachability. About secondary use, the aim is to ensure ethical

practices fostering both the progress of data science and the protection of the right of individuals and groups, as pointed in Leonelli [2016]. An example of the question of privacy and secondary use of data in health research is given in Lowrance [2003] by considering three different levels: informed consent, anonymity, and public interest mandate. In health research, the reuse of clinical data is a fast-growing field, recognized as essential to: i) realize the potentials for high-quality healthcare, ii) improve healthcare management, iii) reduce healthcare costs, and iv) perform effective clinical research (Meystre et al. [2017]). In particular, one of the main issues in this field is the trade-off between the need of keeping personal data anonymous and the need of exploiting data to achieve results that could be useful for the citizens, according to the notion of public interest. An example is available in Kaplan [2016], where authors describe two court cases (appeared in US and UK) about selling prescription data and the related questions of what constitutes privacy and what public interest. Balancing privacy, public interest and open access raises ethical and juridical questions in the legal field as well, because Criminal Courts declare in their decisions what is forbidden and what is allowed. Thus, according to the European Court of Human Rights, criminal argumentation reported in CDs has to be published, accessible, and known by individuals. The CRIKE system has a scientific research aim only and it respects the GDPR rules for scientific research purposes. We mine CDs in order to extract the legal argumentation and the juridical terms application, by considering the diffusion of the legal knowledge as a positive element. For these reasons, we aim at facilitating the access to legal knowledge without pursuing goals of judge profiling or similar.

6.2.4 Next Challenge: Justice Prediction

Future ethics questions may concern evolutions in argumentation mining and recent approaches to justice prediction. About evolutions in argumentation mining, Moens [2018] addresses that when sufficient explicit discourse markers are present in the language, the argumentation can be interpreted by the machine with an acceptable degree of accuracy. However, in many real settings, the mining task is difficult due to the lack or the ambiguity of the discourse markers, and the fact that a substantial amount of knowledge needed for the correct recognition of the argumentation, its composing elements and their relationships is not explicitly present in the text, but makes up the background knowledge that humans possess when interpreting language. For this reason, in

this work the author surveys on the topic of how the machine can automatically acquire the needed common sense and world knowledge, proposing representation learning and deep learning as possible solutions. About justice prediction, in Medvedeva et al. [2019], the potential of using language analysis and automatic information extraction in order to facilitate statistical research in the legal domain is discussed. More in detail, authors demonstrate the possibilities of NLP techniques for automatically predicting judicial decisions of the European Court of Human Rights (ECHR). The possibilities of analyzing weights assigned to different phrases by the machine learning algorithm, and how these may be used for identifying patterns within the texts of proceedings has been analyzed, thanks to the obtainable information provided from the ECHR. Moreover, recently in Li et al. [2019] experiments on the topic Legal Judgement Prediction (LJP) are discussed. More in particular, in this work authors conduct experiments on four real-world datasets containing large-scale criminal cases in mainland China. Inspired by the impressive success of deep neural networks in a wide range of application scenarios, authors propose a multichannel attentive neural network model (MANN), which learns from previous judgment documents and performs the integrated LJP task in a unified framework, permitting to predict, as outcomes: charge, articles and prison terms. Such an approach has been developed thanks to a large dataset, pre-annotated, and obtainable of real-world criminal CDs, filtered and adjusted to their aims. These preconditions impact in terms of applicability or replicability in other jurisdictions.

Bibliography

Brooke Abrahams, Peter Condliffe, and John Zeleznikow. Using an owl ontology to support legal negotiation about owners corporation disputes. In *Proc. of the 13th Int. Conf. on Artificial Intelligence and Law (ICAIL 2011)*, pages 194–198, Pittsburgh, Pennsylvania, USA, 2011.

Wendy Ann Adams. A transdisciplinary ontology of innovation governance. *Artificial Intelligence and Law*, 16(2):147–174, 2008.

Sagri M.T. (2014) Agnoloni T., Bacci L. Legal keyword extraction and decision categorization: a case study on italian civil case law. In *Proc. of the 5th Workshop on Semantic Processing of Legal Texts (SPLeT 2014)*, pages 1–7, Reykjavik, Iceland, 2014.

Gianmaria Ajani, Guido Boella, Luigi Di Caro, Livio Robaldo, Llio Humphreys, Sabrina Praduroux, Piercarlo Rossi, and Andrea Violato. European legal taxonomy syllabus: A multi-lingual, multi-level ontology framework to untangle the web of european legal terminology. *Applied Ontology*, 11(4):1–51, 2017.

Radboud Winkels Alexander Boer, Rinke Hoekstra. The clime ontology. In *Proc. of the 2nd Workshop on Legal Ontologies at the 14th Int. Conf. on Legal Knowledge and Information System (JURIX 2011)*, pages 37–47, Amsterdam, The Netherlands, 2001.

Denis Araujo, Sandro Rigo, and Jorge Barbosa. Ontology-based information extraction for juridical events with case studies in brazilian legal realm. *Artificial Intelligence and Law*, 25(4):379–396, 2017.

Carmelo Asaro, Maria Biasiotti, Paolo Guidotti, Maurizio Papini, Maria Teresa Sagri, and Daniela Tiscornia. A domain ontology: Italian crime ontology. In *Proc. of*

the Workshop on Legal Ontologies and Web Based Legal Information Management (LegOnt 2003), pages 1–7, Edinburgh, UK, 2003.

Kevin D Ashley. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017.

Kevin D. Ashley and Stefanie Brüninghaus. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17(2):125–165, 2009.

John Bagby and Tracy Mullen. Legal ontology of sales law application to e-commerce. *Artificial Intelligence and Law*, 15(2):155–170, 2007.

Solon Barocas and Helen Nissenbaum. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, chapter Big Data's End Run around Anonymity and Consent, page 44–75. Cambridge University Press, 2014.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

Cesare Bartolini, Robert Muthuri, and Cristiana Santos. *New Frontiers in Artificial Intelligence*, chapter Using Ontologies to Model Data Protection Requirements in Workflows, pages 233–248. Springer International Publishing, 2017.

Christian Baumann and Christian Loës. Formalizing copyright for the internet of services. In *Proc. of the 12th Int. Conf. on Information Integration and Web-Based Applications Services (iiWAS2010)*, pages 714–721, New York, NY, USA, 2010.

Kathleen Benitez and Bradley Malin. Evaluating re-identification risks with respect to the hipaa privacy rule. *Journal of the American Medical Informatics Association : JAMIA*, 17(2):169–77, 2010.

Steven C Bennett. The right to be forgotten: Reconciling eu and us perspectives. *Berkeley Journal of International Law*, 30(1):161–195, 2012.

T. D. Breaux, M. W. Vail, and A. I. Anton. Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. In *Proc. of the 14th Int. Conf. on Requirements Engineering (RE2006)*, pages 49–58, Minneapolis/St.Paul, Minnesota, USA, 2006.

Joost Breuker and Rinke Hoekstra. Core concepts of law: Taking common-sense seriously. In *Proc. of the 3rd Int. Conf. on Formal Ontologies in Information Systems (FOIS 2004)*, pages 210–221, Torino, Italy, 2004.

Joost Breuker, André Valente, and Radboud Winkels. *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, chapter Use and Reuse of Legal Ontologies in Knowledge Engineering and Information Management, pages 36–64. Springer, 2005.

Amedeo Cappelli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Carlo Biagioli. Modelization of domain concepts extracted from the italian privacy legislation. In *Proc. of the 7th Int. Workshop on Computational Semantics (IWCS-7)*, pages 1–4, Tilburg, The Netherlands, 2007.

Benjamin N Cardozo. *The nature of the judicial process*. Yale University Press, 1925.

Núria Casellas. *Legal Ontology Engineering: Methodologies, Modelling Trends, and the Ontology of Professional Judicial Knowledge*, volume 3. Springer Science & Business Media, 2011.

Nuria Casellas, Juan-Emilio Nieto, Albert Meroño-Peñuela, Antoni Roig, Sergi Torralba, Mario Reyes, and Pompeu Casanovas. Ontological semantics for data privacy compliance: The neurona project. Technical report, AAAI Press technical reports serie, 2010.

Silvana Castano, Mattia Falduti, Alfio Ferrara, and Stefano Montanelli. Crime knowledge extraction: an ontology-driven approach for detecting abstract terms in case law decisions. In *Proc. of the 17th Int. Conference on Artificial Intelligence and Law (ICAIL19)*, pages 179–183, Montreal, Canada, 2019.

Marcello Ceci and Aldo Gangemi. An owl ontology library representing judicial interpretations. *Semantic Web Journal*, 7(3):229–253, 2016.

Claudia Cevenini, Giuseppe Contissa, Migle Laukyte, Régis Riveret, and Rossella Rubino. Development of the alis ip ontology: Merging legal and technical perspectives. In *Proc. of the the 2nd Topical Session on Computer-Aided Innovation (CAI2008)*, pages 169–180, Boston, Massachusetts, Usa, 2008.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proc. of the 23rd Int. Conference on Knowledge Discovery and Data Mining (KDD 2017)*, pages 797–806, Halifax, Nova Scotia - Canada, 2017.

Oscar Corcho, Mariano Fernández-López, Asunción Gómez-Pérez, and Angel López-Cima. *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, chapter Building Legal Ontologies with METHONTOLOGY and WebODE, pages 142–157. Springer, Berlin, Heidelberg, 2005.

Emile de Maat, Kai Krabben, and Radboud Winkels. Machine learning versus knowledge based classification of legal texts. In *Proc. of the 23rd Int. Conf. on Legal Knowledge and Information Systems (JURIX2010)*, pages 87–96, Liverpool, UK, 2010.

Cleyton Mário de Oliveira Rodrigues, Frederico Luiz Gonçalves de Freitas, Emanoel Francisco Spósito Barreiros, Ryan Ribeiro de Azevedo, and Adauto Trigueiro de Almeida Filho. Legal ontologies over time: A systematic mapping study. *Expert Systems with Applications*, 130(1):12 – 30, 2019.

Jaime Delgado, Isabel Gallego, Silvia Llorente, and Roberto García. Ipronto: An ontology for digital rights management. In *Proc. of the 16th Int. Conf. on Legal Knowledge and Information Systems (JURIX2003)*, pages 1–10, Utrecht, The Netherlands, 2003.

Judith P. Dick. Representation of legal text for conceptual retrieval. In *Proc. of the 3rd Int. Conf. on Artificial Intelligence and Law (ICAIL1991)*, pages 244–253, Oxford, England, 1991.

Isabella Distinto, Mathieu d'Aquin, and Enrico Motta. Loted2: an ontology of european public procurement notices. *Semantic Web Journal*, 7(3):267–293, 2016.

Timothy AO Endicott. Linguistic indeterminacy. *Oxford Journal of Legal Studies*, 16 (4):667–697, 1996.

EU Parliament and Council of European Union. General Data Protection Regulation, May 2016. URL http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC.

EU Parliament and Council of European Union. Official Journal of the European Union: C 362, 2018.

Biralatei Fawei, Jeff Z. Pan, Martin J. Kollingbaum, and Adam Zachary Wyner. A semi-automated ontology construction for legal question answering. *New Generation Computing*, 37(4):453 – 478, 2019.

Luciano Floridi. Open Data, Data Protection, and Group Privacy. *Philosophy & Technology*, 27(1):1–3, 2014.

Luciano Floridi and Mariarosaria Taddeo. What is data ethics? *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, 374:1–5, 2016.

Enrico Francesconi. An approach to legal rules modelling and automatic learning. In *Proc. of the 22nd Int. Conf. on Legal Knowledge and Information Systems (JURIX2009)*, pages 59–68, Rotterdam, The Netherlands, 2009.

Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. *Integrating a Bottom–Up and Top–Down Methodology for Building Semantic Resources for the Multilingual Legal Domain*, volume 6036, pages 95–121. Springer, 2010.

Enrico Francesconi and Andrea Passerini. Automatic Classification of Provisions in Legislative Texts. *Artificial Intelligence and Law*, 15(1):1–17, 2007.

Keil Geert and Ralf Poscher. *Vagueness and Law: Philosophical and Legal Perspectives.* Oxford University Press, 2016.

Mirna El Ghosh, Hala Naja, Habib Abdulrab, and Mohamad Khalil. Towards a legal rule-based system grounded on the integration of criminal domain ontology and rules. *Procedia Computer Science*, 112:632 – 642, 2017.

Teresa Gonçalves and Paulo Quaresma. Is linguistic information relevant for the text legal classification problem? In *Proc. of the 10th Int. Conf. on Artificial Intelligence and Law (ICAIL2005)*, pages 168–176, Bologna, Italy, 2005.

Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. In *Proc. of the 15th Int. Conf. on Artificial Intelligence and Law (ICAIL2015)*, pages 69–78, San Diego, California, USA, 2015.

Cristine Griffo, João Paulo A Almeida, and Giancarlo Guizzardi. A systematic mapping of the literature on legal core ontologies. In *Proc. of the 7th Brazilian Symposium on Ontology Research (ONTOBRAS 2015)*, pages 79–90, Sao Paolo, Brazil, 2015.

Herbert Lionel Adolphus Hart. *The Concept of Law*. Oxford University Press, 1961.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

A. Isaac and E. Summers. SKOS Simple Knowledge Organization System Primer. Technical report, Working Group Note, W3C, 2009.

Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150(1-2): 239–290, 2003.

Mustafa Jarrar and Robert Meersman. *Advances in Web Semantics I: Ontologies, Web Services and Applied Semantic Web*, chapter Ontology Engineering - The DOGMA Approach, pages 7–34. Springer, Berlin, Heidelberg, 2009.

Mustafa Jarrar, Ruben Verlinden, and Robert Meersman. Ontology-based customer complaint management. In *Proc. of the Workshop on Regulatory Ontologies and the Modelling of Complaint Regulations (WORMCoRe 2003)*, pages 594–606, Catania, Sicily, Italy, 2003.

Jinhyung, Myunggwon Hwang, Hanmin Jung, and WonKyung Sung. ilaw: Semantic web technology based intelligent legislation supporting system. *International Journal of Information Processing and Management*, 3(1):45–49, 2012.

Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Annual Machine Learning Conf. of Belgium and The Netherlands (BeneLearn10)*, pages 1–6, Leuven, Belgium, 2010.

Bonnie Kaplan. How should health data be used? privacy, secondary use, and big data sales. *Cambridge Quarterly of Healthcare Ethics*, 25(2):312–329, 2016.

Kenneth L Karst. "the files": Legal controls over the accuracy and accessibility of stored personal data. *Law and Contemporary Problems*, 31(2):342–376, 1966.

Daniel Martin Katz, Michael J Bommarito II, and Josh Blackman. A general approach for predicting the behavior of the supreme court of the united states. *PLoS One*, 12 (4):1–18, 2017.

John D Kelleher and Brendan Tierney. *Data science*. MIT Press, 2018.

Koen Kerremans, Rita Temmerman, and Jose Tummers. Representing multilingual and culture-specific knowledge in a vat regulatory ontology: Support from the termontography method. In *Proc. of the Workshop on Regulatory Ontologies and the Modelling of Complaint Regulations (WORMCoRe 2003)*, pages 662–674, Catania, Sicily, Italy, 2003.

Sylvia Kierkegaard. Open access to public documents–more secrecy, less transparency! *Computer Law & Security Review*, 25(1):3–27, 2009.

Nadzeya Kiyavitskaya, Nicola Zeni, Travis D. Breaux, Annie I. Antón, James R. Cordy, Luisa Mich, and John Mylopoulos. Automating the extraction of rights and obligations for regulatory compliance. In *Proc. of the 27th Int. Conf. on Conceptual Modeling (ER2008)*, pages 154–168, Barcelona, Spain, 2008.

Felicitas Kraemer, Kees Van Overveld, and Martin Peterson. Is There an Ethics of Algorithms? *Ethics and Information Technology*, 13(3):251–260, 2011.

Ken Kress. Legal indeterminacy. *California Law Review*, 77(2):283–337, 1989.

Ronny Laarschot, Wouter Steenbergen, Heiner Stuckenschmidt, Arno Lodder, and Frank Harmelen. The legal concepts and the layman's terms bridging the gap through ontology-based reasoning about liability. In *Proc. of the 18th Int. Conf. on*

Legal Knowledge and Information Systems (JURIX2005), pages 115–125, Brussels, Belgium, 2005.

Guiraude Lame. *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, chapter Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations, pages 169–184. Springer Berlin Heidelberg, 2005.

Di Caro Luigi Villata Serena Leone, Valentina. Taking stock of legal ontologies: a feature-based comparative analysis. *Artificial Intelligence and Law*, 28(2), 2020.

Sabina Leonelli. Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083): 1–12, 2016.

Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, Transparent, and Accountable Algorithmic Decision-Making Processes. *Philosophy & Technology*, 31(4):611–627, 2018.

S. Li, H. Zhang, L. Ye, X. Guo, and B. Fang. MANN: A multichannel attentive neural network for legal judgment prediction. *IEEE Access*, 7:151144 – 151155, 2019.

Liu, Yong-Nan and Li, Jian-Zhong and Zou, Zhao-Nian. Determining the real data completeness of a relational dataset. *Journal of Computer Science and Technology*, 31(4):720–740, 2016.

Lee Loevinger. Jurimetrics the next step forward. *Jurimetrics Journal*, 12(1):3–41, 1971.

S. Lovrencic and I. J. Tomac. Managing understatements in legislation acts when developing legal ontologies. In *Proc. of the 10th Int. Conf. on Intelligent Engineering Systems (INES2006)*, pages 69–73, London, UK, 2006.

William Lowrance. Learning from experience: Privacy and the secondary use of data in health research. *Journal of health services research & policy*, 8(1):2–7, 2003.

Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6(3):175–183, 2004.

Masha Medvedeva, Michel Vols, and Martijn Wieling. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):1–30, 2019.

SM Meystre, Christian Lovis, T Bürkle, G Tognola, A Budrionis, and CU Lehmann. Clinical data reuse or secondary use: Current status and potential future progress. *Yearbook of medical informatics*, 26(1):38–52, 2017.

Imen Mezghanni and Faiez Gargouri. Crimar: A criminal arabic ontology for a benchmark based evaluation. In *Proc. of the 21st Int. Conf. on Knowledge-Based and Intelligent Information Engineering Systems (KES2017)*, pages 653–662, Marseille, France, 2017.

Hugo A. Mitre, Ana Isabel González-Tablas, Benjamín Ramos, and Arturo Ribagorda. A legal ontology to support privacy preservation in location-based services. In *Proc. of the On the Move to Meaningful Internet Systems 2006 (OTM) Int. Workshop on Web Semantics (SWWS)*, pages 1755–1764, Montpellier, France, 2006.

S. Mittal, K. P. Joshi, C. Pearce, and A. Joshi. Automatic extraction of metrics from slas for cloud service management. In *Proc. of the Int. Conf. on Cloud Engineering (IC2E)*, pages 139–142, Berlin, Germany, 2016.

Raquel Mochales and Marie-Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, 2011.

Marie-Francine Moens. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument Computation*, 9(1):1–14, 2018.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *Proc. of the 11th Int. Conf. on Artificial Intelligence and Law (ICAIL2007)*, pages 225–230, Stanford, California, USA, 2007.

José Félix Muñoz-Soro, Guillermo Esteban, Oscar Corcho, and Francisco Serón. Pproc, an ontology for transparency in public procurement. *Semantic Web*, 7(3): 295–309, 2016.

Helen Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.

Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B Norton, N Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web*, 9(2):185–203, 2018.

Rob Opsomer, Geert De Meyer, Chris Cornelis, and Greet Van Eetvelde. Exploiting properties of legislative texts to improve classification accuracy. In *Proc. of the 22nd Int. Conf. on Legal Knowledge and Information Systems (JURIX2009)*, pages 136–145, Rotterdam, The Netherlands, 2009.

Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. *PrOnto: Privacy Ontology for Legal Reasoning*, chapter Electronic Government and the Information Systems Perspective, pages 139–152. Springer International Publishing, 2018.

Harshvardhan J Pandit, Kaniz Fatema, Declan O'Sullivan, and Dave Lewis. *The Semantic Web*, chapter GDPRtEXT-GDPR as a Linked Data Resource, pages 481–495. Springer International Publishing, 2018.

Cleyton Mário Rodrigues, Ryan Ribeiro de Azevedo, Frederico Luiz Gonçalves de Freitas, Eunice Palmeira da Silva, and Patrícia Vieira da Silva Barros. An ontological approach for simulating legal action in the brazilian penal code. In *Proc. of the 30th Annual ACM Symposium on Applied Computing (SAC15)*, pages 376–381, Salamanca, Spain, 2015.

Víctor Rodríguez-Doncel, Cristiana Santos, and Pompeu Casanovas. Ontology-driven legal support-system in air transport passenger domain. In *Proc. of the Semantic Web for the Law and 2nd Doctoral Consortium Workshops Co-located with 27th Int. Conf on Legal Knowledge and Information Systems (JURIX2014)*, pages 1–10, Krakow, Poland, 2014.

Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

Alan Rubel and Kyle M. L. Jones. Student privacy in learning analytics: an information ethics perspective. *The Information Society*, 32(2):143–159, 2016.

A Salam. Design and implementation of semantic decision support system for supplier performance contract monitoring and execution: Integrating description logics, semantic web rules and service-oriented computing in the context of the extended enterprise. In *Proc. of the Americas Conf. on Information Systems (AMCIS2007)*, pages 293–307, Keystone, Colorado, USA, 2007.

M Saravanan and Balaraman Ravindran. Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment. *Artificial Intelligence and Law*, 18(1):45–76, 2010.

Jaromir Savelka and Kevin D Ashley. Extracting case law sentences for argumentation about the meaning of statutory terms (ArgMining2016). In *Proc. of the 3rd Int. Workshop on Argument Mining*, pages 50–59, Berlin, Germany, 2016.

Jaromír Savelka, Gaurav Trivedi, and Kevin D. Ashley. Applying an interactive machine learning approach to statutory analysis. In *Proc. of the 28th Int. Conf. on Legal Knowledge and Information Systems (JURIX2015)*, 2015.

Jaromir Savelka, Huihui Xu, and Kevin D. Ashley. Improving sentence retrieval from case law for statutory interpretation. In *Proc. of the 17th Int. Conf. on Artificial Intelligence and Law (ICAIL2019)*, pages 113–122, Montreal, Canada, 2019.

Bart Schermer. The limits of privacy in automated profiling and data mining. *Computer Law & Security Review*, 27(1):45–52, 2011.

Gaurav Kant Shankhdhar, V. K. Singh, and M. Darbari. Legal semantic web - a recommendation system. *International Journal of Applied Information Systems*, 7(3): 21–27, 2014.

Chiseung Soh, Seungtak Lim, Kihyun Hong, and Young-Yik Rhim. *AI Approaches to the Complexity of Legal Systems*, chapter Ontology Modeling for Criminal Law, pages 365–379. Springer International Publishing, 2018.

Byron Spice. Carnegie mellon transparency reports make ai decision-making accountable. Technical report, Carnegie Mellon University School of Computer Science, 2016.

Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. JRC eurovoc indexer JEX - a freely available multi-label categorisation tool. In *Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC2012)*, Istanbul, Turkey, 2012.

Piotr Stolarski and Tadeusz Tomaszewski. Modeling and using polish legal knowledge - commercial companies code ontology. In *Proc. of the 11th Int. Conf. on Business Information Systems*, pages 83–94, Innsbruck, Austria, 2008.

Kenji Takano, Makoto Nakamura, Yoshiko Oyama, and Akira Shimazu. Semantic analysis of paragraphs consisting of multiple sentences - towards development of a logical formulation system. In *Proc. of the 23rd Int. Conf. on Legal Knowledge and Information Systems (JURIX2010)*, pages 117–126, Liverpool, UK, 2010.

Tantisripreecha Tanapon and Soonthornphisaj Nuanwan. A study of thai succession law ontology on supreme court sentences retrieval. In *Proc. of the Int. Multi Conf. of Engineers and Computer Scientists (IMECS2010)*, pages 146–151, Hong Kong, 2010.

Paul Thompson. Automatic categorization of case law. In *Proc. of the 8th Int. Conf. on Artificial Intelligence and Law (ICAIL2017)*, pages 70–77, St. Louis, Missouri, USA, 2001.

Matteo Turilli and Luciano Floridi. The Ethics of Information Transparency. *Ethics and Information Technology*, 11(2):105–112, 2009.

Lita Van Wel and Lambèr Royakkers. Ethical Issues in Web Data Mining. *Ethics and Information Technology*, 6(2):129–140, 2004.

Günter Wilms. Good data protection practice in research. Technical report, European University Institute (EUI), 2019.

R. Winkels and R. Hoekestra. Automatic extraction of legal concepts and definitions. In *Proc. of the 25th Int. Conf. on Legal Knowledge and Information Systems (JURIX2012)*, pages 156–165, Amsterdam, Netherlands, 2012.

Adam Wyner and Guido Governatori. A study on translating regulatory rules from natural language to defeasible logic. In *Proc. of the 7th Int. Symposium on Rules (RuleML2013)*, pages 1–8, Seattle, USA, 2013.

Adam Z. Wyner and Wim Peters. On rule extraction from regulations. In *Proc. of the 24th Int. Conf. on Legal Knowledge and Information Systems (JURIX2011)*, pages 113–122, Vienna, Austria, 2011.

Yutaka Yoshida, Kozo Honda, Yuichi Sei, Hiroyuki Nakagawa, Yasuyuki Tahara, and Akihiko Ohsuga. Towards semi-automatic identification of functional requirements in legal texts for public administration. In *Proc. of the 26th Int. Conf. on Legal Knowledge and Information Systems (JURIX2013)*, pages 175–184, Amsterdam, Netherlands, 2013.

Jiansong Zhang and Nora El-Gohary. Automated information transformation for automated regulatory compliance checking in construction. *Journal of Computing in Civil Engineering*, 29(4):907–916, 2015.