Master in
Artificial Intelligence for Public Services
AI4Gov

UNIVERSIDAD
POLITÉCNICA
DE MADRID

POLITECNICO
MILANO 1863
DIPARTIMENTO DI DESIGN

Master’s Degree in Artificial Intelligence for Public Services

A Method for the Semantic Annotation of Legal Documents with Doccano

Supervisor Victor Rodriguez Doncel

Student

Alessio Nardin

Academic Year

2022/2023

Abstract

This project seeks to make a valuable contribution to the field of legal informatics by creating an enhanced and efficient approach for document conversion and annotation within the legal domain. To achieve this goal, the project involves the development of a Python library that empowers seamless data processing. By extending the functionality of the open-source annotation tool, Doccano, the proposed method allows for enriching the label functionality of the tool with existing ontologies and controlled vocabularies, enabling comprehensive end-to-end processing. Through this approach, the project aims to streamline and improve the document annotation process, benefiting legal professionals and researchers alike.

Contents

Introduction

Motivation

The process of annotating legal texts is often a time-consuming endeavor, difficult to scale up due to its dependence on factors like the specific objectives of the annotation, the complexity of the language used in legal text, and the impact that an incorrect interpretation can have, among others. The importance of annotated legal data has become more prominent to improve the performance of both sub-symbolic and symbolic AI algorithms (Francesconi, 2019). For these AI models to be trained effectively, high-quality data is crucial. As a result, good quality semantically annotated legal data has become an objective in its own right.

The literature does point out a gap in methods that are both accurate, economical and efficient in the annotation process, emphasizing the need for streamlined production. Our work aims to increase the tools at disposal to the research and public servant communities by developing a versatile end-to-end data transformation pipeline for legal documents.

Objectives

For the purpose of this project, we lay the ground work for a semantic annotation pipeline that leverages legal ontologies and machine-readable legal resources. Specifically, the work focuses on two strands:

- Developing a data transformation pipeline that integrates the Doccano annotation tool (Nakayama et al., 2018) and outputs the annotated documents according to the Lynx specification (Rodríguez-Doncel et al., 2023)
- Demonstrating the application of the data transformation pipeline by annotating the Italian Copyright legislation with the Copyright ontology (García and Gil, 2006)
- Package and make available the data transformation pipeline for reuse, reproducibility and contributions from the open-source community

In summary, we propose a reproducible and general-purpose end-to-end data transformation pipeline which leverages the Doccano annotation tool for the manual semantic annotation of legal texts and provides hooks to their integration in knowledge graphs through the Lynx ontology.

Methodology

The project involves the development of a Python library for end-to-end data processing. By extending the functionality of the Doccano open-source annotation tool (Nakayama et al., 2018), the proposed method allows for enriching the label functionality of the tool with existing ontologies and controlled vocabularies.

Doccano provides a web interface that allows for the integration of both manual annotations and automatic services via APIs. It supports a variety of annotation techniques, for both images and text, but in the context of this work, we extended the named entity recognition functionality by integrating semantic annotation based on existing legal ontologies or controlled vocabularies. By improving the support for Akoma Ntoso (Palmirani and Vitali, 2011) documents with Doccano, we can enable the customization of the annotation process as per user requirements.

The annotated documents can then be exported from Doccano in an offset-based format, where ontological classes are represented as *entities* and properties are represented as *relations*. Additional processing then takes place to transform the annotated documents into a knowledge graph based on the Lynx specification (Rodríguez-Doncel et al., 2023), that enables the use of such data for powering ML models.

About this document

This thesis is structured as follows:

- Chapter 1 presents the theoretical background and previous work on the semantic annotation of legal texts
- Chapter 2 provides an overview of the data and tools used during the successful iteration of this project, as well some background information on the legal domain we have decided to tackle (Italian Copyright Legislation).
- Finally, in Chapter 3 we describe more in detail the data transformation pipeline, reflect on its potentials and limitations, compare it with other existing tools and propose future research and development work.

Chapter 1

Semantic annotation

The maturation of digital technologies has opened up countless new opportunities for legal activities. Across the years, they have enabled for improved methodologies for handling, exploring, analyzing and implementing legal resources, greatly impacting the legal sector at different levels (Katz et al., 2021; Sharma et al., 2021). In this context, the field of legal informatics has thrived, with more and more scholars focusing on the different aspects of innovative legal technologies, from the development of legal ontologies and semantic annotation of legal texts, to automated regulatory compliance checks and enhanced legal reasoning (Katz et al., 2021; Loutsaris and Charalabidis, 2020; Rodriguez-Doncel, 2023).

This was largely enabled by the development of new ways of storing legal information, which allowed for an easier retrieval and reuse of the legal concepts present in the natural language texts such as the competent jurisdiction, temporal elements, and deontic rules (Sharma et al., 2021; Loutsaris and Charalabidis, 2020).

1.1 What is semantic annotation?

Among the array of techniques that have emerged in recent years, a prominent approach involves semantically annotating legal texts (Soavi et al., 2022b; Nazarenko et al., 2021; Tang et al., 2020; Ghijsen et al., 2013). This technique entails *"connecting these texts with established ontologies [by] to identify[ing] concepts from that ontology that are relevant to the document or that are referred to by it, as well as identifying specific passages in the document where the concepts in question are mentioned" (Brank et al., 2018).*. In other words, semantic annotation can be defined as the process of mapping texts to ontologies, enabling further processing by machines based on meaning instead of syntax (Kiryakov et al., 2003).

According to Adebayo et al. (2018) semantic annotation can be formalized *"as a 4-tuple (a, b, c, d), where a is the subject of the annotation, b is the object of the annotation, c is the predicate which defines the type of relationship between a and b, while d signifies the context in which the annotation is made."*

1.1.1 Benefits of semantic annotation

Semantically annotating legal texts has several advantages. In particular, it can power the retrieval of legal documents using detailed natural language queries. Traditional annotation-based retrieval systems rely on word matching and they often fall short in supporting complex legal query types. With a semantic understanding of query sentences and legal annotations, retrieval can become more precise and efficient (Soavi et al., 2022a).

Additionally, semantic annotations minimizes the ambiguity of the meaning of legal terms by making them explicitly defined (Soavi et al., 2022a). However, this is often complex to achieve, as in many cases legislators draft legislation which is "ambiguous-by-design" in order to strike a deal.

Moreover, semantic annotations could be reused in data-intensive applications downstream. An example of such applications is *Kantoorbelasting*, a proof-of-concept of the Flemish Government that leverages semantically annotated municipal regulations to enable citizens to estimate the taxes on commercial buildings depending on the city they plan to set their business in (Digitaal Vlaanderen, 2023). Automated compliance checks are also another well-explored use-case (Kiyavitskaya et al., 2006). Among many, a notable example is the Lynx project, focused on enhancing access to digital regulatory compliance documents that target SMEs. The pilot developed regulatory compliance assurance services in three areas: contract compliance, labour law and fossil fuels (Rodriguez-Doncel, 2023). These applications are built on top of the semantic information contained in the legal corpora, thus increasing the integration among legal resources and the business logic and data access layers of the IT systems involved (Digitaal Vlaanderen, 2023; Rodriguez-Doncel, 2023).

1.1.2 Challenges of semantic annotation

However, annotating legal documents is a complex task due to the discursive nature of legal information. This is often referred to as the "natural language barrier," which involves translating natural language sentences into semantic interpretations (Ceci et al., 2012; Soavi et al., 2022a). Transforming these texts into solid formal specifications often requires a multi-step process with multiple iterations that ideally would need to be validated by the legislators (Soavi et al., 2022a).

Additionally, as mentioned above, legal resources often purposefully leave room for diverse and/or conflicting interpretations, so that the issue is dealt in secondary legislation such as implementing acts or by the competent judicial court.

Finally, achieving high-level semantic meanings in the legal domain often requires integrating extracted relations with external legal knowledge sources, which can be vast and varied.

1.2 Considerations for the semantic annotation of legal texts

As described above, annotating a legal text is an open field of research. However, with the evolution of legal informatics, some well-known concepts and considerations have been pinpointed by the scientific community as crucial for achieving solid results.

1.2.1 Legal isomorphism

The unique characteristics of the legal domain demand accountability, traceability, and explainability of legally binding decisions and provisions. Consequently, it is essential that a machine-readable representation of a legal text can be traced back to its original form in natural language. This element is crucial because, according to prevailing legal provisions in most legal systems, the natural language version stands as the sole authoritative source of truth recognized by the courts. As a consequence, much effort has been devoted to finding ways of encoding legal information in a way that is both human readable and machine readable (Santos et al., 2018; Rodriguez-Doncel et al., 2018).

This correspondence is referred to with the term *isomorphism*, which can be defined as the precise correspondence between the natural language of a text and its representation (Bench-Capon and Coenen, 1992). Karpf and Københav (1989) identifies five conditions for achieving this correspondence:

1. The individual representation of each legal source.
2. The preservation of each source's structure.
3. The maintenance of mutual relations, references, and connections between sources.
4. A distinct representation of legal sources from other system components.
5. The inclusion of both material and procedural rules (if the model covers procedural law).

In other words, isomorphism demands a clear link between the source material items and items in the knowledge base. Originally, the key benefits of isomorphism were believed to be in knowledge base verification, validation, and maintenance. However, further application revealed broader advantages, impacting the system's entire lifecycle (Bench-Capon and Coenen, 1992).

Legal knowledge bases, due to the dynamic nature of laws and regulations, necessitate frequent updates. An isomorphic approach allows for pinpointed updates, meaning that minimal changes in the source material lead to localized adjustments in the knowledge base (Bench-Capon and Coenen, 1992).

Lastly, from a user's perspective, isomorphism brings forth several advantages, such as simplified system learning and clarity in understanding the system's rationale. If users base their understanding on specific source documents, an isomorphic system aligns with this perspective, making interactions more intuitive (Bench-Capon and Coenen, 1992).

1.2.2 Provision-centric approach

A notable perspective emerging from multiple studies is viewing legislation through a "provision-centric" lens. Legislation, in essence, transports rules or provisions, carried by linguistic acts. By focusing on provisions, researchers aim to offer a more systematic view of legal systems. Provisions, in their various forms depending on legislative intent, can be described using metadata schemes (Biagioli et al., 2005; Francesconi and Passerini, 2007).

In contractual and legal contexts, a provision is often understood as a stipulation, term, or condition that delineates specific aspects of an agreement. These provisions, as highlighted by Francesconi and Passerini (2007), serve as the cornerstone of many agreements, outlining the obligations, rights, and responsibilities of the agents involved. By focusing on provisions as the central component rather than mere isolated terms, this approach offers a holistic perspective on the entire contractual structure (Francesconi and Passerini, 2007). Provisions, when semantically annotated, transcend from being static textual entities to dynamic, interconnected, and easily navigable data points. This amalgamation not only streamlines the processing and management of legal documents but also ensures greater clarity and reduced ambiguities (Biagioli et al., 2005).

1.2.3 Deontic logic

Deontic logic can be defined as *"the logic of normative expressions: expressions pertaining to the obligations, permissions and rights of agents"* (Jones and Sergot, 1992; von Wright, 1951). As such, it is highly relevant for the formal representation of legal knowledge, as it is one of the key elements that could be leveraged for modeling legal rules. Its significance becomes even more pronounced when we delve into the semantic annotation of legal texts, a process that attaches deeper, structured meanings to textual content.

This is not to say that legal provisions cannot be modeled without taking into account deontic logic. Jones and Sergot (1992) reflects on the fact that while many legal fragments can be modeled without such operators, certain situations, especially those involving potential violations and consequential states, demand its inclusion for achieving an accurate and reusable representation.

The semantic richness of legal texts often goes beyond mere linguistic expressions. The real challenge lies in discerning whether a provision can be violated and if such a violation leads to a state regulated by another legal provision (Jones and Sergot, 1992). This discernment is crucial for semantic annotation, ensuring that the underlying obligations, permissions, or prohibitions are captured accurately.

1.2.4 Interpretation

Semantically annotating legal texts means diving deep into the meanings and contexts of legal language. Instead of just labeling words, sentences or paragraphs, this process aims to understand and map out the deeper ideas, relationships, and intentions in legal documents. It's about making sense of complex legal terms and determining their relationships within the broader legal framework (Ceci et al., 2012; Soavi et al., 2022b).

Legal texts, more often than not, are characterized by having an open texture (Waismann, 1947), that is the natural uncertainty or lack of clarity in various legal rules and guidelines. This ambiguity emerges because the nuances of language can't always accurately reflect all facets of human experiences. When framing laws, it's impossible for lawmakers to predict every circumstance or situation where a law might come into play. Consequently, these legal documents frequently include terms that invite interpretation. Such undefined spaces or "grey zones" in the text offer adaptability, ensuring that the law remains relevant in evolving societal situations and unexpected events. As a consequence, while the goal is to remain as objective as possible, this inherent nature of legal language means that some degree of interpretation is inevitable. Annotators must discern the latent intentions behind provisions, predict potential areas of contention, and preemptively address ambiguities (Stegmeier et al., 2021; Ma and Wilson, 2021).

However, the ultimate authority on interpretation traditionally rests with the courts. Annotators, regardless of their expertise, are navigating a terrain where they must tread carefully. Every legal provision or statute is a product of extensive deliberations, negotiations, and compromises. Thus, the language used often carries with it historical, political, and socio-economic nuances that might not be immediately evident (Stegmeier et al., 2021; Athan et al., 2014).

Moreover, court decisions, which serve as primary interpretive guides, evolve over time. Precedents may be overturned, and interpretations might shift based on societal changes, evolving legal philosophies, or the particular composition of a judicial bench. Annotators, therefore, face the challenge of ensuring their annotations remain relevant and accurate in light of an ever-evolving body of case law. They might be confronted with situations where different courts provide varied interpretations of the same provision. Deciding which interpretation to align with, or how to present multiple interpretations without bias, can be daunting (Stegmeier et al., 2021; Athan et al., 2014).

Additionally, the risk of personal bias is ever-present. Annotators, being human, come with their own set of beliefs, values, and experiences that can inadvertently influence their interpretation. While the goal is always to remain neutral, the subjective nature of some legal provisions can make this challenging. For instance, terms like *"reasonable"* or *"undue hardship"* in legal statutes are inherently open to interpretation and don't have fixed, universally accepted definitions (Ma and Wilson, 2021; Athan et al., 2014). This requires a traceability of the agent that performed the annotation in the annotation itself, in order to be able to ask for clarifications or interpretative keys, should the need arise.

Lastly, the stakes are high. Misinterpretation or oversimplification during annotation can lead to the misrepresentation of legal provisions, that can have tangible consequences. This is especially true for systems that rely on these annotations for the automation of decision-making processes.

1.3 An overview of semantic annotations techniques

So far, we have explored some considerations that needs to be taken into account when establishing an annotation strategy. However, we have not dwelt yet into how the annotation itself can take place and the role of automated systems in supporting or carrying out the annotation effort. Broadly speaking, the techniques that have been employed for this task can be subdivided in three main categories: manual annotation, assisted or semi-automated annotation, and automated annotation.

1.3.1 Manual Annotation

Legal texts, as repositories of historical, cultural, and social contexts, require an interpretative finesse that automated approaches often lack. Because of this, semantic annotation requires the support of legal scholars and domain experts who engage in a meticulous examination of the texts. Through this process, they assign relevant semantic tags or metadata based on their expertise, interpretation and understanding of the legal text (Nazarenko et al., 2018; Governatori, 2005).

The main advantage of manual annotation lies in its precision. The complexity of the legal language often elude computational algorithms. As a result, manual annotation ensures that annotations encompass both the overt and covert meanings enshrined within legal provisions. Furthermore, as mentioned above, certain legal provisions or judgments can be open to multiple, potentially conflicting interpretations (Stegmeier et al., 2021).

While the meticulous nature of manual annotation underscores its strength, it is simultaneously a source of limitation. The time and effort it demands make it less viable for managing extensive databases of legal documents, a concern amplified in an era characterized by exponential digital expansion (Nazarenko et al., 2018).

Nazarenko et al. (2021) worked on a methodology to semantically annotate legal documents. They introduce a coarse-grained, interpretation-neutral annotation layer to enrich legal texts, striking a balance between traditional statistical retrieval and full textual rule formalization. Their methodology is exemplified through the semantic annotation of the French version of the GDPR, utilizing the Core Legal Annotation Language (CLAL) formalized in XML, establishing a gold standard for future semantic annotations in the legal domain.

Another issue with manual annotation is that, although it enriches the depth of analysis, it introduces an element of subjectivity. Given the interpretational diversity among annotators, disparities in annotations can arise, particularly when dealing with multiple experts (Stegmeier et al., 2021).

In light of its reliance on human expertise, manual annotation necessitates substantial resources, both in terms of time and specialized knowledge. This, in turn, can render it less tenable for projects constrained by limited resources or stringent timelines.

1.3.2 Semi-automated or assisted annotation

Semi-automated annotation seeks to strike a balance between human expertise and the speed of automated tools. Generally, by using this approach, computational tools propose initial annotations, which are subsequently reviewed and refined by human experts. By amalgamating human judgment with computational speed, semiautomatic annotation offers an efficient means of annotating texts, ensuring a harmonious blend of accuracy and efficiency.

Soavi et al. (2022b) developed a tool named *"ContrattoA"* that semi-automatically conducts semantic annotation of legal contract text leveraging a domain ontology. To achieve this, they first leveraged lexical patterns to recognize some ontological concepts in the legal text. Then, the effectiveness of these patterns was evaluated in an empirical study where one group of subjects was asked to annotate legal text manually, while a second group edited the annotations generated by ContrattoA. Subsequently, they focused on the core contract concepts of obligation and power where the results from the first iteration were mixed. Using an extended set of sample contracts, new lexical patterns were derived. These patterns were shown to significantly improve the performance of *ContrattoA*.

Ceci et al. (2012), worked on designing a system to assist human experts in the annotation of normative modifications. As in the previous example, the automated system proposes a label, which is then validated or modified by the human annotator.

1.3.3 Automated annotation

The domain of legal texts presents unique challenges for automated semantic annotation due to its intricate language, layered references, and the high stakes involved in accurate interpretation. The most salient advantage of automatic annotation is its speed. Algorithms can process vast volumes of data at rates unattainable by humans. Yet, the very strength of automatic annotation is also its weakness. The absence of human oversight can lead to misinterpretations, especially given the nuanced and context-sensitive nature of legal language. Algorithms, while efficient, often lack the depth to fully grasp the multifaceted interpretations inherent in legal provisions at the current state of technology. Over the years, several researchers have delved into the task of automating various aspects of legal text processing.

Palmirani et al. (2003) worked on the automatic extraction of normative references in legal texts from the Italian acquis. They claimed that standardized legal documents, marked up under uniform formats and structures, can facilitate a seamless integration between distinct legal texts, making them easier to reference, find, and process. A significant contribution of their work lies in the identification and marking of various components of legal texts, such as partitions (paragraphs, articles and sections), and the normative references they contain. Their project sought to craft a model adept at recognizing, understanding, and normalizing these normative references, a step forward in ensuring semantic interoperability among various legal information systems.

Recognizing the magnitude and intricacy of legal documents, Biagioli et al. (2005) emphasized the need for automated tools that can support the annotation process. They worked on the identification of regulatory provisions by detecting actors, rights, obligations and other entities integral to the legal discourse. They structured their approach in two steps: a provision identification module, that detects and classifies fragments of normative texts automatically based on their provision type, and an argument extraction module, that extracts the associated arguments from these provisions. They encountered several challenges tied to the unique lexicon and structure of the legal language, as they often carry weighty consequences.

Zeni et al. (2008, 2013) worked on automating the annotation of the Stanca law on web accessibility by adapting the Cerno framework in the Italian language. This framework provides a structured method, encompassing the identification of specific text fragments in regulatory documents, constructing a cohesive semantic model from these annotations, and subsequently transforming this model into distinct functional and non-functional requirements by focusing on concepts of obligations, rights, anti-rights, actor, and others.

Asooja et al. (2015) focused on creating an automated approach for annotating financial regulations. Their aim was to enrich a formal ontology using segments from legislative texts. The method employed not only domain-centric semantics but also general ones, and integrated deontic logic for prescriptive clauses. Their technique utilized a multi-label classification strategy, powered by several binary classifiers. This system was trained using provision types manually annotated by subject matter experts, adopting a supervised learning approach.

Adebayo et al. (2018) leveraged the TextTiling algorithm, enhancing it with Latent Dirichlet Allocation (LDA) for refined topic modeling. By integrating Semantic Textual Similarity principles, legal texts are segmented to mirror inherent subtopics. Their aim was to semantically connect concept descriptors with topical segments across a document. Their methodology unfolds in three stages: topical document segmentation, concept profiling, and mapping of concepts to segments.

Humphreys et al. (2020) harness advanced techniques to automate the semantic annotation of legal documents, specifically focusing on extracting norms and their elements. Central to their approach is combination of paraphrasing tools within unsupervised information extraction systems to produce semantically cohesive components from natural language. They applied Semantic Role Labeling (SRL) to identify and annotate specific elements such as "Situation", "Result", and "Condition" in legal texts. These annotated roles are then systematically mapped to slots in legal ontologies.

In Sleimi et al. (2020), an automated system that utilizes NLP and ML to extract designated metadata from texts was developed. They tested the system’s effectiveness through two case studies centered on Luxembourgish legislation. The outcomes were encouraging: precision rates reached 97.2% and 82.4%, with recall at 94.9% and 92.4%. This evaluation was anchored on 200 randomly selected statements from traffic laws. Their approach could be divided into three components: semantic metadata extraction (at both phrase and statement levels) and a named entity recognition algorithm used to identify the agents involved.

(Commission et al., 2021)

In the study by Savelka (2023), the use of generative pre-trained transformers (GPT) was explored for the semantic annotation of short text snippets from various legal documents. While there have been discussions regarding the potential applications of this technology in the legal domain, a thorough analysis was missing, especially concerning sentence-level semantic annotation in zero-shot learning settings. Savelka (2023)'s research addressed this gap, evaluating the LMM's ability to annotate small batches of short text snippets based solely on concise definitions of the semantic types. The results showed that the GPT model performed in zero-shot settings across different document types, achieving F_1 scores of 73% on court opinions, 86% for contracts, and 54% for statutes and regulations.

1.4 Conclusion

The complex nature of legal resources, rich in subtle linguistic patterns and vague language, compels a well-though approach to semantic annotation. While manual annotation offers a depth unattainable by machines, it often falls short in scalability. On the other hand, while automatic annotation offers unparalleled speed, it sometimes lacks the finesse required for precise interpretation. Semi-automatic annotation, on the other hand, offers a promising middle ground.

As the domain of legal informatics evolves, a holistic understanding of these techniques is imperative. Future endeavors in this realm will undoubtedly seek to further harmonize the strengths of each method, paving the way for more refined, efficient, and accurate semantic annotations.

In conclusion, the semantic annotation of legal texts remains an evolving field, characterized by diverse methodologies and techniques. These methodologies, while differing in their approaches, collectively underscore the significance of structured, systematic, and semantically rich representation of legal texts for enhanced accessibility, interpretation, and applicability.

Chapter 2

Data and tools

In the previous chapter, we discussed research on the semantic annotation of legal resources, along with crucial factors to consider in this endeavor. Building on that theoretical foundation, we will now delve into the core components of this project: the data available to us and a preliminary analysis of its structure. In particular, we choose to provide a demonstration of the data transformation pipeline by annotating the Italian copyright legislation (Gazzetta Ufficiale, 1941) with the copyright ontology (García and Gil, 2006).

The decision to annotate Italian copyright legislation was influenced by the significant harmonization already present within EU legislation. This uniformity offers an interesting case study that could be expanded across different Member States. While our focus is on this particular legislation, the transformation pipeline has a versatile nature, making it adaptable to diverse domains.

2.1 The Italian Copyright legislation

Copyright is a legal instrument designed to safeguard *"any literary, dramatic, musical or artistic original work, provided that is recorded, either in writing or otherwise"* from imitation (Spence, 2007). Unlike patents, which cover broader concepts, copyright specifically defends against direct copying and doesn't account for independent creation. It provides authors with exclusive rights concerning the reproduction, performance, adaptation, and translation of their work. The criteria for determining if an intangible item qualifies for copyright are broad. A "work" encompasses any set of materials organized intentionally. The term 'literary' doesn't assess the content's quality: originality doesn't depend on innovation but merely requires some form of intellectual endeavor (Spence, 2007). Moreover, copyright law also recognize a spectrum of related rights, extending to creative outputs not directly tied to the original author. This encompasses rights of artists, audio recording producers, broadcasting entities, and, within the European framework, rights of movie producers, database designers, semiconductor creators, and industrial design professionals. Given these overarching criteria, the realm of copyright protection is vast. It spans sectors like visual arts, publishing, performing arts, and areas like source code and databases.

2.1.1 Historical overview

Italy's copyright law is rooted in Law no. 633 of 22 April 1941 (Gazzetta Ufficiale, 1941), which is further complemented by select provisions from the Italian Civil Code of 1942. The digital revolution and the subsequent challenges it posed to traditional copyright frameworks necessitated adaptations in legal structures across Europe.

Initially, this law adhered to the minimum protection requirements outlined in the Berne Convention for the Protection of Literary and Artistic Works of 1886 (WIPO, 1886). Over the years, the law has undergone several revisions to comply with various directives from the European Union and to align with changes following the establishment of the Italian Republic. For example, the legal protection of software was addressed through the issuance of Legislative Decree No. 518 of December 29, 1992 (Gazzetta Ufficiale, 1992), later amended by Legislative Decree No. 205 of March 15, 1996 (Gazzetta Ufficiale, 1996), in response to the EU Directive 91/250/EEC of May 14, 1991 (EURLex, 1991).

In recent years, Italy, like other European nations, has integrated European Union (EU) directives designed to harmonize and modernize copyright law in the digital context. These directives address the complexities introduced by digital networks, which have transformed communication, education, creativity, and professional development.

2.1.2 Structural Analysis of Italian Copyright Law in Akoma Ntoso

In order to build the Akoma Ntoso parser, we first need to identify the key components of this particular Akoma Ntoso file, as the specification allows for different implementations, according to the needs of the use-case

Some words about Akoma Ntoso

The Akoma Ntoso XML framework, often referred to simply as Akoma Ntoso, serves as a platform-independent XML description of legal resources. In Akoma Ntoso, these documents are enriched with detailed structural annotations, enabling machine-readable processing. This enhanced readability facilitates the development of advanced legislative information systems, promoting better efficiency and transparency in parliamentary, legislative, and judicial settings. Moreover, this framework allows for the creation of software tools that can interpret documents not just as basic text, but in relation to their inherent structure and meaning (Palmirani and Vitali, 2011).

In 2019, Italy embarked on a journey to transform its entire legislative repository and judicial decisions into Akoma Ntoso format through the Lexdatafication project (Palmirani, 2021). Before this, Italy utilized different standards for encoding such data (Biagioli et al., 2004). However, these standards had their limitations, such as only extending to the article level and having inaccuracies, notably in the preamble and conclusions. They also missed details on consolidated versions and normative references.

Currently, through Italy's official portal for legal information, Normattiva, one can download legal documents that align with the Akoma Ntoso guidelines.

Metadata

The copyright legislation is enriched with a comprehensive metadata structure that aids in discerning the provenance, lifecycle, and interrelations of the legislative document. The *identification* section bestows a granular level of granularity by offering identifiers rooted in the FRBR standard (on the Functional Requirements for Bibliographic Records, 1998). Each context is meticulously annotated with the Uniform Resource Identifier (URI), date of creation or modification, authorship, and linguistic medium. Augmenting this, preservation metadata according to ELI ontology (European Union, 2023), ensures interoperability with other European legislation.

In tandem with identifying attributes, the metadata encapsulates key milestones in the document's lifecycle, providing a chronological roadmap of significant events. The *analysis* section delves into the active and passive modifications undergone by the document, spotlighting insertions and their textual specifics. Complementing this, the *references* XML element elucidates the foundational links to the document's genesis and affiliated passive citations. Finally, in the *proprietary* tag, information related to the document format are provided.

Body

The body of the copyright legislation contains the actual legislative text. It is subdivided into 49 chapters, serving as broader thematic divisions. Within these chapters, the document elaborates on its provisions through 306 articles. These articles are then detailed out in 915 paragraphs.

The legislation presents itself as a comprehensive document, organized to address the various dimensions of copyright law. It begins with key chapters such as *"DISPOSIZIONI SUL DIRITTO DI AUTORE"* and *"Soggetti del diritto,"* which lay down definitions and foundational principles of the law. Such an initiation provides clarity on the legislation's scope, the types of creative works protected, and the individuals or entities entitled to these protections.

As the document progresses, it delves into specialized domains with chapters dedicated to *"Programmi per elaboratore," "Banche di dati,"* and *"Diritti audiovisivi sportivi."* These chapters were added at a later stage to adapt the legislation to technological advancements and cultural phenomena.

The latter sections of the legislation are indicative of a holistic approach, addressing both the rights and the potential exceptions or limitations. Furthermore, chapters discussing penalties, enforcement mechanisms, and advisory bodies support a comprehensive view of rights and the procedural rules for their enforcement and guidance.

2.2 The copyright ontology

Another key element for our analysis is the choice of annotations. In this research, our emphasis is on linking the rights articulated in the legal text and their corresponding exceptions to an established domain ontology. Future endeavors could shift towards the modeling of procedural rules using alternative ontological standards like LegalRuleML (Palmirani and Governatori, 2021).

The Copyright Ontology is a formalization of concepts and relationships of copyright legislation within the domain of content rights management (García and Gil, 2006; García et al., 2007). The copyright ontology was originally released in 2006 and built upon a more comprehensive work on IPRonto, that encompasses the larger field of Intellectual Property Rights (Delgado and Garcia, 2003). Its primary objective is to enable automated and computer-supported copyright management throughout the entire content value chain, in strict accordance with copyright law. Unlike conventional rights languages and ontologies that only consider end-users' content consumption permissions, this ontology comprehensively addresses all aspects of content life-cycle. It is implemented leveraging W3C standards, such as RDF and OWL. This implementation ensures wide accessibility and ease of integration into existing systems and workflows. The ontology is available in Turtle RDF serialization format, facilitating data exchange and interoperability with other ontologies and systems.

To tackle the complexities of the copyright domain, the ontology is subdivided in three smaller sets of concepts and relationships:

- Creation Model, which captures the diverse manifestations of copyright creations as they progress through their life-cycle. It encompasses different forms of original content and tracks their transformations over time.

- Rights Model, which represents the legal constructs that regulate content actions. It accommodates various legal systems, encompassing both broad global rights frameworks advocated by organizations like WIPO and specific rights delineated in regional legal regimes.

- Action Model, which defines a comprehensive set of actions that govern the life-cycle of content. From creation to distribution, each action is meticulously outlined, facilitating content management and utilization.

By combining these three interrelated models, the Copyright Ontology empowers content creators, managers, and consumers to navigate the complexities of copyright law (García et al., 2007). In the following sections, we will explore these three subdivisions more in detail.

2.2.1 Creation model

As mentioned above, the creation model aims to capture the various forms a creative work takes during its life-cycle. This is not a new issue area. In particular, extensive research was undertaken in the context of bibliographic studies to represent works.

FRBR Creation Model

The FRBR (Functional Requirements for Bibliographic Records) model, proposed by the International Federation of Library Associations and Institutions (IFLA) (on the Functional Requirements for Bibliographic Records, 1998), is the most widely recognized model for modeling intellectual creations. Its main components are:

- Work: An abstract intellectual or artistic creation, recognized through its various expressions and characterized by its shared content.
- Expression: The realization of a Work in forms like alpha-numeric notation, sound, image, and more.
- Manifestation: The physical form of an Expression, ranging from books and maps to films and multimedia kits. It represents physical objects with shared content and characteristics.
- Item: A single instance or copy of a Manifestation.

MPEG-21 Media Value Chain Ontology

The MPEG-21 Media Value Chain Ontology (MVCO) is delineated in ISO/IEC standard MPEG-21 Part 19 and provides a robust framework for media value chains (Rodriguez-Doncel and Delgado, 2009; Gauvin et al., 2010). Key MVCO concepts include:

- Work: A creation's essence, independent of its Manifestations.
- Adaptation: A Work inspired by another.
- Manifestation: A tangible or perceivable expression of a Work, like digital files or performances.
- Instance: An example of a recognized Manifestation, like a specific file.
- Copy: A reproduction of an IP Entity. Digital Copies are near-identical, while analog Copies can vary.

While MVCO offers a granular approach, it falls short in discerning between object-based and event-based manifestations and instances, blurring distinctions in the creation process.

Overcoming limitations

The Copyright Ontology attempts to address the gaps of previous creation models to the specific needs of the domain.

Key elements and their relationships are:

- Work: An intellectual or artistic creation with forms ranging from literature, art, and music to software and databases.

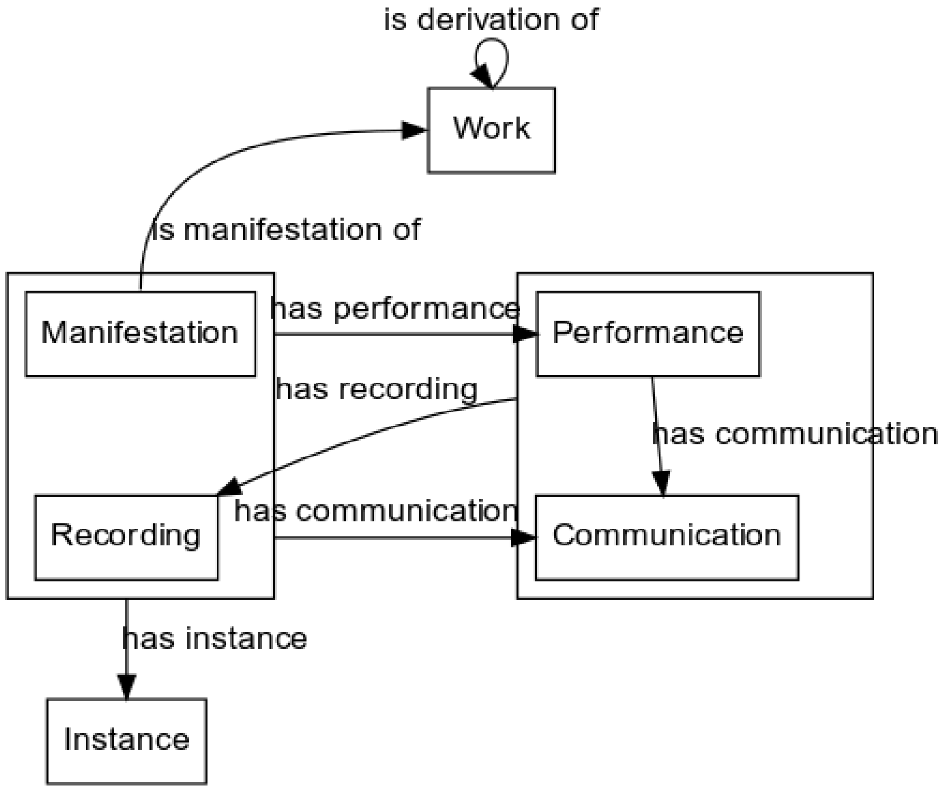

Figure 2.1: Copyright's Ontology Creation Model (García and Gil, 2006)

- Manifestation: The tangible or digital embodiment of a Work, with subtypes like Sound and Audiovisual Recordings.
- Instance: A reproduction of a Manifestation or another Instance, either physical or digital.
- Performance: The temporal expression of a Work, involving performers or technical means.
- Improvisation: A spontaneous, unrehearsed expression of a Work, distinct from a pre-existing Manifestation.
- Communication: The distribution of a Work, either as BroadcastEvent (broad) or OnDemandEvent (tailored, such as streaming).
- Live Communication: Direct relay of a performance, devoid of non-transient recordings.

These elements delineate a Work's progression to end-user consumable forms like Instance, Communication, or Performance, emphasizing directly accessible instances for users.

2.2.2 Rights model

The Rights Model of the Copyright Ontology captures the legal aspects of copyright, encompassing various rights granted to creators and other parties involved in the exploitation of works, as well as exceptions. The Rights Model is designed to be flexible and adaptable to different legal systems worldwide and it takes as reference international treaties deposited at WIPO such as the Berne Convention (WIPO, 1886) and the Beijing Treaty (WIPO, 2012). It follows recommendations from the World Intellectual Property Organisation (WIPO) and covers both economic and moral rights, as well as copyright-related rights.

The model's primary focus is on the economic rights as they pertain to the production and commercial aspects of copyright. These economic rights include Reproduction, Distribution, Public Performance, Fixation, Communication, and Transformation. Each of these rights governs a specific action on copyrighted content, such as the right to reproduce or distribute the work.

Moral rights, on the other hand, are distinct from economic rights and are inherently non-commercial. They are always retained by the creator of a work and are primarily concerned with the reputation of the author. These include the Attribution Right, which allows the creator to claim authorship, and the Integrity Right, which enables the creator to object to any modifications that might tarnish their reputation. Notably, while moral rights are integral to many legal systems, they are not universally recognized.

In addition to the aforementioned rights, the copyright ontology also addresses related rights, that are referred to as Neighbouring Rights. These rights are tailored for other key stakeholders involved in the production and dissemination of copyrighted works, such as performers, producers, and broadcasters. Their roles in bringing creations to the audience are significant, and thus they are granted exclusive rights over their contributions. For instance, performers have exclusive rights over their performances, while broadcasters have rights over their broadcasts.

The model also addresses Copyright Exceptions, serving as right limitations. These include Quotation, Education, Reporting, Official Act, Private Copy, Parody, and Temporary Reproduction exceptions, accommodating specific uses and differing per jurisdiction.

2.2.3 Actions model

The Action Model in the Copyright Ontology plays a crucial role in capturing the dynamic aspects of creation value chains. By focusing on actions performed by various actors involved in the copyright ecosystem, this "Action-Oriented Modeling" approach creates a cohesive and comprehensive framework.

At the heart of the Action Model are actions that "move" creations along their value chain, allowing for a clear understanding of how a creation progresses from a conceptual Work to a perceivable Manifestation or a dynamic Performance. The Manifest action represents the embodiment of a Work into a tangible object or Manifestation, while the Perform action involves the direct embodiment of a Work into a Performance, without the need for a prior Manifestation.

These actions are not isolated; they are interconnected with the rights identified in the Rights Model, which govern the actions and the creations they involve. For instance, the Copy action, governed by the Reproduction Right, generates replicas of a Manifestation, Recording, or Instance. The Perform action, on the other hand, is not constrained by copyright unless it takes place in public, in which case it falls under the Public Performance Right.

To capture the various nuances and complexities of copyright agreements, the Action Model includes

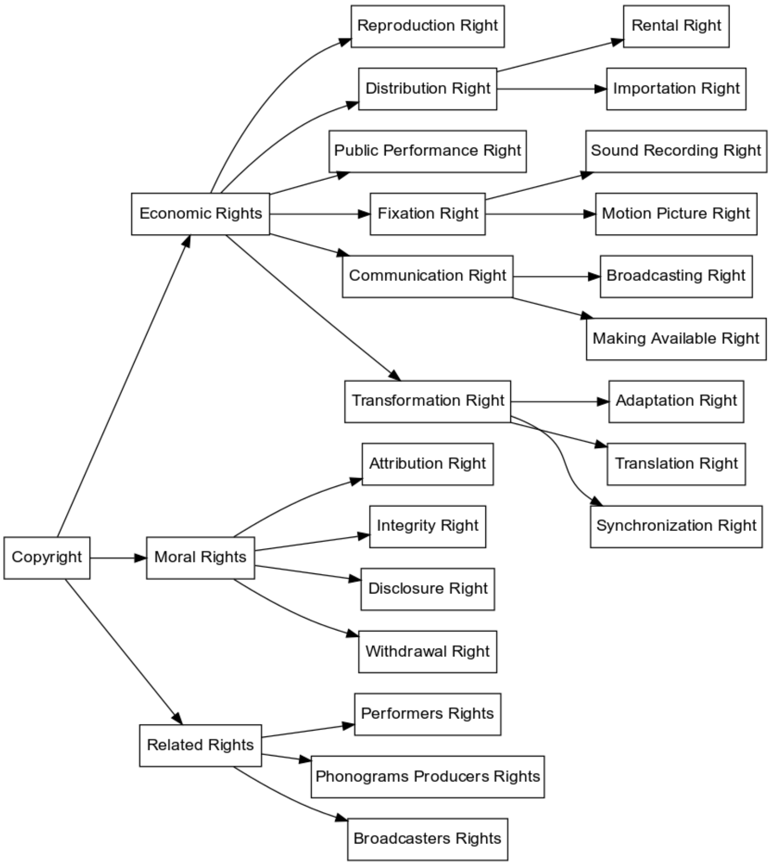

Figure 2.2: Copyright Ontology's Rights Model

specific actions for managing rights and agreements. The Agree action represents the mutual agreement of parties to abide by constrained actions related to the use or exploitation of copyrighted content. Conversely, the Disagree action revokes an existing agreement by referring to the corresponding Agree action.

For end-users of Works, the Action Model includes actions related to content consumption. The Use action encompasses the consumption of copyrighted content by end-users, with specific kinds of uses like Access, or Copy, allowing users to consume content via on-demand access, live performances, or broadcasts.

2.3 The Doccano Annotation Tool

Another key element of our data transformation pipeline is Doccano (Nakayama et al., 2018), an open-source web application that supports a variety of annotation needs. Its flexibility and user-friendly interface make it a popular choice among researchers and developers in the fields of natural language processing (NLP) and machine learning.

2.3.1 Features

Doccano offers a range of features that cater to different annotation needs:

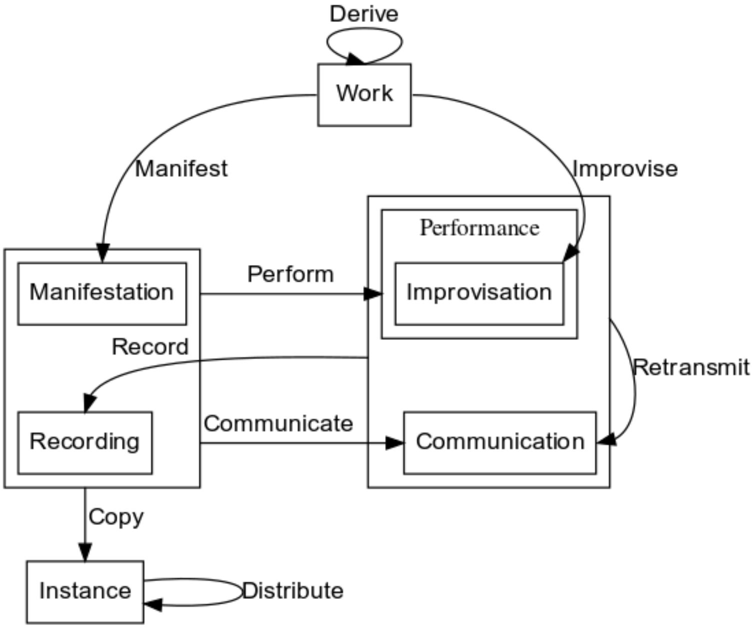

Figure 2.3: Copyright Ontology's Action Model

- Labeling Formats: Supports various annotation types including named entity recognition (NER), text classification, and sequence-to-sequence tasks.
- Collaboration: Multiple annotators can work on the same dataset simultaneously, facilitating collaborative projects.
- Custom Label Creation: Users can define their own labels tailored to specific projects.
- Import/Export: Easy import and export of data in formats like JSON and CSV.
- User Management: Role-based access ensures data security and project management.
- Semi-automated or automated annotation support

2.3.2 Applications in Legal Text Annotation

Doccano's adaptability makes it a valuable tool for various domains, including legal text annotation. Its ability to handle large volumes of text and support for custom labels allows for efficient annotation of complex legal documents. For instance, labels specific to legal entities, acts, or jurisdictions can be created to extract pertinent information from legal texts.

2.4 Conclusion

This chapter discussed the data and tools that are being used for the project work. Initially, we examined the structure of the Italian copyright legislation, formatted in Akoma Ntoso (Gazzetta Ufficiale, 1941). In the second section, we explored the previous work by García and Gil (2006) on the development of the Copyright Ontology. The ’Creation model’ explained how works qualify for copyright. The ’Rights model’ detailed the protections given to creators. The ’Actions model’ showed the different activities and interactions possible within copyright. In the third section, we briefly presented the Doccano annotation tool and its main features (Nakayama et al., 2018).

In the next chapter we will go through the various components of the data transformation pipeline and how they interact with each other.

Chapter 3

Doccano-Semantics

Having outlined the key ingredients for the data transformation pipeline[1], this chapter focuses on the various system components, up to the final transformation of the legal provisions in the Lynx format specification (Rodríguez-Doncel et al., 2023). Finally, it discusses the results and compares the data transformation pipeline with other existing tools for the semantic annotation of legal texts.

3.1 System description

The data transformation pipeline, as illustrated in Figure 3.1, is designed to enable seamless end-to-end semantic annotation. It encompasses the following steps:

- Parse files from Akoma Ntoso format: In this initial stage, we extract the legal provisions at the paragraph level from files structured in Akoma Ntoso, while preserving relevant metadata, such as paragraph, article, and chapter number, the presence of insertions or reference, and others. This components outputs the provisions in a JSONLines file[2], required for the upload via the Doccano web interface.
- Parse relevant ontologies and controlled vocabularies into a Doccano-compatible format (JSON): Doccano does not support ontological classes and controlled vocabularies out of the box, but it has an "import labels" functionality that requires JSON files with the following attributes: label, suffix_key, text_color, and background_color.
- Annotate using the Doccano tool: Doccano provides a user-friendly interface where labels and relationships are easy to find in the respective ontologies. Appropriate user-interface and database adaptations of Doccano's open-source code were made necessary, as the tool did not initially support more than 26 different labels. This increased the complexity of the work, as it required an

[1] Namely, a legislative text (Law 633 of 1942, (Gazzetta Ufficiale, 1941)), a set of potential labels (the Copyright Ontology (García and Gil, 2006)) and the annotation tool (Doccano (Nakayama et al., 2018))

[2] A file containing lines of JSON files

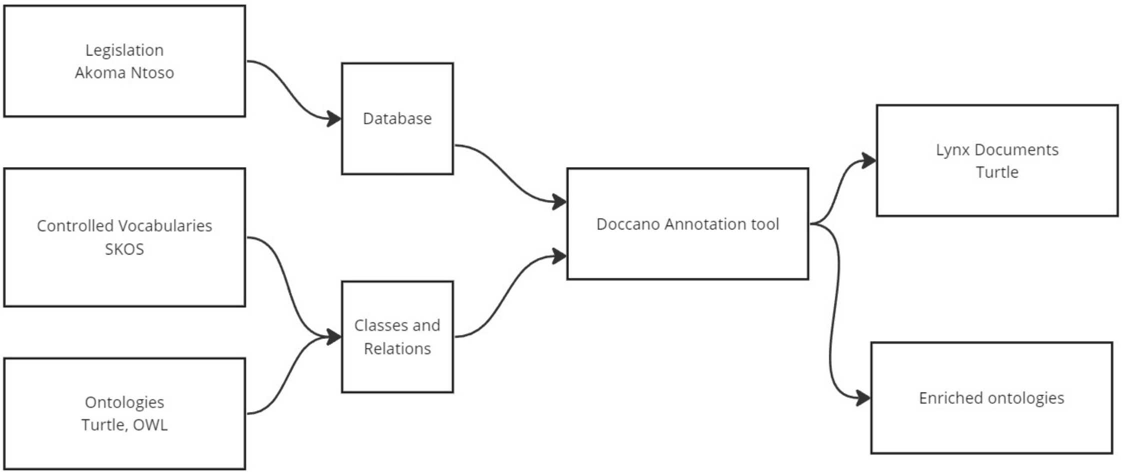

Figure 3.1: Pipeline flowchart

analysis of the architecture of the tool. The tool's user-friendly interface enables the annotator to mark and categorize important information according to their needs.

- Map the Doccano annotated output file to a Lynx format: we transform the annotated JSONL file from Doccano into the Lynx format, an offset-based document format for legal compliance.

3.2 System's components

In this section, we will present in detail the various components of the data transformation pipeline.

3.2.1 Akoma Ntoso preprocessing

Akoma Ntoso is one of the major standards used for structuring machine-readable legal documents across a variety of legislation. While some applications support it out-of-the box, Doccano is not among them. Therefore, it is key to transform these files in a format suitable for it (JSONLines). This takes place by calling the AkomaNtosoParser class, which is instantiated with the file path of the Akoma Ntoso file and then leverages the ElementTree library to parse it. The component undergoes the following sequential steps:

- Parsing Chapters: the extract_chapters method parses chapters' titles and ids and stores them as a DataFrame in an attribute of the class instance
- Parsing Articles: the extract_articles method parses the Akoma Ntoso file at the article level and stores them as a list in an attribute of the class instance
- Parsing Paragraphs: the extract_paragraphs method parses the paragraph attribute of the class instance and stores them as a DataFrame in an attribute of the class instance

- Creating the DataFrame: the create_dataframe method merges all the information regarding chapters, articles and paragraph in a single DataFrame and stores it in an attribute of the class instance
- Data Export: the export_jsonl method converts the stored DataFrame in JSON writes it to a JSONLines file.

3.2.2 Label preprocessing: OWL and SKOS

SKOS

SKOS is a widely-used RDF vocabulary for representing and managing knowledge organization systems. SKOS files can contain valuable information, and extracting top concepts is crucial for effective knowledge representation and visualization. This work introduces SKOSParser, a Python component designed to parse SKOS files and identify top concepts.

The SKOSParser class is instantiated with the file path to the SKOS file in XML format. The component employs the rdflib library to create a graph object, which parses and represents the SKOS file. Subsequently, the SKOSParser performs the following key functionalities:

- Parsing and Extracting Top Classes: The parse_skos method parses the SKOS file, extracting essential concepts, and filters out top classes by considering SKOS:Concept instances without SKOS:topConceptOf relationships. The parser retrieves the concept's URI, preferred label (in English), and an associated prefix. The top classes are then stored in a structured format as a list of dictionaries.
- Enrichment with Random Colors: The enrich_json method processes the extracted top concepts, appending a human-readable label by concatenating the concept's prefix and preferred label. The method also assigns random background colors to each SKOS file to enhance visual representation during visualization. The enriched data is stored in a standardized JSON format.
- Data Export: The SKOSParser class includes a write_to_file method that exports the enriched data to a JSON file compatible for the upload in the Doccano annotation tool.

The SKOSParser component demonstrates a proficient solution for extracting top concepts from SKOS controlled vocabularies. By leveraging Python's rdflib library, the component efficiently navigates and processes RDF graph data, yielding structured information about top classes.

OWL

The OntologyParser is a Python class which is instantiated with the file path to the RDF ontology in Turtle format. The parser utilizes rdflib to create a graph object, subsequently parsing the ontology file to populate the graph. The parser then employs the random_color method to assign a random color to each ontology.

The key functionalities of the OntologyParser are as follows:

- Class Extraction: The component traverses the ontology graph to identify classes with RDF type OWL.Class. Utilizing a predefined set of namespaces, the parser associates each class with a human-readable and concise class name, while preserving its full URI reference. The parser stores class information in a structured format, including class names, suffix keys, background colors, text colors, and URIs.

- Property Extraction: The parser further identifies RDF Object Properties in the ontology, extracting and organizing them similarly to classes. Utilizing the graph's triples, the component extracts property names and stores them alongside relevant attributes, such as suffix keys and background colors.

- Data Export: The Ontology Parser provides methods to obtain and export extracted class and property data as JSON files, supported for their loading in the Doccano annotation tool.

3.2.3 Adaptation of the Doccano annotation tool

To overcome the challenge of accommodating numerous classes and properties stored in ontologies, we had to adapt a small portion of the Doccano source code. The original web app only supports a limited number of labels (26).

To address this, the Django model *LabelType* had to be modified by extending the the maximum length of the *suffix_key* field to 254 to allow for up to *256!* different combinations.

This enabled the adoption of the URIs of ontological classes as unique identifiers in the database, thus promoting compatibility with ontological data sources.

3.2.4 Manual annotation

Since the development of a semi-automated system for semantic annotation is beyond the scope of this project, annotations were manually done by the researcher. The main goal was to pinpoint elements in the legislation offering a legal foundation for the copyright ontology. This entailed a methodical approach, aided by the thematic organization of the Italian legislation. For instance, the section "Opere Protette" (Protected Works) outlines what qualifies as a subject of copyright protection, essentially, what is viewed as a creative work. In this section, only two ontological classes were assigned to specific provisions: schema1:CreativeWork and cro:Work. However, the subsequent section, "Soggetti del Diritto" (Subjects of Law), delves into the act of creation, remaining within the Creative Model of the Copyright ontology.

Key observations from the annotation process include the realization that while most declarative provisions aligned directly with Copyright Ontology classes, those indicating specific rules and procedures necessitated a distinct representation, especially concerning elements tied to deontic logic. This indicates a potential need to integrate with other ontologies like LegalRuleML or LKIF (Legal Knowledge Interchange Format). Such integration implies a layered, multi-step annotation process, enabling diverse machine-readable representations to accurately capture nuances in natural language. Legal texts' inherent ambiguity allows them to abstractly convey intricate information, presenting a challenge when trying to express such content explicitly.

The resulting annotated dataset primarily aids in targeted retrieval. Enriching the original Copyright Ontology with foundational details from Italian legislation streamlines the search process, though the benefits currently appear modest. However, expanding the annotation to multiple legislations would likely prove useful, granting legal scholars direct access to shared concepts across different jurisdictions.

3.2.5 Representing Legal concepts in Lynx

After having completed the semantic annotation of the examined legal documents, the last step of the pipeline concerns their export in a machine-readable a reusable format. In this context, we identified the Legal Knowledge Graph Ontology (Rodríguez-Doncel et al., 2023), developed in the context of the Lynx project (Rodriguez-Doncel, 2023) as a promising candidate format for encoding the annotations in a machine-readable format.

The Lynx specification was developed to provide a suitable standard for representing documents related to compliance, through the class LynxDocument. Lynx Documents adhere to the NIF (NLP Interchange Format) specification and extensively incorporate ELI metadata elements. Additionally, they can be assembled into collections and augmented with annotations.

This system component is centered around the LynxDocument class. It extracts entities and their metadata from annotated text documents and convert them into a RDF format according to the Lynx specification. The key functionalities are as follows:

- Annotation Conversion: LynxDocument reads and processes Akoma Ntoso and SKOS elements from a JSON file, converting them into a format suitable for RDF graph representation. The Akoma Ntoso elements define metadata, while the SKOS elements represent predefined concepts and their URIs.

- Graph Initialization: LynxDocument initializes an RDF graph, binding relevant prefixes to namespaces used in the graph representation. The SKOS and Akoma Ntoso elements are then loaded into the graph.

- Document Representation: LynxDocument generates a URI for each document based on its identifier and type. The tool adds document metadata, such as language and identifiers at the paragraph level, if available in the Akoma Ntoso elements.

- Text and Annotation Addition: LynxDocument extracts the text from the annotated document and adds it to the RDF graph. The tool annotates entities within the text by creating RDF nodes for each entity span.

- Serialization: LynxDocument serializes the generated RDF graph into Turtle format.

3.3 Discussion

3.3.1 Limitations

The data pipeline named doccano-semantics, has several limitations:

- Annotation capabilities: as of now, the system supports only manual annotation. While this ensures precision and attention to detail, we recognize the potential for automated or semi-automated annotation tools in the future.
- User interface: the current user interface in Doccano is functional, but there's room for further refinement, in particular for the management of the ontology labels and the integration of the command line utilities directly in the user interface.
- Limited support for Akoma Ntoso documents: the system is tailored to parse Italian Akoma Ntoso documents, as there are still slighly different implementations of the standard across jurisdiction.
- Limited support for output format: currently, the only output format available is the Lynx format. This can be restrictive for users who are looking for different markup languages. Expanding our output format options would increase the overall utility of the system.

3.3.2 Overview of existing platforms

Several annotation tools exist for the legal domain, therefore extensive comparison with equivalent platforms would be a complex and time-intensive task, therefore for the purpose of this project work, it was conducted a literature review and comparison of functionalities and markup languages supported.

The research group working in the University of Bologna has been developing a series of utility tools[3] supporting their sponsored specifications[4] as well as annotation platforms.

RAWE is a web-based WYSIWYG editor designed for legal experts. It enables the modeling of Akoma Ntoso legal texts in deontic rules structured as valid LegalRuleML, as well as the inclusion of other external rules (Palmirani et al., 2013). While promising, it is does not allow the upload of user-defined ontologies, as doccano-semantics.

Notable is also the LIME platform, which is able to parse Akoma Ntoso documents and provide some form of NER-based automatic annotation. However, there is little support for manual annotation and the inclusion of user-defined ontologies or vocabularies, a core feature of doccano-semantics.

An interesting development is the University of Auckland's "LegalRuleML editor with transformer-based autocompletion" (Fuchs et al., 2023), showcased in 2023 at the European Conference on Computing in Construction. However, although this platform provides semantic-aware translation of legal rules in a machine-consumable language, it does not really annotate the legal texts (as doccano-semantics does),

[3] Among the conversion tools we can list Formex2AKN, Norme in Rete to AKN, Text2AKN

[4] Akoma Ntoso and LegalRuleML

which is a key element of explainability. In any case, the use of LLMs for semantic annotation is a promising field of research.

Finally, the Maastrict University's Law & Tech lab is carrying on the Lawnotation project (2022-2024), which aims to develop an infrastructure that allows researchers to making legal data and annotation schemes (current and future) accessible for annotation and analysis purposes, and to develop an annotation platform for analyzing the linguistic and legal characteristics of legal documents. There is a proof-of-concept but access to external users is still limited. However, the tool shares the same goals are in line with the ones of doccano-semantics and it leverages the same frontend framework (Nuxt.js),

While these platforms are considered promising, the complexities of legal texts are numerous and at this stage, we lack a unified platform that adequately addresses them (Huang et al., 2022; Reeve and Han, 2005).

3.3.3 Ethical aspects

The doccano-semantics system, at its current stage, does not necessitate an ethical impact assessment as it is not fundamentally based on AI technologies, rather on manual annotation. The system primarily aims to facilitate best practices of data quality for subsequent AI-based applications. Its focus is on data preparation and organization rather than autonomous decision-making or predictive analytics.

However, should the system evolve to include semi-automatic or automatic semantic annotation functionalities in the future, an ethical impact assessment may become required. In the case of the usage of data produced by doccano-semantics in downstream AI systems, an AI ethical impact assessment may be considered.

3.3.4 Future Work

In this study, we present a comprehensive end-to-end pipeline, yet several enhancements warrant future exploration. Specifically:

- Akoma Ntoso Version Support: The current AkomaNtosoParser class is tailored to the Italian legislative implementation. However, preliminary trials with the Luxembourg and France versions have indicated discrepancies. Refinement in extraction methodologies is essential.

- User Interface Integration: Currently, the conversion to Doccano-compatible formats and file exports take place through a command-line interface. Integrating these processes within the Doccano web interface would significantly enhance user experience.

- Semi-automated Annotation: The Doccano platform permits automated annotations via external API services. Implementing a diverse set of semantic annotation algorithms could expedite the annotation rate and increase the time experts can spend on high-value decisions.

- Scaling up: A plug and play extension model that connects the academics and open source communities to share legal semantic annotation algorithms at scale could largely improve the interchange of semantic annotation algorithms. APIs guarantee the technical interoperability of such a model. Some identified algorithms are the followings: Fuchs et al. (2023); Ceci et al. (2012); Lesmo et al. (2009); Savelka (2023); Camilleri et al. (2016)
- Expand the set of export representations: While the selected annotation specification is Lynx, optimized for embedding the semantic annotations of legal documents in Turtle format, it may not be the optimal tool for delineating procedural rules and deontic logic. Exploring alternative specifications, such as Akoma Ntoso (Palmirani and Vitali, 2015), LegalRuleML (Palmirani and Governatori, 2021), or Catala (Merigoux et al., 2021), could offer more comprehensive representations.

3.3.5 Results

This research adheres to the FAIR principles (Wilkinson et al., 2016) and it produced the following results:

- This thesis: this document describes the research objectives, methodology, and results of the project work.
- A data pipeline for semantic annotation of legal texts in Akoma Ntoso format, as well as the description of the internal components. The software code is freely accessible in the Github Repository at the following address https://github.com/AlessioNar/semanticannotation.
- A modified version of doccano, named doccano-semantics. The fork of the doccano repository is available at the following address https://github.com/AlessioNar/doccano-semantics
- A partial annotation of the Italian copyright legislation, the dataset is available for download in Zenodo at the following address https://doi.org/10.5281/zenodo.8312755
- A partial representation of the Italian copyright legislation in the Lynx format, the dataset is available for download in Zenodo at the following address https://doi.org/10.5281/zenodo.8312759

3.4 Conclusion

Throughout this chapter, we have presented the data transformation pipeline built for streamlining the semantic annotation of legal documents by leveraging the Doccano tool and the Lynx specification. We presented the different components of the system and provided an initial evaluation of the system. Finally, we presented and discussed the results of the work and compared it with other existing platforms.

Conclusions

This project work focused on the development of a streamlined approach for the semantic annotation of legal texts. The primary objective was to enhance support for legislative documents and to augment an existing ontology by associating specific legal provisions to their representation in Turtle format.

Chapter 1 provided an in-depth examination of the theoretical underpinnings of semantic annotation in legal texts. The investigation encompassed topics such as legal isomorphism, deontic logic, common methodologies for annotation, and prior work in semi-automated and automated annotation.

In Chapter 2, the theoretical foundation was leveraged to establish the groundwork for the data pipeline. This section detailed the core components of the project, including the available data and an initial analysis of its structure. As a practical example, the Italian copyright legislation was annotated with the copyright ontology, illustrating the data transformation pipeline (Gazzetta Ufficiale, 1941; García and Gil, 2006).

Chapter 3 was dedicated to describe in detail the data transformation pipeline, outlining the strategy employed for manual annotation. Additionally, it identified the system's limitations and proposed potential avenues for future research. The process involved the enhancement of data flow from Akoma Ntoso, SKOS, and OWL to Doccano (Nakayama et al., 2018), an open-source annotation tool. Subsequently, the data was transformed according to the Lynx specification, a format tailored for legal documents that accommodates offset-based annotation and integrates seamlessly with other pertinent ontologies, including ELI (European Union, 2023) and NIF.

Transforming legislative resources into machine-readable formats is not merely a technical endeavor but a vital research area that affects various sectors of governance. Supported by multiple governments, particularly in the European Union, this transformation seeks to enhance interoperability among not only public administration systems but also businesses and legal entities.

The way in which legal resources interact with the different areas of society is a complex process that permeates all levels of public and private life. Public services and businesses often face a labyrinth of information that can be difficult to navigate, leading to inefficiencies in public service delivery and unsustainable administrative burden.

Recognizing this challenge, the European Union has launched several initiatives aiming to streamline these processes and reduce the burden on legal entities. The focus is not just on the reduction of administrative complexities but also on enhancing the very fabric of legal interactions.

As argued in this thesis, one promising approach is the semantic annotation of legal texts, which can enable the automation of parts of the implementation of the legal provisions enshrined in the texts. By fostering shared interpretations of legal provisions right at the policy design stage, the Member States can align their legal frameworks. This alignment can reduce the effort needed in the subsequent implementation, promote transparency, and ensure uniform applicability across the regional block.

However, the complexity of legal systems requires a multifaceted approach. Shared conceptualizations and semantic interoperability are vital but constitute only one part of a larger puzzle. This layer of interoperability, while essential, must be integrated with other aspects to form a cohesive framework that truly addresses the complexities of legal systems.

Enhancements in semantic interoperability can indeed have a cascading effect on other areas, facilitating legal interoperability across diverse legislations. Different applications and interpretations of legal texts are made explicit, allowing for a more nuanced understanding of existing legislation. Gaps can be bridged, and bottlenecks in legislation can be identified more efficiently.

Such an integrated approach offers a roadmap to ease legislative complexities, allowing policymakers to assess whether existing laws need to be updated or amended. It opens new possibilities for more transparent, uniform, and effective governance, potentially revolutionizing the way legal systems interact and operate. By considering legislation not merely as static texts but as dynamic, interconnected systems, this transformation can contribute to a more efficient, transparent, and responsive legal framework across the European Union and potentially beyond.

Appendix A

On the use of LLM-powered AI tools in this thesis

For this work, we employed cutting-edge LLM-based technologies to support the package development and assist in academic writing. These tools proved invaluable in analyzing the structure of legal texts, restructuring the initial code, identifying pertinent papers, and refining the English language for improved readability.

Nevertheless, the intellectual rigor and effort invested in the research and documentation remain undiminished. While these tools aided the process, the content is original, shaped by critical thinking and the researcher's intentions.

Since the end of 2022, the widespread diffusion of generative AI tools has opened several discussion within academia and traditional dilemmas related to authorship and plagiarism have resurfaced, sooner than expected.

I strongly believe that these dilemmas could be solved by adopting the legal lenses of copyright legislation and specifically, in the definition of creative work. As already mentioned in chapter 2, *a "work" encompasses any set of materials organized intentionally* and *originality doesn't depend on innovation but merely requires some form of intellectual endeavor*.

Under this light, the use of generative AI tools, when purposefully directed and organized, does not represent a threat to academia, rather an opportunity for increasing the allot of time available to the researcher for the profound reflection on the motivations, objectives, and scope of the research, and thus the quality of the intellectual output.

Bibliography

Adebayo, K. J., Di Caro, L., and Boella, G. (2018). Towards annotation of legal documents with ontology concepts. In Pagallo, U., Palmirani, M., Casanovas, P., Sartor, G., and Villata, S., editors, *AI Approaches to the Complexity of Legal Systems*, pages 337–349, Cham. Springer International Publishing.

Asooja, K., Bordea, G., Vulcu, G., O'Brien, L., Espinoza, A., Abi-Lahoud, E., Buitelaar, P., and Butler, T. (2015). Semantic annotation of finance regulatory text using multilabel classification. *LeDA-SWAn (to appear, 2015)*, 8.

Athan, T., Governatori, G., Palmirani, M., Paschke, A., and Wyner, A. Z. (2014). Legal interpretations in legalruleml. In *SW4LAW+DC@JURIX*.

Bench-Capon, T. J. M. and Coenen, F. P. (1992). Isomorphism and legal knowledge based systems. *Artificial Intelligence and Law*, 1.

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., and Soria, C. (2005). Automatic semantics extraction in law documents. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ICAIL '05, page 133–140, New York, NY, USA. Association for Computing Machinery.

Biagioli, C., Francesconi, E., Spinosa, P.-L., and Taddei, M. (2004). Xml documents within a legal domain: Standards and tools for the italian legislative environment. In *International Workshop on Document Analysis Systems*.

Brank, J., Leban, G., and Grobelnik, M. (2018). Semantic annotation of documents based on wikipedia concepts. *Informatica (Slovenia)*, 42.

Camilleri, J. J., Gruzitis, N., and Schneider, G. (2016). Extracting formal models from normative texts. In *International Conference on Applications of Natural Language to Data Bases*.

Ceci, M., Lesmo, L., Mazzei, A., Palmirani, M., and Radicioni, D. P. (2012). Semantic annotation of legal texts through a framenet-based approach. *Lecture Notes in Computer Science*, page 245–255.

Commission, E., Directorate-General for Financial Stability, F. S., and Union, C. M. (2021). *Implementing dictionaries of regulatory concepts and reporting obligations by assisted machine learning – Final report*. Publications Office.

Delgado, J. and Garcia, R. (2003). Intellectual Property Rights Ontology Specification — dmag.ac.upc.edu. https://dmag.ac.upc.edu/ontologies/ipronto/. [Accessed 31-07-2023].

Digitaal Vlaanderen (2023). Kantoorbelasting — kantoorbelasting.aeco.cloud. https://kantoorbelasting.aeco.cloud/. [Accessed 16-08-2023].

EURLex (1991). Council directive 91/250/eec of 14 may 1991 on the legal protection of computer programs. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:31991L0250.

European Union (2023). ELI - EU Vocabularies - Publications Office of the EU — op.europa.eu. https://op.europa.eu/en/web/eu-vocabularies/eli. [Accessed 20-08-2023].

Francesconi, E. (2019). Ai and law: Semantic annotation of legal texts.

Francesconi, E. and Passerini, A. (2007). Automatic classification of provisions in legislative texts. *Artificial Intelligence and Law*, 15:1–17.

Fuchs, S., Dimyadi, J., Ronee, A. S., Gupta, R., Witbrock, M., and Amor, R. (2023). A legalruleml editor with transformer-based autocompletion. In *Proceedings of the 2023 European Conference on Computing in Construction and the 40th International CIB W78 Conference*, volume 4 of *Computing in Construction*, Heraklion, Greece. European Council on Computing in Construction.

García, R. and Gil, R. (2006). An owl copyright ontology for semantic digital rights management. In Meersman, R., Tari, Z., and Herrero, P., editors, *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, pages 1745–1754, Berlin, Heidelberg. Springer Berlin Heidelberg.

García, R., Gil, R., and Delgado, J. (2007). A web ontologies framework for digital rights management. *Artificial Intelligence and Law*, 15(2):137–154.

Gauvin, M., Delgado, J., Rodriguez-Doncel, V., and Choi, M. (2010). Media value chain ontology. Retrieved from https://dmag.ac.upc.edu/ontologies/mvco/.

Gazzetta Ufficiale (1941). Legge n. 633 del 22/04/1941. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:regio.decreto:1941-04-22;633.

Gazzetta Ufficiale (1992). Legge n. 518 del 29/12/1992. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:1992-12-29;518.

Gazzetta Ufficiale (1996). Legge n. 205 del 15/03/1996. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:1996;205.

Ghijsen, M., van der Ham, J., Grosso, P., Dumitru, C., Zhu, H., Zhao, Z., and de Laat, C. (2013). A semantic-web approach for modeling computing infrastructures. *Computers & Electrical Engineering*, 39(8):2553–2565.

Governatori, G. (2005). Representing business contracts in ruleml. *Int. J. Cooperative Inf. Syst.*, 14:181–216.

Huang, Y.-T., Lin, H.-R., and Liu, C.-L. (2022). Toward an integrated annotation and inference platform for enhancing justifications for algorithmically generated legal recommendations and decisions. In *International Conference on Legal Knowledge and Information Systems*.

Humphreys, L., Boella, G., Di Caro, L., Robaldo, L., van der Torre, L., Ghanavati, S., and Muthuri, R. (2020). Populating legal ontologies using semantic role labeling. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2157–2166, Marseille, France. European Language Resources Association.

Jones, A. J. I. and Sergot, M. J. (1992). Deontic logic in the representation of law: Towards a methodology. *Artificial Intelligence and Law*, 1:45–64.

Karpf, J. and Københav, H. (1989). Quality assurance of legal expert systems. In *III International Congress "Logica, Informica, Diritto" Expert Systems in Law, Florence, Italy, November 2 - 5, 1989*.

Katz, D. M., Dolin, R., and Bommarito, M. J. (2021). *Legal Informatics*. Cambridge University Press.

Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. (2003). Semantic annotation, indexing, and retrieval. In *International Workshop on the Semantic Web*.

Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, J. R., and Mylopoulos, J. (2006). Text mining through semi automatic semantic annotation. In *Practical Aspects of Knowledge Management*.

Lesmo, L., Mazzei, A., and Radicioni, D. P. (2009). Extracting semantic annotations from legal texts. In *ACM Conference on Hypertext & Social Media*.

Loutsaris, M. A. and Charalabidis, Y. (2020). Legal informatics from the aspect of interoperability: A review of systems, tools and ontologies. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, ICEGOV '20, page 731–737, New York, NY, USA. Association for Computing Machinery.

Ma, M. and Wilson, B. (2021). The legislative recipe: Syntax for machine-readable legislation. *Nw. J. Tech. & Intell. Prop.*, 19:107.

Merigoux, D., Chataing, N., and Protzenko, J. (2021). Catala: A programming language for the law. *Proc. ACM Program. Lang.*, 5(ICFP).

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Nazarenko, A., Lévy, F., and Wyner, A. Z. (2018). An annotation language for semantic search of legal sources. In *International Conference on Language Resources and Evaluation*.

Nazarenko, A., Lévy, F., and Wyner, A. Z. (2021). A pragmatic approach to semantic annotation for search of legal texts - an experiment on gdpr. In *International Conference on Legal Knowledge and Information Systems*.

on the Functional Requirements for Bibliographic Records, I. S. G., editor (1998). *Functional Requirements for Bibliographic Records, Final report*. K. G. Saur, Berlin, Boston.

Palmirani, M. (2021). Lexdatafication: Italian Legal Knowledge Modelling in Akoma Ntoso. In Rodríguez-Doncel, V., Palmirani, M., Araszkiewicz, M., Casanovas, P., Pagallo, U., and Sartor, G., editors, *AI Approaches to the Complexity of Legal Systems XI-XII*, pages 31–47, Cham. Springer International Publishing.

Palmirani, M., Brighi, R., and Massini, M. (2003). Automated extraction of normative references in legal texts. In *International Conference on Artificial Intelligence and Law*.

Palmirani, M., Cervone, L., Bujor, O., and Chiappetta, M. (2013). Rawe: An editor for rule markup of legal texts. In *International Web Rule Symposium*.

Palmirani, M. and Governatori, G. (2021). LegalRuleML Core Specification Version 1.0 — docs.oasis-open.org. `https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/os/legalruleml-core-spec-v1.0-os.html`. [Accessed 25-07-2023].

Palmirani, M. and Vitali, F. (2011). *Akoma-Ntoso for Legal Documents*, pages 75–100. Springer Netherlands, Dordrecht.

Palmirani, M. and Vitali, F. (2015). Akoma Ntoso Version 1.0 Part 1: XML Vocabulary — docs.oasis-open.org. `https://docs.oasis-open.org/legaldocml/akn-core/v1.0/csprd01/part1-vocabulary/akn-core-v1.0-csprd01-part1-vocabulary.html#_Toc417639619`. [Accessed 25-07-2023].

Reeve, L. H. and Han, H. (2005). Survey of semantic annotation platforms. In *ACM Symposium on Applied Computing*.

Rodriguez-Doncel, V. (2023). Lynx - Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe — lynx-project.eu. `https://lynx-project.eu/`. [Accessed 17-08-2023].

Rodriguez-Doncel, V. and Delgado, J. (2009). A media value chain ontology for mpeg-21. *IEEE MultiMedia*, 16(4):44–51.

Rodriguez-Doncel, V., Santos, C., Casanovas, P., Gómez-Pérez, A., and Gracia, J. (2018). A linked data terminology for copyright based on ontolex-lemon. In Pagallo, U., Palmirani, M., Casanovas, P., Sartor, G., and Villata, S., editors, *AI Approaches to the Complexity of Legal Systems*, pages 410–423, Cham. Springer International Publishing.

Rodríguez-Doncel, V., Karampakis, S., Maganza, F., Bernardos, S., and Moreno-Schneider, J. (2023). Legal knowledge graph ontology. `https://lynx-project.eu/doc/lkg/`. [Accessed 25-07-2023].

Santos, C., Casanovas, P., Rodríguez-Doncel, V., and van der Torre, L. (2018). Reuse and reengineering of non-ontological resources in the legal domain. In Pagallo, U., Palmirani, M., Casanovas, P., Sartor, G., and Villata, S., editors, *AI Approaches to the Complexity of Legal Systems*, pages 350–364, Cham. Springer International Publishing.

Savelka, J. (2023). Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. *ArXiv*, abs/2305.04417.

Sharma, S., Gamoura, S., Prasad, D., and Aneja, A. (2021). Emerging legal informatics towards legal innovation: Current status and future challenges and opportunities. *Legal Information Management*, 21(3-4):218–235.

Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L. C., Ceci, M., and Dann, J. (2020). An automated framework for the extraction of semantic legal metadata from legal texts. *Empirical Software Engineering*, 26.

Soavi, M., Zeni, N., Mylopoulos, J., and Mich, L. (2022a). From legal contracts to formal specifications: A systematic literature review. *SN Computer Science*, 3.

Soavi, M., Zeni, N., Mylopoulos, J., and Mich, L. (2022b). Semantic annotation of legal contracts with contrattoa. *Informatics*, 9:72.

Spence, M. (2007). *Intellectual Property*. Oxford University Press.

Stegmeier, J., Hartig, J., Levstáková, M., Logan, K. T., Bartsch, S., Rapp, A., and Pelz, P. F. (2021). Development of an annotation schema for the identification of semantic uncertainty in din standards. In *Uncertainty in Mechanical Engineering*.

Tang, M., Su, C., Chen, H., Qu, J., and Ding, J. (2020). Salkg: A semantic annotation system for building a high-quality legal knowledge graph. *2020 IEEE International Conference on Big Data (Big Data)*, pages 2153–2159.

von Wright, G. H. (1951). Deontic logic. *Mind*, 60(237):1–15.

Waismann, F. (1947). Verifiability. *Journal of Symbolic Logic*, 12(3):101–101.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. O. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S. C., Evelo, C. T. A., Finkers, R., González-Beltrán, A. N., Gray, A. J. G., Groth, P., Goble, C. A., Grethe, J. S., Heringa, J., ’t Hoen, P. A. C., Hooft, R. W. W., Kuhn, T., Kok, R. G., Kok, J. N., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R. C., Sansone, S.-A., Schultes, E. A., Sengstag, T., Slater, T., Strawn, G. O., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E. M., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

WIPO (1886). Berne convention for the protection of literary and artistic works.

WIPO (2012). Beijing treaty on audiovisual performances.

Zeni, N., Kiyavitskaya, N., Cordy, J. R., Mich, L., and Mylopoulos, J. (2008). Annotating regulations using cerno: An application to italian documents - extended abstract. *2008 Third International Conference on Availability, Reliability and Security*, pages 1437–1442.

Zeni, N., Kiyavitskaya, N., Mich, L., Cordy, J. R., and Mylopoulos, J. (2013). Gaiust: supporting the extraction of rights and obligations for regulatory compliance. *Requirements Engineering*, 20:1 – 22.