Managing Legal Data: A Framework for Document Annotation and Retrieval

Valerio Bellandi[1], Stefano Siccardi[2]

[1]Università Degli Studi di Milano, Department of Computer Science, Via Celoria 18, Milano, Italy
[2]Consorzio Interuniversitario Nazionale per l'Informatica, Via Ariosto, 25, Roma, Italy

Abstract

This paper presents a structured approach to legal document management and analysis, focusing on the integration of structured and unstructured data for enhanced legal investigations and decision-making. The proposed framework incorporates Natural Language Processing (NLP) techniques, Named Entity Recognition (NER), and an Entity Registry (EReg) to annotate, disambiguate, and retrieve information from legal documents. Additionally, the system leverages graph databases and distributed architectures to facilitate semantic exploration, statistical reporting, and document retrieval. By combining rule-based and machine learning approaches, the framework provides legal professionals with advanced tools for data extraction, knowledge discovery, and pattern recognition within complex legal datasets. The paper also discusses scalability challenges and future research directions for optimizing legal data processing and information retrieval.

Keywords

Legal Documents, Entity Registry, annotation

1. Introduction

The present paper describes a work, lasted several years, for the Italian Ministry of Justice, to manage data legal documents and data. We will concentrate on data from civil cases, even if we also dealt with criminal proceedings, especially investigations.

Our starting aim, actually, was to support investigations, see [1]. In today's world we are confronted with increasing amounts of information every day coming from a large variety of sources. People and corporations are producing data on a large scale, and since the rise of the internet, e-mail and social media the amount of produced data has grown exponentially. From a law enforcement perspective we have to deal with these huge amounts of data when a criminal investigation is launched against an individual or company. Relevant questions need to be answered like who committed the crime, who were involved, what happened and on what time, who were communicating and about what? Not only the amount of available data to investigate has increased enormously, but also the complexity of this data has increased. These communication patterns need to be combined with the objective to extract entities, relations and events. Furthermore, the information management processes within crime investigations are very complex and often delegated to the investigator's computer skills. Despite that, the application of natural language processing techniques based on crime data can prove to be beneficial in several processes of the criminal justice industry. Up to date, it is not feasible for the law enforcement agencies to get into the detail of these available massive crime reports and get the answers and furthermore is not available a system that permits to integrate information coming from text document with structured data. Starting from these considerations, we proposed a system to support the prosecutors to identify suspicious activities managing information of different kinds of format and sources.

*Corresponding author.
†These authors contributed equally.
valerio.bellandi@unimi.it (V. Bellandi); stefano.siccardi@unimi.it (S. Siccardi)
0000-0003-4473-6258 (V. Bellandi); 0000-0002-6477-3876 (S. Siccardi)

published 2026-03-13

As a result of the analyses carried out, we identified some key facts, and recognized that they apply to civil trials data too:

1. each case consists of data and a huge number of textual documents in natural language
2. documents must be annotated to highlight relevant entities and facilitate text navigation
3. entities must be disambiguated and uniquely identified in a so called *Entity Registry (EReg)*
4. a powerful and flexible retrieval and statistical system must be provided

Moreover, we realized that graph databases can be a very useful tool; however, as they have been used in particular to support investigations, here we will only touch on them briefly.

The idea of a "semantic document management system for PA," was described in [2]. Information systems of PAs are to evolve into platforms where unstructured data can be exploited and integrated with structured data to enhance and add value to the digital services provided by the PA; governance processes can be conducted using all knowledge expressed in documents and other forms of unstructured data. The judicial organization, proceedings, processes, user needs, functional structure of the Datalake, and implementation architecture have been described, aiming towards a design and production pathway directed at all PAs.

We described architectures to implement such functionalities in several paper, as we will discuss more in depth later: in [3] and [4] we defined the Entity Registry; the concept of service infrastructure to manage legal documents was introduced in [5] and [6]; graph applications to this context were described in [7] and applications to document exploration, retrieval and deriving statistics were considered in [8], [9], [10] and [11]. Being service based, each task can be performed using different techniques. In particular, performances of several algorithms have been compared in [4]; we also studied, however in a different application, possible interplay of rule based and machine learning algorithms in [12].

2. State of the Art

In the literature several systems have been proposed for **legal document management**. For instance [13] analyzes the state-of-the-art in software architecture for NLP, in the field of law documents. The authors identified several architectural approaches, including pipeline, service-oriented, and microservices architectures and reviewed the last important studies, that focus on the pipeline architecture. It consists of a sequence of NLP modules that process text in a predefined order. Moreover, the advantages of service-oriented and microservices architectures have been highlighted. However, they do not identify a generic infrastructure to manage entities in the legal domain. [14] describes a document management system that combines NLP and ontologies in the legal domain. It manages paper documents, automatically transforming them into RDF statements, for indexing, retrieval and long term preservation. [15] describes an architecture to analyze and extract a set of specific data from natural language documents; the system has later been improved in [16] with microservices and message brokers. It is based on an ontology containing information about the types of documents, their properties and sections, entities from each section and their relations. This implementation, that works on a real document management system which performs intensive processes, shares some characteristics with our design for districts. However, like in most of the architectures quoted above, entities do not play a central role like in the system we propose. [17] describes a specialist knowledge management system to semi-automate the extraction of norms to populate legal ontologies, by relying on general-purpose NLP modules, with pre and post-processing tasks based on domain knowledge rules. At a difference with the above works, an entity-centric architecture has been described in [18] to manage court judgements and other legal documents. In the present discussion we go a step further introducing the hierarchy concept.

Many other systems have been described to store and make available legal documents, however without specific focus on any analytical capabilities. For instance, [19] describe a framework to regulate the preparation, transfer, and use of electronic justice case files and [20] a system to manage various types of legal documents, their hierarchy and status, relations between legal documents and the parties involved.

From the **architectural point of view**, distributed systems have been studied from several angles across the ICT evolution, focusing on their design, development, and deployment, as well as the evaluation of their non-functional behavior. Recently, particular emphasis has been given to big data architectures, focusing on the design and implementation of big data systems, their deployment on the edge-cloud continuum, as well as the evaluation of their performance and scalability (e.g., [21, 22]). The impact of distributed systems on the safety, security, and privacy of humans has then been considered, with particular reference to system trustworthiness in terms of governance, risk, and compliance. Several assurance techniques [23] have been defined with the aim of proving a specific system behavior in terms of non-functional properties support. Today, certification is considered by policymakers, regulators, and industrial stakeholders as the most suitable assurance technique for the verification of non-functional properties (e.g., availability, confidentiality, privacy) of distributed systems [24].

On the other hand, many studies refer to **specific aspects of the legal document analysis and NLP applications**. A good overview can be found in [25], which emphasizes the role of Named Entity Recognition (NER) techniques and Relation Extraction (RE). Usage of ontologies and of widely used NLP models like BERT in the legal domain has been reported (e.g. [26] and [27]). NLP methods have been applied to support legal information extraction and retrieval (see e.g. [28]); contributions to the Competition on Legal Information Extraction/Entailment (COLIEE), organized since 2017 [29], describe several studies in this area. Question answering systems have been implemented, see [30] for a survey.

Finally, we quote some **machine learning related** aspects, even if this is not the main goal of our system. In this context, architectures where local systems store their data and cooperate to build general models are known under the heading of *federated learning*. For instance, in [31] an iterative framework has been described, that proceeds through periodic communication between the central server and local systems. However, many important issues typical of federated learning are of no relevance for the present study: for instance we do not deal with heterogeneous devices, networks and architectures as described in [32] and, on the contrary, concepts like entities are not considered in federated learning. In the last years many applications of LLMs to the legal domain have been described; we quote for instance [33] for a review, [34] and [35] for applications to the Italian civil code.

3. Functional architecture

The management of legal documents requires a robust and structured framework that ensures efficient handling of the entire document lifecycle. This includes ingestion, annotation, retrieval, and statistical analysis. The proposed architecture is designed to support the creation, storage, and exploration of legal texts while maintaining compliance with security and access control policies.

A key feature of this architecture is its ability to integrate structured and unstructured data, allowing legal professionals to navigate vast datasets efficiently. Documents are categorized based on case types and legal domains, facilitating easy access and organization. The system ensures secure storage through role-based permissions, preventing unauthorized access to sensitive legal information.

Starting from these general considerations and during these years of collaboration with the Italian Ministry of Justice, further and more specific key functions have been elicited which are necessary, in details we identified:

1. *Data Ingestion.*The system collects, standardizes, and indexes legal documents along with metadata such as case numbers, court decisions, parties involved, and legal references. This step ensures that data is entered in a structured format, allowing for seamless downstream processing.
2. *Pre-Processing and Indexing.* Before analysis, documents undergo pre-processing, including text normalization, tokenization, stop-word removal, and syntactic parsing. This step enhances the accuracy of document retrieval and annotation. Indexing mechanisms create structured representations that facilitate quick searches.
3. *Annotation services.* Through NLP and Named Entity Recognition (NER), the system automatically identifies and highlights critical entities such as people, organizations, legal terms, dates, and

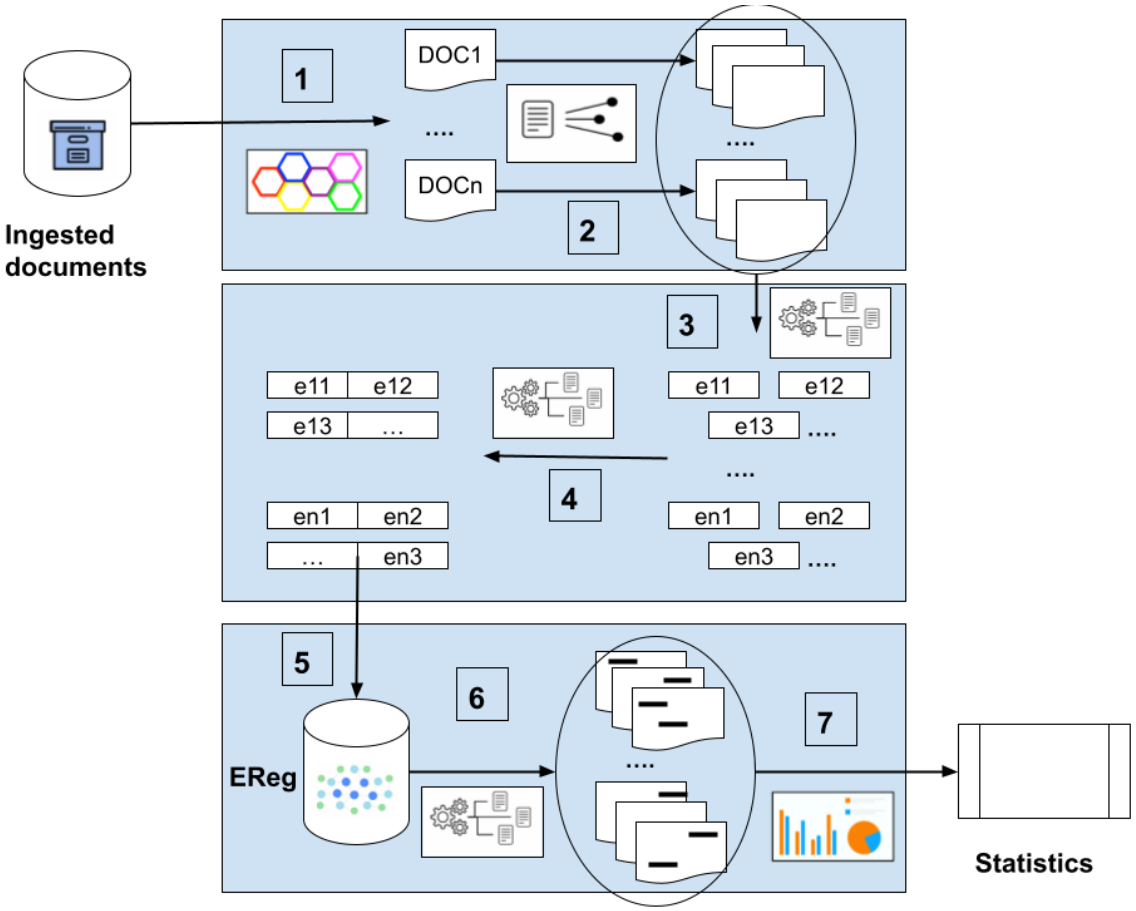

Figure 1: Illustration of the entity extraction pipeline. Arrows representing actions have been labeled with numbers for readability: 1. Data filtering using queries on the ingested documents; 2. Document tokenization to find sections; 3. Named Entity Recognition; 4. Named Entity disambiguation and grouping of their quotations; 5. EReg entries creation; 6. Insertion of EReg Ids in the annotations; 7. Creation of statistics.

case references. These annotations enhance document navigation and enable intelligent retrieval (see section 4).

4. *Entity Management and Disambiguation.* Extracted entities are mapped to a centralized Entity Registry (EReg) to ensure consistency across different legal documents. This process minimizes ambiguity, as multiple mentions of an entity (such as a judge's name or a company) are linked to a single, verified identity.
5. *Data Storage and Retrieval.* The architecture includes a scalable database system capable of handling large volumes of legal texts. It maintains original documents, extracted metadata, annotations, and indexes, enabling structured querying.
6. *Statistical Analysis and Reporting.* To support legal research and decision-making, the system provides statistical insights. Users can generate reports on case trends, citation patterns, case durations, and judicial decision patterns, which aid in policy-making and case strategy formulation.
7. *User Interface and Search Capabilities* A robust front-end interface enables legal professionals to retrieve, filter, and analyze documents using keyword searches, semantic queries, and entity-based exploration. The system also supports advanced search functionalities, such as similarity-based retrieval and contextual exploration of case laws.

Fig. 1 represents our pipeline proposed to identify entities, create annotations and prepare statistical reports.

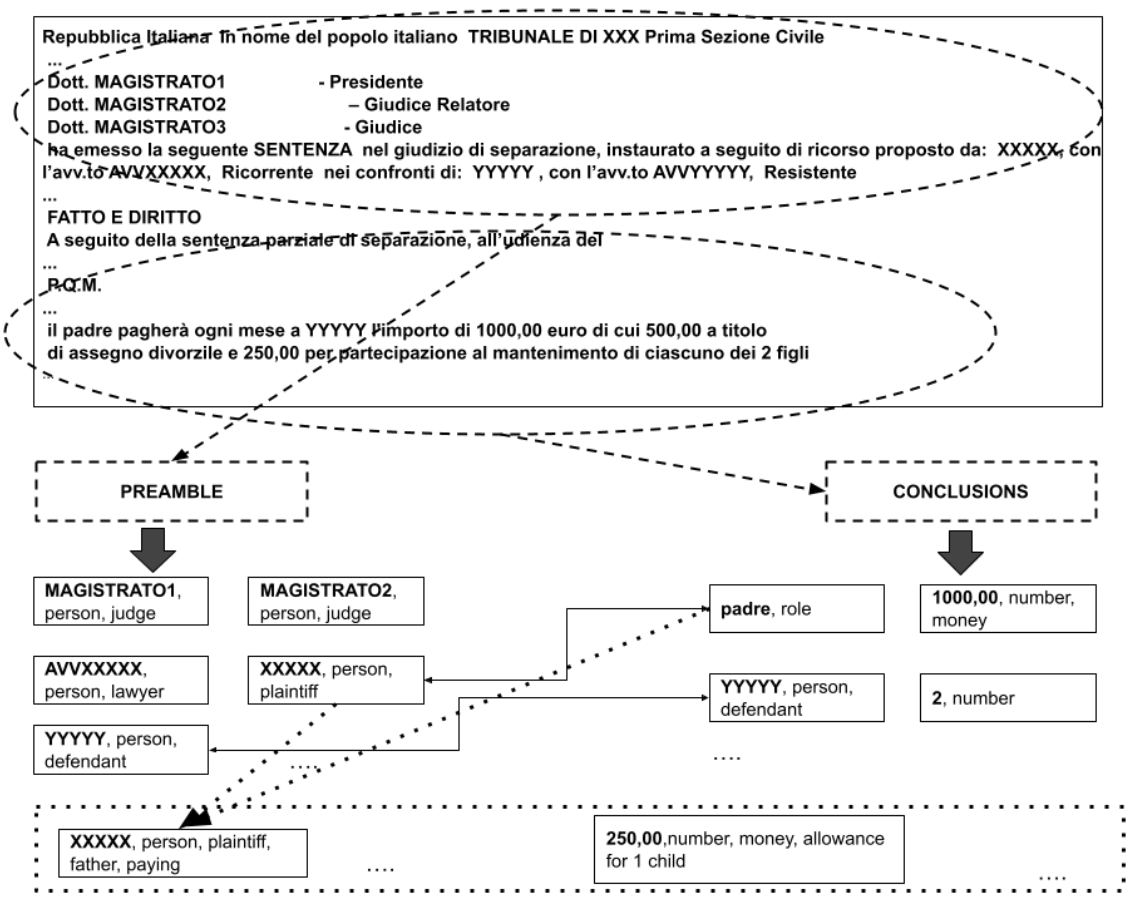

Figure 2: Example of annotations. Top: the document. Intermediate: sections and annotations. Bottom: entries in the EReg.

4. Documents and Annotations

Formally, a document is a textual item ingested in the data storage, like a law, a judgment, a court decision. It is defined as a pair $d = \langle M, T \rangle$, where M is a set of document metadata, and T is the textual document content. An item $m \in M$ is expressed as $m = (mn_i, mv_i)$, where mn_i is the name of the metadata item and mv_i is the corresponding value. For example, a metadata item can be the date of publication of a judgment document (i.e., $mn = publication\_date$). The text T of a document is represented in terms of *sections* and *chunks* that are the result of tokenization mechanisms of the document text T at different levels of granularity. A *section* $s \in T$ is defined as $s = (sn_i, sv_i, sc_i)$, where sn_i is the name of the section, sv_i is the corresponding content (the section text), and sc_i is a set of *chunks* in which the section content is segmented. For instance, considering a court decision, possible sections are the incipit (i.e., $sn = incipit$), the facts (i.e., $sn = fact$), and the conclusions (i.e., $sn = conclusion$).

Annotations are traditionally created by Named Entity Recognition (NER) services, that deal with the identification of entities featured by a name in the textual content of a document such as people, organizations, and locations. The overall output of the NER step is composed by a set of entities e and a set of annotations a, where each annotation a references the position where entity e occurs in the document d.

Fig. 2 illustrates a document, in the top part of the picture, that has been divided into sections (preamble, conclusions, etc.). Sections have been annotated locating mentions of entities like persons and more specifically judges (*magistrato*), plaintiff, defendant, father (*padre*) and so on (middle part of the picture). Finally, unique entries for entities are created in the EReg (bottom part of the picture).

In [4], we presented an entity-centric infrastructure to manage legal documents, especially court

judgments, based on the organization of a textual document repository and on the annotation of these documents to serve a variety of downstream tasks. Documents are pre-processed and then iteratively annotated using a set of NLP services that combine complementary approaches based on machine learning and syntactic rules. We present a framework that has been designed to be developed and maintained in a sustainable way, allowing for multiple services and uses of the annotated document repository and considering the scarcity of annotated data as an intrinsic challenge for its development. The design activity is the result of a cooperative project where a scientific team, institutional bodies, and companies appointed to implement the final system are involved in co-design activities. We describe experiments to demonstrate the feasibility of the solution and discuss the main challenges to scaling the system at a national level. In particular, we report the results we obtained in annotating data with different low-resource methods and with solutions designed to combine these approaches in a meaningful way. An essential aspect of the proposed solution is a human-in-the-loop approach to control the output of the annotation algorithms in agreement with the organizational processes in place in Italian courts. Based on these results we advocate for the feasibility of the proposed approach and discuss the challenges that must be addressed to ensure the scalability and robustness of the proposed solution.

We stress that documents can be annotated not only with concrete entities mentions, but also with abstract entities like concepts.

5. The Entity Registry

Usually annotations created by NER systems contain only generic entity specifications, like *person* or *location* and so on. On the contrary, we need to relate each mention to a well identified entity, that is a specified person or location and so on. We therefore introduced the Entity Registry

5.1. Definition

In [3] we defined a conceptual structure for a repository of entities. These can be found by usual procedures of Natural Language Processing, like the search for entities mentioned in text, their identification, possibly through the link to entries in Background Knowledge Basis (BKG) and the construction of a Knowledge Basis or Graph to host the information found in this process. Such approach is especially important for applications where a BKG is of little help, because the involved entities are not so relevant to be included in any, being for instance ordinary people or small companies. Therefore, we rely on the entities' attributes and relationships for unique identification, disambiguation, knowledge checking and any other relevant operation. One of the final goals achieved by the proposed method is the ability to merge knowledge collected in separate bases, once they are referred to the same Entity Registry.

In the EReg, an entity e is defined as a pair $e = \langle E, F \rangle$, where E is an entity type, and F is a set of features with related values that qualify the entity e; F depends on the type E. An entity type E is defined as $E = (en, ef, ek)$, where en is the entity name, ef is a set of features that characterize the entity type, and ek is a set of *keys*. Keys are combinations of features within ef whose values uniquely identify the entity. In general, several sets of features can denote the same entity. For instance, if $en = person$, with featurees ssn, firstname, lastname, birthdate, and birthplace, two keys can be defined: the ssn and the set of the other four features. For an entity e, a feature $f \in F$ is defined as $f = (fn, fv)$, where $fn \in ef$ is a feature name, and fv is the corresponding value.

5.2. Distributed EReg

In practical applications in the legal area, as many different courts and districts exist, data and documents are distributed and a distributed system must be considered. In [8] we presented a real-world application scenario involving a distributed repository of documents and metadata. The application entails a network of edge document repositories analyzing textual content and metadata to extract entities, thereby offering enhanced semantic exploration functionalities for the Italian Ministry of Justice. The

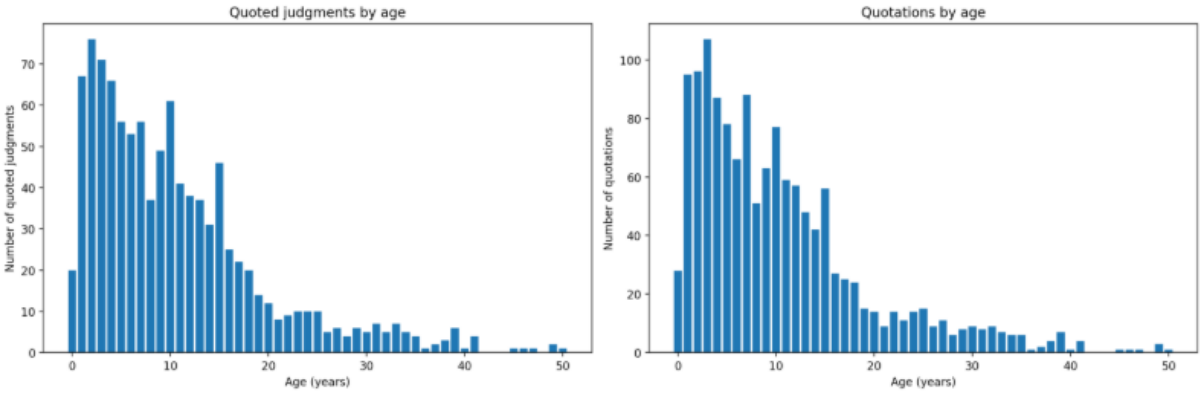

Figure 3: Age of Supreme Court sentences quoted in divorce allowance judgments. Left panel: age distribution of sentences, right panel: age distribution of quotations.

system has a hierarchical architecture, whose goal is to make information about entities available at all districts, according to user access rights. It is provided with a top level EReg, and the structure of both the top level and local ERegs is implemented in order to store i) all the entities' attributes ii) the entities' relationships iii) ids of entities in local ERegs. In other terms, the top level EReg collects available information about entities, helping users to merge or disambiguate entities. The hierarchy can be further implemented to consider three levels: courts, districts and the top level, each with specific services and user roles.

6. Document Retrieval and Statistics

Based on the annotations and the EReg described above, a number of services can be provided to users.

In [9] we presented a knowledge-based approach for legal document retrieval based on the organization of a textual data repository and on document embedding models. Pre-processed and embedded documents are iteratively classified at sentence level through a terminology extraction and concept formation cycle, using a zero-knowledge approach that offers a high degree of flexibility with regard to the integration of external knowledge and the variability of inputs, suitable to face the scarcity of annotated data and the specificity of terminology that feature the Italian legal domain document corpora.

Concept-driven data exploration was described in [10] and [5] considering a case-study about "unfair competition" as subject matter and invoking a knowledge extraction pipeline with the aim to explore the concepts extracted from the corpus on such a subject. The user can enforce a preliminary filtering step over the document metadata to select the set of court decisions to consider for concept exploration. A concept graph is returned by the knowledge extraction pipeline for describing the filtered dataset on unfair competition. Most of the graph concepts pertain to the domain of trade justice (e.g., "consortium", "partnership", "transaction"), by also describing specific aspects concerned with unfair competition. Through links, it is possible to move from specific concepts (e.g., "sponsorship") to more general ones (e.g., "business"), and vice-versa.

Statistics based on entities have been described in several papers. In [6] and [5] we described a statistics of plaintiff gender in divorce trials; in [11] we considered ages of people involved in employment related cases, taking into account also the missing data problem.

More recently, we computed statistics of Supreme Court citations in first and second degree judgments and of outcomes of second degree trials. As an example, fig. 3 shows the age distribution of Supreme Court decisions quoted in divorce allowance judgments; ages have been computed as the difference between the year of the judgment and the year of the quoted sentence.

We note that such statistics can be combined together and with operational data to search for patterns and correlations. For instance we explored correlations between the number of Supreme Court citations

(that may represent how much the subjects has been debated and clarified) and: 1) case durations for first and second degree trials, 2) confirmation or reformation of first degree judgments in second degree.

Among services provided to users, we quote also the anonymization, that can be performed on the fly, and is related to the user rights to access specific entities.

Finally, we quote the possibility to extract graphs, describing both relationships between entities and events involving entities. Such graphs have been extensively used to support investigations (see e.g. [7])

7. Conclusion and future work

This paper presented a structured approach to managing legal data through a framework that integrates Natural Language Processing (NLP), Named Entity Recognition (NER), and an Entity Registry (EReg). The system aims to enhance document annotation, retrieval, and statistical analysis, particularly in the legal domain. By leveraging both machine learning and rule-based methods, the proposed approach enables more efficient organization, exploration, and analysis of legal documents.

The framework addresses critical challenges in legal data management, such as entity disambiguation, structured data integration, and large-scale document retrieval. The use of graph databases and distributed architectures further enhances the system's ability to process complex legal datasets efficiently. Additionally, statistical reporting functions provide valuable insights into legal trends, supporting informed decision-making for legal professionals.

One of the key contributions of this work is the development of a scalable and distributed entity registry that allows for a more precise and context-aware retrieval of legal information. This ensures that legal entities, whether individuals, organizations, or case-related terms, are accurately linked and retrievable across various legal documents and jurisdictions.

Future research should focus on refining NLP techniques, improving automation in legal text analysis, and expanding the system's applicability beyond Italian judicial cases. Overall, this study demonstrates that integrating advanced data management techniques into legal workflows significantly enhances document accessibility, accuracy, and efficiency, paving the way for more intelligent and scalable legal information systems.

Acknowledgment

This work is partially supported by i) the Next Generation UPP project within the PON programme of the Italian Ministry of Justice, ii) the Università degli Studi di Milano within the program "Piano di sostegno alla ricerca", iii) the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strenghtening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D, and iv) the project SERICS (PE00000014) under the MUR NRRP funded by the EU - NextGenerationEU.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Batini, V. Bellandi, P. Ceravolo, F. Moiraghi, M. Palmonari, S. Siccardi, Semantic data integration for investigations: Lessons learned and open challenges, in: 2021 IEEE International Conference on Smart Data Services (SMDS), 2021, pp. 173–183. doi:10.1109/SMDS53860.2021.00031.

[2] C. Batini, G. Santucci, M. Palmonari, V. Bellandi, E. Fersini, B. Pernici, F. Zanzotto, G. Vecchi, S. Ronchi, Towards a semantic document management system for public administration, in: Proceedings of the Ital-IA Intelligenza Artificiale - Thematic Workshops co-located with the 4th CINI National Lab AIIS Conference on Artificial Intelligence (Ital-IA 2024), volume 3762, 2024, p. 360 – 365. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205566532&partnerID=40&md5=a5d11de53eb680f9853d7b1a34041956.

[3] V. Bellandi, S. Siccardi, An entity registry: A model for a repository of entities found in a document set, in: NIAI, MoWiN, AIAP, SIGML, CNSA, ICCIoT - 2023, AIRCC Publishing Corporation, 2023, p. 1–12.

[4] V. Bellandi, C. Bernasconi, F. Lodi, M. Palmonari, R. Pozzi, M. Ripamonti, S. Siccardi, An Entity-centric Approach to Manage Court Judgments based on Natural Language Processing, Computer Law & Security Review 52 (2024). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175050823&doi=10.1016%2fj.clsr.2023.105904&partnerID=40&md5=4a095f2a44c8d584960b66eeb4906eb1. doi:10.1016/j.clsr.2023.105904.

[5] V. Bellandi, S. Castano, S. Montanelli, D. Riva, S. Siccardi, A service infrastructure for the italian digital justice, in: R. Chbeir, D. Benslimane, M. Zervakis, Y. Manolopoulos, N. T. Ngyuen, J. Tekli (Eds.), Management of Digital EcoSystems, Springer Nature Switzerland, Cham, 2024, pp. 179–192.

[6] V. Bellandi, S. Castano, A. Ferrara, S. Montanelli, D. Riva, S. Siccardi, A service infrastructure for management of legal documents, in: CEUR WORKSHOP PROCEEDINGS, volume 3606, CEUR-WS, 2023, pp. 1–8.

[7] V. Bellandi, P. Ceravolo, S. Maghool, S. Siccardi, Graph embeddings in criminal investigation: Extending the scope of enquiry protocols, in: Proceedings of the 12th International Conference on Management of Digital EcoSystems, MEDES '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 64–71. URL: https://doi.org/10.1145/3415958.3433102. doi:10.1145/3415958.3433102.

[8] V. Bellandi, P. Ceravolo, S. Siccardi, Enhancing semantic exploration: A distributed repository approach, in: Proceedings of the International Workshop on Big Data in Emergent Distributed Environments, BiDEDE '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3663741.3664792. doi:10.1145/3663741.3664792.

[9] V. Bellandi, S. Castano, P. Ceravolo, E. Damiani, A. Ferrara, S. Montanelli, S. Picascia, A. Polimeno, D. Riva, Knowledge-based legal document retrieval: A case study on italian civil court decisions, in: CEUR Workshop Proceedings, volume 3256, 2022. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142439411&partnerID=40&md5=d71543f8c88985f21f0a44f6bc924d02.

[10] V. Bellandi, S. Castano, S. Montanelli, D. Riva, A service architecture for ai-based legal knowledge extraction, in: CEUR WORKSHOP PROCEEDINGS, volume 3478, 2023, p. 110 – 119.

[11] V. Bellandi, S. Maghool, S. Siccardi, An nlp-based statistical reporting methodology applied to court decisions, in: 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2023, pp. 108–111. doi:10.1109/SEAA60479.2023.00025.

[12] V. Bellandi, S. Siccardi, Gender discriminatory language identification with an hybrid algorithm based on syntactic rules and machine learning., in: SEBD, 2022, pp. 578–585.

[13] Z. Pauzi, A. Capiluppi, Applications of natural language processing in software traceability: A systematic mapping study, Journal of Systems and Software 198 (2023) 111616. URL: https://www.sciencedirect.com/science/article/pii/S0164121223000110. doi:https://doi.org/10.1016/j.jss.2023.111616.

[14] F. Amato, A. Mazzeo, A. Penta, A. Picariello, Using nlp and ontologies for notary document management systems, in: Proceedings of the 2008 19th International Conference on Database and Expert Systems Application, DEXA '08, IEEE Computer Society, USA, 2008, p. 67–71. URL: https://doi.org/10.1109/DEXA.2008.86. doi:10.1109/DEXA.2008.86.

[15] M. G. Buey, A. L. Garrido, C. Bobed, S. Ilarri, The ais project: Boosting information extraction from legal documents by using ontologies, in: Proceedings of the 8th International Conference on Agents and Artificial Intelligence, ICAART 2016, SCITEPRESS - Science and Technology Publications, Lda, Setubal, PRT, 2016, p. 438–445. URL: https://doi.org/10.5220/0005757204380445.

doi:10.5220/0005757204380445.
[16] M. D. Ruiz, C. Román, A. L. Garrido, E. Mena, uais: An experience of increasing performance of nlp information extraction tasks from legal documents in an electronic document management system, in: International Conference on Enterprise Information Systems, 2020. URL: https://api.semanticscholar.org/CorpusID:219322285.
[17] L. Humphreys, G. Boella, L. van der Torre, L. Robaldo, L. D. Caro, S. Ghanavati, R. Muthuri, Populating legal ontologies using semantic role labeling, Artificial Intelligence and Law 29 (2021) 171–211.
[18] V. Bellandi, C. Bernasconi, F. Lodi, M. Palmonari, R. Pozzi, M. Ripamonti, S. Siccardi, An entity-centric approach to manage court judgments based on natural language processing, Computer Law & Security Review 52 (2024) 105904. URL: https://www.sciencedirect.com/science/article/pii/S0267364923001140. doi:https://doi.org/10.1016/j.clsr.2023.105904.
[19] H. Qin, L. Chen, L. Mou, The development of china's electronic case file regulations and its future implications, Computer Law & Security Review 52 (2024) 105930. URL: https://www.sciencedirect.com/science/article/pii/S0267364923001401. doi:https://doi.org/10.1016/j.clsr.2023.105930.
[20] D. Ridwandono, M. I. Afandi, E. D. Wahyuni, E. Simaremare, I. Sinaga, Legal documents repository systems, Nusantara Science and Technology Proceedings 2023 (2023) 477–481. URL: https://nstproceeding.com/index.php/nuscientech/article/view/983. doi:10.11594/nstp.2023.3377.
[21] J. Dongarra, B. Tourancheau, D. Balouek-Thomert, E. G. Renart, A. R. Zamani, A. Simonet, M. Parashar, Towards a computing continuum: Enabling edge-to-cloud integration for data-driven workflows, Int. J. High Perform. Comput. Appl. 33 (2019) 1159–1174. URL: https://doi.org/10.1177/1094342019877383. doi:10.1177/1094342019877383.
[22] J. C. S. Dos Anjos, K. J. Matteussi, P. R. R. De Souza, G. J. A. Grabher, G. A. Borges, J. L. V. Barbosa, G. V. González, V. R. Q. Leithardt, C. F. R. Geyer, Data processing model to perform big data analytics in hybrid infrastructures, IEEE Access 8 (2020) 170281–170294. doi:10.1109/ACCESS.2020.3023344.
[23] C. Ardagna, R. Asal, E. Damiani, Q. Vu, From Security to Assurance in the Cloud: A Survey, ACM CSUR 48 (2015).
[24] C. A. Ardagna, N. Bena, Non-functional certification of modern distributed systems: A research manifesto, in: Proc. of IEEE SSE 2023, Chicago, IL, USA, 2023.
[25] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How does NLP benefit legal system: A summary of legal artificial intelligence, 2020.
[26] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020.
[27] D. Licari, G. Comandè, Italian-legal-bert models for improving natural language processing tasks in the italian legal domain, Computer Law & Security Review 52 (2024) 105908. URL: https://www.sciencedirect.com/science/article/pii/S0267364923001188. doi:https://doi.org/10.1016/j.clsr.2023.105908.
[28] S. Castano, M. Falduti, A. Ferrara, S. Montanelli, A knowledge-centered framework for exploration and retrieval of legal documents, Information Systems 106 (2022).
[29] J. Rabelo, R. Goebel, M.-Y. e. a. Kim, Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021, The Review of Socionetwork Strategies 16 (2022).
[30] J. Martinez-Gil, A survey on legal question–answering systems, Computer Science Review 48 (2023) 100552. URL: https://www.sciencedirect.com/science/article/pii/S1574013723000199. doi:https://doi.org/10.1016/j.cosrev.2023.100552.
[31] J. Oh, D. Lee, T. Ha, Y. Jeon, W. Noh, S. Cho, Federated flowchart: Overview of state-of-the-arts based on federated learning process, in: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), 2022, pp. 1076–1081. doi:10.1109/ICTC55196.2022.9952536.
[32] M. Ye, X. Fang, B. Du, P. C. Yuen, D. Tao, Heterogeneous federated learning: State-of-the-art

and research challenges, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3625558. doi:10.1145/3625558.

[33] M. Siino, M. Falco, D. Croce, P. Rosso, Exploring llms applications in law: A literature review on current legal nlp approaches, IEEE Access 13 (2025) 18253–18276. doi:10.1109/ACCESS.2025.3533217.

[34] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code, Artificial Intelligence and Law 30 (2022) 417–473. doi:10.1007/s10506-021-09301-8.

[35] A. Simeri, A. Tagarelli, Exploring domain and task adaptation of lamberta models for article retrieval on the italian civil code, in: 19th IRCDL Conference on Information and Research science Connecting to Digital and Library science, CEUR Workshop Proceedings, Bari, Italy, 2023. URL: http://ceur-ws.org/Vol-3365/paper4.pdf. doi:N/A, use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).