Towards a Semantic Document Management System for Public Administration

Carlo Batini[1,*,†], Gaetano Santucci[1], Matteo Palmonari[3,*], Valerio Bellandi[2,*], Elisabetta Fersini[3], Barbara Pernici[5], Fabio Zanzotto[4], Giancarlo Vecchi[5] and Stefano Ronchi[5]

[1]Consorzio Interuniversitario Nazionale di Informatica (CINI), Italy
[2]Università degli Studi di Milano, Italy
[3]University of Milan-Bicocca, Italy
[4]University of Rome Tor Vergata, Italy
[5]Polytechnic University of Milan, Italy

Abstract

To deliver services to users, central and local Public Administrations (PA) make extensive use of data. Various qualitative estimates suggest that databases contain 10-20

This work has two objectives: to summarize the experiences carried out over the past four years by the National Interuniversity Consortium for Informatics (CINI) in the Datalake project funded by the CRUI in collaboration with the Directorate General of Automated Information Systems (DGSIA) of the Ministry of Justice, in synergy with other related projects of the Ministry; and to demonstrate how the experiences, Proof of Concepts, and functional specifications produced can serve as a repository of functionalities for a "semantic document management system for PA," which aims to evolve the information systems of PAs 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, and where 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 are described, aiming towards a design and production pathway directed at all PAs.

Keywords

Semantic Document Management, Data Lake, Legal AI, Civil Trials, Criminal Trials

1. Proceedings, Trials, Organization, Justice Information Systems

The Ministry of Justice performs administrative functions in both the civil and criminal fields. The judiciary is a complex of structures and institutions aimed at the administration of justice, overseen by individual judges. The primary activities of the Ministry of Justice and the judges (collectively referred to as Justice) concern criminal and civil proceedings. The criminal proceeding includes preliminary investigations, activities of cognition in the three levels of judgment in the criminal process, and the execution of penalties or alternative activities in juvenile and community justice. Similarly, the civil proceeding consists of a cognition phase, which includes three levels of judgment, and an execution phase. The cognition phase of the civil proceeding has long been subject to automation within the On-Line Civil Trial (in Italian abbreviated as PCT) information system. Consequently, the digitization of structured data and documents in the civil proceedings files is significantly more advanced than in the preliminary investigations and criminal proceedings files. The digital file of a civil proceeding consists of acts and documents, and, for concluded proceedings, the judgment. An act of the civil proceeding is a documentary artifact related to a file, whose content and form are prescribed by regulations. A document is any artifact (text, audio recording, image, video, etc.) related to the file and attached to acts. The progress of the civil proceeding is represented in terms of states and events. In the first phase of the civil proceeding and, to a greater extent, in the preliminary investigative phase of the criminal proceeding, numerous documentary sources of evidence are acquired, including telephone records, credit card traces, inspections, transcribed telephone interceptions, and many others. The primary activities of preliminary investigations and cognition of the criminal proceedings have only recently become the subject of automation. The execution phase of criminal proceedings is characterized by greater automation compared to civil proceedings, with the Judiciary Record (in Italian "Casellario Giudiziale") and databases of the Department

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding author.

† ChatGPT was used to translate original content written by the authors in Italian; the authors have read and revised the translation, ultimately agreeing on the final content.

carlo.batini@unimib.it (C. Batini); matteo.palmonari@unimib.it (M. Palmonari); valerio.bellandi@unimi.it (V. Bellandi)

of Penitentiary Administration and the Department of Juvenile and Community Justice being the main realizations. The Datalake project was initiated in 2019 by the DGSIA, which entrusted the CRUI and, subsequently, the Consorzio Interuniversitario Nazionale per l'Informatica (CINI), with a renewed line of research over the years, investigating the adoption of technologies based on natural language processing (NLP), knowledge graphs, machine learning, and, recently, generative AI, which can be most useful in carrying out the primary processes of civil and criminal cognition and execution. In 2022, Justice included Datalake among the projects funded with PNRR funds, whose implementation was entrusted through a tender to a group of three companies: Almaviva, Almawave, and Accenture.

2. User Needs in Civil and Criminal Trials

The Datalake project focused on the preliminary investigation phase of criminal proceedings, criminal enforcement, and the cognition phase of civil proceedings.

The functionalities developed originate from a user requirements elicitation activity, which for the criminal procedure involves Prosecutors and the Judicial Police, and for the civil procedure involves Judges. The requirements were collected in the preliminary investigation phase, through the sharing of Proof of Concepts on defined procedures, and subsequently through consultations on ongoing procedures. In the civil domain, interviews were conducted with Judges of the Court of Appeal of Milan, who will soon experiment with a first set of functionalities developed by the supplier Almawave. The outcome of the experiment will result in a new version of the system that can be adopted in all Courts of Appeal. The user needs are briefly described below.

Primary Activities – Preliminary Investigations in Criminal Proceedings

- Specific searches and semantic aggregations for the discovery and confirmation of clues and evidence.
- Integrated analysis of relational knowledge with visualization.
- Selection of node clusters in a semantic graph with certain properties.
- Reconstruction of relationships maintained by suspected individuals, composing the entire relational network of the suspect.
- Transcriptions and semantic enrichment of audio messages.

Primary Activities – Civil Proceedings

- Enrichment and exploration of legal knowledge during the process.
- Semantic search for nominal entities, mentions, concepts, terms, phrases, and simple sentences, to analyze seriality and search for precedent cases.
- Search for relevant judgments and case law with a single integrated access point.
- Selection of judgments concerning topics (e.g., damages for privacy violations) not included in the system metadata.
- Decision support in the cognition phase of the judgment and linking relevant acts and documents with the judgment.
- Extraction from civil judgments of the outcome of the process (e.g., damage compensation, maintenance allowance in judicial divorce proceedings) and correlated salient features.

Governance Activities in the PCT (On-line Civil Trial)

- Predictive models of the expected durations and variability of the proceedings based on their characteristics.
- Assessment of the expected complexity of the proceedings for the distribution of workload among judges, identification of "bottlenecks", identification of signals and events that significantly impact the duration of the proceedings, and analysis of the impact of changes in laws, regulations, and practices.
- Descriptive statistics on structured data and judgments.
- Correlation analysis between salient characteristics and outcomes in civil proceedings for uniformity purposes (so-called "tabulation").
- Comparative analysis of trial durations in different sections and districts.

3. Functionalities

The Proof of Concept developed and the functional specifications produced within the Datalake project concern the following macro-functionalities: Preparation, Semantic Enrichment and Knowledge Integration, Semantic Search and Analysis, Knowledge Base Management, and Quality Control. The following are the detailed functionalities.

F1 - Preparation

1. Document pre-processing (removal of special characters, correction of accented letters, removal

of headers, removal of stamps, punctuation management)
2. OCR and generation of interpretable documents.
3. Identification of sections of the judgments: preamble, case description, and decision.
4. Classification of texts within the files.

F2 - Semantic Enrichment and Knowledge Integration

Semantic enrichment is performed by extracting information from documents, especially named entities and terms, and persisting the result of this extraction process into semantic annotations. This process is in use in legal AI to a large extent. The peculiar characteristic of the proposed approach lies in the effort to consolidate the knowledge extracted by linking different mentions that refer to the same entities (exploiting background knowledge bases like Wikipedia and clustering mentions of the entities - of course, the majority - that are not present in Wikipedia) [1, 2, 3]. The impact of this approach is particularly noticeable during document search (see functionality F3).

Civil Trials. Various NLP techniques have been applied to extract, link, and consolidate entity mentions from judgments and produce semantic annotations that associate the extracted entities with specific token sequences in the judgments. In particular, the current pipeline combines the following techniques [1, 2]:

- *Named Entity Recognition (NER):* Utilizes rule-based and neural approaches, tuned to the data distribution in the domain (sequential classifiers on features from a BERT-based encoding transformer).
- *Named Entity Linking (NEL):* Based on the BLINK entity retrieval algorithm trained on the Italian Wikipedia within the project.
- *NIL Prediction:* Decides whether to link an entity mention to the entity associated with it by NEL or label it as a new entity not present in the knowledge base (NIL); for this task, an internal classifier based on features is used. To perform NEL and NIL prediction at once, an extended named entity disambiguation algorithm has also recently been explored to predict NIL as a class.
- *NIL Clustering:* Groups entity mentions referring to the same real-world entities (typically applied to mentions labeled as NIL because entities linked to a knowledge base are implicitly grouped).
- *Entity Registry Construction:* The Entity Registry is a component where each entity, enriched with attributes deduced during the linking phase, corresponds to a unique entry, avoiding duplicates and disambiguating homonyms and synonyms. Text annotations are updated with entity identifiers.
- *Refinement of Decisions:* Final decisions made at the end of the pipeline are refined based on some domain-specific rules (especially for the classification of specific and fine-grained entities).
- *Relation Extraction:* Extraction of relationships such as victim-offender relationship, or based on the expression "against". We used pre-trained transformer models for text representation, with training conducted according to a cross-validation policy and an extraction model based on the entity-relationship paradigm and REBEL [4].
- *Features & values Extraction:* Aims to extract values associated with features (e.g., the economic value of maintenance payments). Two available open-source models, Camoscio and Stambecco (versions of LLAMA trained on the English language and adapted for the Italian language), and the pay-per-use model known as ChatGPT were considered. Techniques based on prompt engineering were experimented with, using the following types of prompts: Direct Instruction Prompts, Contextual Prompts, Bridging Prompts, Socratic Prompts.
- Few-shot Fine-grained Entity Typing: Assignment of specific types from taxonomies to entity mentions. We used a neuro-symbolic method, where the taxonomy is explicitly modeled, and a method based on LLM with implicit prompts.

Criminal Trials - Preliminary Investigations. For documents related to preliminary investigations, a very similar pipeline was applied for entity extraction and subsequent document annotation, a similar semantic search paradigm. A first discussion of the application of entity-centric approaches to manage documents in preliminary investigations can be found in [5]. However, other functionalities and techniques were applied such as:

- Extraction of graph representations from instant messaging applications (IMA) data, e.g., WhatsApp dumps, and storage in a graph DB (Neo4J); messages can be queried using a structured language that supports graph-based data analysis.
- Content enrichment with speech-to-text technology; OpenAI's Whisper was used to transcribe audio messages and make these contents searchable. All messages and chats are analyzed using small adaptations of the NLP pipeline described earlier, supporting semantic search powered by entity-based annotations.
- Semantic enrichment and specialization of entity annotation ontologies relative to specific taxonomy (is-overlapping, is-within, ordering).

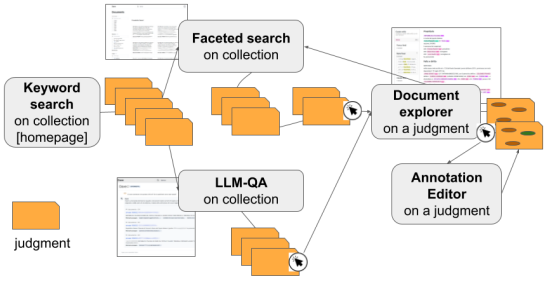

Figure 1: Architecture of the semantic search interface

Other developed functionalities include domain concept extraction, text summarization, and georeferencing of spatial entities. For all functionalities, accuracy analyses were conducted based on scientific methodologies. For the entity extraction pipeline, some results are reported in [1, 2]. As examples of accuracy measured for relation and feature extraction capability, we report accuracy for the Relationship "against", 83.5%, and for the extraction of the maintenance payment in favor of children in separation cases, 77.52%.

F3 - Semantic Search and Data Analysis

Common Search Functionalities for Preliminary Investigations and Civil Proceedings. Search functionalities are inspired by the well-known faceted and semantic search paradigms, with additional and more experimental Question Answering (QA) capabilities based on the Retrieval Augmented Generation (RAG) paradigm. Based on the semantic enrichment functionalities shown in the previous point, the entities that appear in the filters during the search phase can refer to mentions present in different documents; moreover, when a user explores, for example, a judgment, they can find all mentions of an entity throughout the document, a feature that can become particularly relevant for long judgments or other documents. The conceptual architecture for semantic search is shown in Fig 1. The components are:

- *Keyword search:* Allows simple keyword searches. This module can be useful as a starting point for the search, before activating the faceted search.
- *Faceted search:* Combines keyword searches and filters based on the attributes of the judgments. The module uses known technologies for indexing and querying document databases (e.g., Elasticsearch).
- *LLM-QA:* Implements a conversational search based on the RAG paradigm. A generative LLM manages the interaction with the user and the generation of responses; a neural retrieval module allows indexing chunks of judgments using embeddings and retrieving the relevant ones for a user's question.
- *Document explorer:* Allows exploring a document, such as a judgment, guiding the search within it for specific entities or mentioned concepts.
- *Annotation editor:* Allows modifying annotations to support a supervised annotation process where users can correct wrong or imprecise annotations and add new annotations.
- *Concept search:* Allows searching or exploring concepts according to domain logic. This module can be useful to help the user select specific concepts of interest in an exploratory or search refinement phase.

The above functionalities have all been demonstrated using DAVE, a prototype open-source application for semantic search developed in the context of this and the PON Next Generation UPP [1] project. A video demonstrating the proposed combination of semantic and conversational search on judgments of criminal trials published online is available at https://www.youtube.com/watch?v=XG7RsI3t-2Q. However, the data enrichment process developed in the project supports also other forms of search, such as *Advanced search.* This functionally supports advanced searches by combining various filters on document attributes. This module is included in many search applications on structured or semi-structured data, to complement the modules based on Keyword search and Faceted search; typically, the function of this module is to construct precise queries based on structured descriptions of documents.

Analysis Functionalities for Governance Activities - Civil. The semantic organization of documents obtained through semantic enrichment and integration functionalities enabled by the Entity registry allows for multiple statistics and correlations on structured data linked to annotated documents, e.g., the number of documents involving natural legal entities, the number and average value of minors involved in divorce decisions, correlation for tabulation purposes of the compensation value and related features in non-pecuniary damage cases. Further analyses concern survival curves of processes and explanatory variables of temporal duration and process complexity. Several analysis functionalities were developed within the PON Next Generation UPP project and other CRUI-funded projects. The following research based on the SICID system registers for the PCT was conducted (see [6, 7]):

- *Variant Analysis:* Clusters of proceedings with the same structure and sequence of states and their evaluation for monitoring purposes. In particular, the factors that have the greatest impact on the

[1] https://www.nextgenerationupp.unito.it/

duration of the processes were analyzed. For this activity, the process mining tool Apromore[2] was used.

- *Identification of Critical Events:* The impact of specific events on the duration of a process execution is evaluated to identify events systematically associated with anomalous situations. Both the phases and the total duration of the proceedings were examined.
- *Predictive Approaches for Alerts:* Predictors were constructed from sequences of states or events in the registers, based on machine learning techniques with LSTM neural networks, to predict the residual duration of processes and states during their course.

A management control dashboard was created for the Court of Cassation. The adopted solution was to create a dashboard directly fed by the underlying database of the Court's SIC register, with data updated four times a day. All data were identified for:

- Feeding the variables and indicators identified as necessary to describe the file path in the various phases and to calculate indices such as the Disposition Time and the turnover index;
- Building the historical series of such data from January 2019.

Analysis Functionalities for Preliminary Investigations. Relational knowledge analysis with visualization (e.g., selection of clusters of nodes with certain properties) and anomaly detection.

Functionalities for Penal Execution. Integration for the social analysis of data relating to liberty restrictions/alternative penalties experienced by detainees during their lives.

F4 - Knowledge Base Management and Quality Control - Main Methodologies and Developed Functionalities

1. *Manual of Pseudonymization Trial Policies:* Different types of pseudonymization are considered, and various types of data and document processing where pseudonymization is relevant (e.g., publication, linking databases, etc.) are identified, along with the properties that must be respected in each case. A general method is provided that can be followed for the different types of data processing relevant to the Datalake project.
2. *Entity Registry Management:* Includes creation, updating, deletion of entities, merging, and splitting of entities.
3. *Extraction of Lexicons:* Involves extracting lexicons of terms based on noun phrases from judgments and organizing them into an ontology, with specialization of the lexicon in the legal field (fine-tuning).
4. *Quality Assessment of NER and NEL:* Evaluation of the quality of Named Entity Recognition (NER) and Named Entity Linking (NEL) [1, 2].
5. *Benchmarking Extraction Models:* Benchmarking extraction models against various levels of taxonomy depth, and annotation tools among different relationship extraction models.
6. *Introduction of Guardrails:* Implementing guardrails to prevent errors or unprocessable judgments.
7. *Quality Manual for Data, Documents, and Diagnostic and Predictive Models:* Covers aspects such as accuracy, completeness, currency, fairness, and explainability (see [8]). For accuracy and fairness, the manual aligns with policy documents issued by the EU (see [9]).

4. Ontologies/Taxonomies and Their Top-Down and Bottom-Up Generation

In the functionalities of the Datalake, the following ontologies are used:

- **Top Ontology of Justice Procedures (cognition and execution):** Consists of about 400 classes, represented through approximately 40 schemas in the Entity-Relationship model at different levels of integration/abstraction.
- **Ontology for Penal Execution:** Consists of about 100 classes and 8 schemas in the Entity-Relationship model, including all the databases related to penal execution.

The following additional ontologies are represented in the form of two-level taxonomies: i) Top ontology of preliminary investigations, ii) Top ontology of the civil trial, iii) Domain ontologies of the civil process: banking, labor, non-patrimonial damage from privacy violation, judicial separation, iv) Ontology for penal-cognition procedure: victim-perpetrator relationship. The top ontology of Justice procedures and penal execution were produced through reverse engineering from logical schemas. The ontologies for the victim-perpetrator relationship and non-patrimonial damage were produced by domain experts. The ontologies for banking and labor were produced from lexicons built through the analysis of judgments.

[2] https://apromore.com/

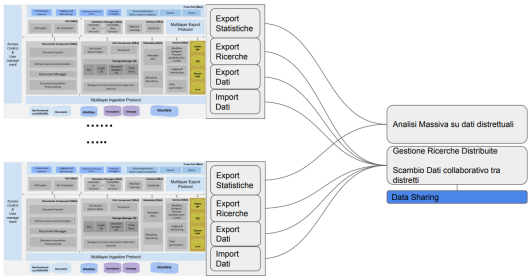

Figure 2: Multi-node Services Architecture.

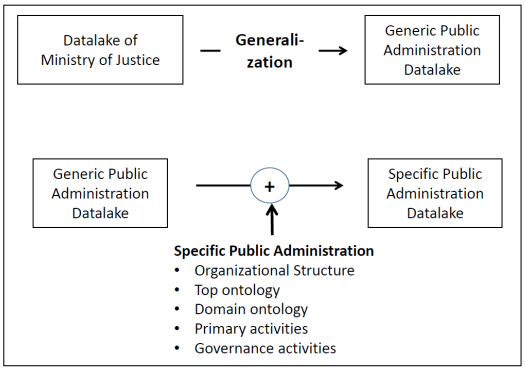

Figure 3: A general semantic document for the Italian Public Administration.

5. Service Architecture

The developed functionalities adopt a service architecture for deployment. The multi-node macro functional architecture is shown in Fig. 2. The components of the single node architecture (red frame) are the Multilayer Ingestion Protocol, Access Control & User Management, Storage Manager, Document Component, Metadata Manager, Service Manager, NLP Service Manager, Analysis, Front End, and Multilayer Export Protocol.

6. Conclusions: Towards a Semantic Document System for the Public Administration

The semantic document system described in the work is potentially useful for all Public Administrations (PAs). A project to disseminate the system should involve two phases: an initial phase of parameterization, concerning the organizational structure, ontologies, and primary and governance processes, and a second phase of customization (see Fig. 3). Such a project requires strong governance, aligning with the strategic path of digital transformation of the country, currently being implemented in the National Strategic Hub. Including services for a semantic document system for the PA in the service architecture of the Hub would require the production of a common top ontology for the PA and high-level modeling of primary and governance processes, with subsequent customization by the individual PAs.

References

[1] 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.

[2] R. Pozzi, R. Rubini, C. Bernasconi, M. Palmonari, Named entity recognition and linking for entity extraction from italian civil judgements, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2023, pp. 187–201.

[3] R. Pozzi, F. Moiraghi, F. Lodi, M. Palmonari, Evaluation of incremental entity extraction with background knowledge and entity linking, in: Proceedings of the 11th International Joint Conference on Knowledge Graphs, 2022, pp. 30–38.

[4] P.-L. H. Cabot, R. Navigli, Rebel: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2370–2381.

[5] 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), IEEE, 2021, pp. 173–183.

[6] A. Campi, S. Ceri, M. Dilettis, B. Pernici, et al., Variants analysis in judicial trials: Challenges and initial results, in: Proc. ECML PKDD Workshop on Knowledge Discovery and Process Mining for Law (KDPM4LAW), 2023, pp. 1–14.

[7] B. Pernici, C. A. Bono, L. Piro, M. Del Treste, G. Vecchi, Improving the analysis of the judiciary performance-the use of data mining techniques to assess the timeliness of civil trials, International Journal of Public Sector Management 37 (2024) 59–76.

[8] C. Batini, Manuale di qualità dei dati, documenti, modelli di giustizia, 2022.

[9] L. Floridi, M. Holweg, M. Taddeo, J. Amaya, J. Mökander, Y. Wen, Capai-a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act, Available at SSRN 4064091 (2022).