Toward a national repository of administrative procedures: Web Crawling and Process Modeling with LLMs
IEEE Computer Society Conferences

Francesca De Luzi
Sapienza Università di Roma, Italy
Email: deluzi@diag.uniroma1.it

Mattia Macrì
Sapienza Università di Roma, Italy
Email: macrì@diag.uniroma1.it

Massimo Mecella
Sapienza Università di Roma, Italy
Email: mecella@diag.uniroma1.it

***Abstract*—This paper presents a scalable approach to the identification, extraction, and modeling of administrative procedures within the Italian public sector. Despite ongoing reform efforts, Italy lacks an official, comprehensive catalogue of such procedures—an absence that hinders administrative simplification and digital transformation. To address this gap, we propose (i) the development of AdmPModeler, a pipeline that uses Large Language Models (LLMs) and prompt engineering techniques to convert legal texts into structured, human-readable process models, and (ii) a nationwide web crawling effort targeting the public administration websites, where procedural documentation is legally mandated to be published. This work lays the foundation for a national repository of administrative procedures—that support transparency, accountability, but also enable advanced policy experimentation, simplification initiatives, and automated compliance verification.**

***Index Terms*—E-Government, Administrative Procedures, Large Language Models, Prompt Engineering.**

1. Motivation

Public administrations operate through well-defined, regulated procedures. However, in practice, especially in Italy, the exact number of such procedures is unknown. While current reform efforts under the PNRR (Italy's Recovery and Resilience Plan) claim a target of simplifying at least 600 procedures, interviews with domain experts reveal a surprising truth: there is no official or comprehensive catalogue of these procedures. In fact, the procedures could be a thousand, two thousand—we don't know. In Italy, no one knows, and that is the incredible part. This ambiguity poses a fundamental challenge to administrative simplification and therefore digitalization. Without a concrete baseline, it is impossible to measure progress or even define the scope of reform. A method is needed to automatically identify, extract, and structure administrative procedures across heterogeneous sources.

2. A scalable approach: AdmPModeler & web-scale crawling

In our work, we collaborated with domain experts involved in the "1000 Esperti" project[1] and we proposed a two-step approach to address the challenges of automatic identification of public administrative procedures and the development of tools to assist experts in digitising these procedures. We are currently working on the first component, which we have named AdmPModeler [1]. It is a pipeline developed to convert legal texts describing administrative procedures into formal, human-readable process models using LLMs.

Firstly, we created a technique to extract generic process models. Secondly, we adapted this technique to the public administration domain. The technique leverages various prompt engineering methods, including prompt chaining, chain of thought, role prompting, and LLMs-as-a-judge. AdmPModeler takes as input one or more documents describing the administrative procedure to be modeled and returns a representation of the procedure in both diagram format and a tabular CSV description. The diagram provides a clear and immediate visualization of the activity flow within the procedure, while the CSV offers a detailed metadata representation for each activity. Moreover, the tool's output can be easily converted into standard process modeling formats, such as BPMN.

The second component, more speculative yet crucial, emerged from expert interviews conducted during the subsequent validation phases of the tool. This highlighted the need for a national data collection effort specifically targeting the "Amministrazione Trasparente" sections of Italian government websites. According to Law 33/2013[2], all public entities are required to publish their active procedures online. These pages are often similarly structured, providing a consistent target for large-scale crawling. Although the quality and consistency of the publications vary (PDFs, HTML

1. This Italian project, cf. https://www.espertipnrr.it/
2. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2013-03-14;33!vig

tables, etc.), they remain the most concrete manifestation of procedural knowledge across thousands of municipalities.

The technique relies on a set of four prompts. Each is described in general terms and then detailed in how it was used to develop AdmPModeler. These prompts share key design principles: *(i)* they were initially structured using the CO-STAR prompting framework; *(ii)* they incorporate Chain-of-Thought (CoT) principles and role prompting to enhance the reasoning capabilities of LLMs within a specific domain; and *(iii)* their descriptions were intentionally kept concise. Figure 1 illustrates the full pipeline of AdmPModeler, highlighting each prompt's role in the process.

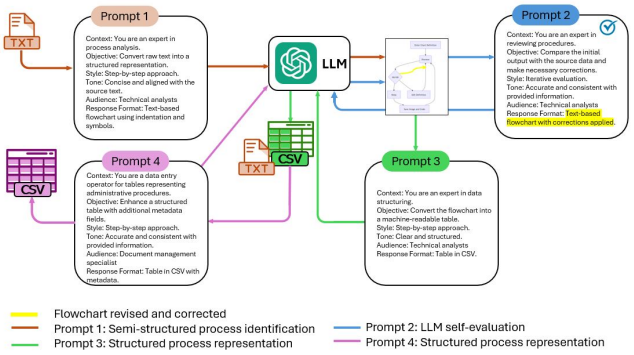

Figure 1. AdmPModeler detailed pipeline.

3. Challenges and discussion points

This dual strategy raises several important questions for the community:

- Scalability - but maintaining differentiation: Italy has 7,986 municipalities, many of which publish overlapping procedures. Can we cluster procedures to estimate their actual variety?
- Quality of input documents: the data available from the Transparent Administration section of the entity's website include both structured tables and PDFs. How can we build robust analysis techniques that fit different formats and quality levels? Are LLMs capable of producing a valuable output?
- Human-in-the-loop correction: Even with LLM-based extraction, expert correction is necessary for accuracy. How can this interaction be optimized to balance scalability with reliability?

The ultimate goal is to create a national repository of administrative procedures—semantically tagged, searchable, and updatable. This repository would not only serve transparency and accountability objectives but also enable advanced work in simplification, policy experimentation, and automated compliance checking. AdmPModeler already demonstrates that LLMs, when properly guided, can generate accurate structured representations from dense legal texts with over 78% average accuracy. Combined with crawling strategies, we believe this work could provide the first concrete steps toward an automated cataloging system for public administration procedures.

4. Positioning

This work is situated at the intersection of *e-Government* and *Service Environments*, addressing the growing need for interoperable, reusable, and semantically rich digital public services within smart cities. The proposed approach contributes to the development of a shared infrastructure for administrative digitalization by designing and exposing public APIs for the structured creation, access, and management of administrative procedures, paving the way for a national repository of administrative services. These APIs are intended not just as technical artefacts but as enablers of organizational innovation, aligning with the broader goals of interoperability and administrative simplification promoted at the European and national levels.

This positioning is further supported by ongoing academic contributions. The methodology, architecture, and early results of the approach have been accepted for publication in an upcoming article at the EGOV 2025 Conference, one of the primary venues for digital government research in Europe. Additionally, a technical abstract [2] describing the service modeling pipeline has been accepted for the Software Technologies: Applications and Foundations (STAF) conference, highlighting the methodological rigor and software engineering contribution of the project.

5. Conclusion

Our work offers a practical and theoretically informed contribution to the development of service-based infrastructures for smart cities, with a specific focus on digitizing and harmonizing public administration services. It embodies the kind of foundational infrastructure that can be reused across domains, supports policy-driven innovation, and facilitates the delivery of smarter, more inclusive public services. It directly addresses the goals of the CINI Smart Cities & Communities Lab by enabling modularity, interoperability, and reusability across vertical domains, supporting the evolution of smart governance ecosystems through actionable, standards-based service components.

References

[1] Macrì, Mattia and De Luzi, Francesca and Mecella, Massimo, "AdmPModeler: Modeling Administrative Processes using Large Language Models. A Case Study," Electronic Government, 2025, Springer

[2] De Luzi, Francesca and Macrì, Mattia and Mecella, "Toward a national repository of administrative procedures: Web Crawling and Process Modeling with Large Language Models," Software Technologies: Applications and Foundations: STAF 2025 Collocated AI4DPS Workshop, 2025, Springer