Automated Analysis with Event Log Enrichment of the European Public Procurement Processes

Automated analysis with event log enrichment of the European public procurement processes

Roberto Nai[1 [0000−1111−2222−3333]], Emilio Sulis[1 [0000−0003−1746−3733]], and Laura Genga[2 [0000−0001−8746−8826]]

[1] Computer Science Department - University of Turin, Italy,
{roberto.nai,emilio.sulis}@unito.com
[2] Eindhoven University of Technology (TU/e), The Netherlands
l.genga@tue.nl

Abstract. The length and extension of legal processes are two of the main problems of contemporary justice. Adopting process analysis techniques can serve to understand the course of these processes, to identify bottlenecks, and to propose improvements. This paper showcases an exploration of legal event logs using process mining techniques. First, we discuss the results obtained by applying state-of-the art process discovery techniques to data obtained from public tenders. Then, we show how to use natural language processing to automatically extract events and dates from the texts of the tenders and leverage this information for improving the results of process mining techniques. As a proof-of-concept, we propose a comparison of the French, Spanish, and Italian cases through process discovery based on calls for tenders from the European TED repository.

Keywords: Public procurement · Process mining · Event log enrichment · Natural Language Processing

1 Introduction

Public procurement processes are often lengthy, which has an impact on the legal and political system as a whole. In addition, special attention is paid to the European landscape, where there is difficulty in standardizing the legislative framework because of the different characteristics and practical applications peculiar to each member state. The increasing adoption of legal information technologies opens the way to exploit digital data, both for process analysis and for text analysis (e.g., public competition notices). In the context of public administration and legal information technologies, organizational aspects leading to the delivery and awarding of public tenders can be examined and supported by legal analysis tools. The areas of application are many in the broader field of Business Process Management (BPM) [6]; however, this type of study is not yet common in the analysis of legal processes. We propose the adoption of automated Process Mining (PM) techniques from legal data. First, we present the preparation steps for constructing a legal event log to be examined by PM techniques. Second, we show how process discovery can be a useful technique for

comparing the procedures (PM models) of public tenders in different member states. Third, we demonstrate how to enrich the event log by extracting features from public tenders by using Natural Language Processing (NLP) techniques. As a case study, we propose to apply this approach to the European public procurement process, starting from the examination of public tenders published only by Tenders Electronic Daily (TED), the online version of the "Supplement to the Official Journal" of the EU, dedicated to European public procurement [15].

We focus on the following research questions (RQs): RQ1) Starting from a legal dataset, can we create a PM model for different countries?; RQ2) Can we compare the legal processes of different countries? RQ3) Starting from a textual legal dataset of a country, can we extract relevant information and events to enrich our PM model?

In the remainder of the article, we describe the background with related work and the case study in Section 2, the methodology adopted in Section 3. Section 4 introduces the results, and Section 5 discusses some concluding remarks.

2 Background

2.1 Related work

Legal aspects have been typically addressed in a BPM perspective by considering regulatory compliance issues. Recently, the PM discipline gains attention for several application areas, from the typical BPM research area [2], healthcare [8], or education [12]. The conformance checking research seems promising to investigate legal compliance issues [5, 16].

Previous works explored real-world cases of process discovery involving public procurement. For instance, a case study focuses on a heuristic algorithm revealing a concept drift in publication of contract in the Philippines [19], while [17] investigates public procurement procedures in Croatia. A comprehensive application of process discovery in the legal field is [21], where authors applied discovery techniques for the extraction of lawsuit processes from the information system of the Court of Justice of the State of Sao Paulo, Brazil.

Artificial intelligence techniques have been explored in legal domain, including machine learning techniques [13], merging dataset [14], explainability [10]. Finally, a relevant area concerns NLP techniques. Applications are finding increasing interest in BPM [1]. Recent works propose to exploit unstructured information in natural language textual descriptions of processes to enable formal process modeling [9]. Other works investigate the enrichment of an event log based on unstructured text [7].

2.2 Case study

Legal data of tenders and processes related to publication procedures can be found in the Tenders Electronic Daily (TED) website[3]; TED publishes 735 thou-

[3] `https://ted.europa.eu/TED/browse/browseByMap.do`

sand procurement notices a year, including 258 thousand calls for tenders which are worth approximately €670 billion.

Dataset overview In the TED dataset, each tender is identified by an alphanumeric value called *document-number* (the key value) and it has the following relevant features: the *sector* to which it belongs ("services", "works", or "supplies"); the *NUTS* (Nomenclature of Territorial Units for Statistics [4]) of the contracting authority (public administration) that issued the procurement; the *type* of the contracting authority (Ministry, European Institution, Regional or local autorithy, etc.); and the *amount* of the tender. The complete schema of the CSV is publicly available[4].

The legal process and the activities We focused on regional authorities of the French (FRA), Italian (ITA), and Spanish (ESP) cases, following the suggestion of domain experts. According to the domain experts, the three member states are similar in the legal review system. There are five main activities in the public procurement process. The initial event is the PUBLICATION of the tender, then the PARTICIPATION of individual entities proposing a bid. Following the adjudication of the tender (AWARD), the contract is issued (CONTRACT-START) with the procedure ending with the deadline provided for in the call (CONTRACT-END). Furthermore, a sixth event can be identified in the notice, i.e. the BID-OPENING which takes place before AWARD.

3 Methodology

3.1 The process dataset construction

The dataset was created from the collection of annual tender and award notices available on the TED portal[5]. For each tender, identified by a document-number, we generated an event log trace. Next, we retrieved the PDF text, and the trace was enriched with the new date found in the text via NLP. Our final collection of French, Italian, and Spanish cases from 2016 to 2022 is publicly available [6].

Event log of legal process We introduce here the terminology for automated process analysis. The starting point of a PM research is an *event log*, i.e. a set of *traces* where each trace stores a sequence of *events*, each representing the execution of an *activity* occurred during a single execution of the process, possibly together with some additional data (e.g., the resource who performed the activity) [2]. Every trace is identified by means of a so-called *case ID*. Different traces in the event log are called *process variants* (*variants*), since they represent

[4] `https://data.europa.eu/api/hub/store/data/ted-csv-data-information-v3-5.pdf`

[5] `https://data.europa.eu/data/datasets/ted-csv?locale=en`

[6] `https://github.com/roberto-nai/JUSMOD2023`

alternative ways to execute the same process (cases may perform activities in a different order before the end).

Three attributes are necessary in order to generate an event log from a set of not process-oriented samples, such as the Tender dataset we start from: first, the *case ID* indicating which case of the process is responsible for each event. Second, the *event class* (or *activity name*) specifying which activity the event refers to. Third, the *timestamp* specifying when the event occurred [3]. An event log may carry additional attributes in its payload; these are called *event-specific attributes* (or *event attributes* for short) [2]. In our case study, document-number is the case identifier. The activities include the date of each activity at the level of granularity of the day on which the event occurred. In order to have a consistent event log, we removed cases too short, which are not meaningful according to domain experts. In addition, we add in the log file as attributes the information on the corresponding sector, the NUTS, and the amount. The script in Python to transform TED's CSV files into event log is publicly available[7].

```
Case ID;Activity;Timestamp;Sector;Amount;Nuts;Country
...
2017106814;PUBLICATION;2017-03-17;S;1035000.0;ITC13;IT
2017106814;PARTICIPATION;2017-05-09;S;1035000.0;ITC13;IT
2017106814;AWARD;2017-06-07;S;1035000.0;ITC13;IT
2017106814;CONTRACT-START;2017-09-01;S;1035000.0;ITC13;IT
2017106814;CONTRACT-END;2022-07-31;S;1035000.0;ITC13;IT
2017107959;PUBLICATION;2017-03-20;S;637622.4;ITE19;IT
2017107959;PARTICIPATION;2017-05-03;S;637622.4;ITE19;IT
2017107959;CONTRACT-START;2017-06-01;S;637622.4;ITE19;IT
2017107959;AWARD;2017-06-15;S;637622.4;ITE19;IT
2017107959;CONTRACT-END;2020-05-31;S;637622.4;ITE19;IT
...
```

Fig. 1. Legal event log example in CSV format. Full size image available at `https://github.com/roberto-nai/JUSMOD2023`

3.2 Process discovery and variant analysis

We imported the event log file in the tool DISCO from Fluxicon[8] to perform an initial analysis of the data involving filtering, exploration with process discovery, and bottleneck search.

Furthermore, we perform variant analysis to discover possible significant differences between subgroups of process executions. Variant analysis is a family of techniques that analyze event logs to identify and explain the differences between two or more processes. In this research, we performed a *manual* comparison of the process models obtained for the identified variants. With this exploratory phase, it is possible to gather useful indications as to whether these variants actually show differences and to determine which properties we want to explore further.

[7] `https://github.com/roberto-nai/JUSMOD2023`
[8] `https://fluxicon.com/disco/`

3.3 Event log enrichment

Information extraction is a challenging task that requires various techniques, including named entity recognition (NER), regular expressions, and text matching, among others [11]. In this research, we dive into date extraction using the Spark NLP `DateMatcher` and `RegexMatcher` *annotators*. Spark NLP[9] is a library built on top of Apache Spark ML [18]. An *annotator* in Spark NLP is a component that performs a specific NLP task on an input text document and produces an output document with additional metadata (e.g., dates or amounts) [20]. To measure the efficiency of these methods, we compared the performance of the pipelines from Spark NLP with the built-in method of Python dedicated to regular expressions (RE). As a proof-of-concept, we extracted dates from the texts of only one state (enriching the ITA log). We used the open source library *PyPDF2*[10] to extract text parts from PDFs to be used as input for NLP tasks[11].

4 Results

4.1 Legal Dataset

As stated in Section 3.1, we focused on the period from 2016 to 2022. Following the indications of the domain experts, the time frame is fairly consistent with the validity of the Italian public procurement code, which came into force in April 2016 and was replaced on 31 March 2023. The dataset contains data for 66,382 French tenders, 16,861 Italian tenders and 33,032 Spanish tenders. We started from the dataset to construct the event logs, with which to provide a comparison between the three cases by mean of standard metrics, i.e. the number of traces, instances and variants.

4.2 Legal event log

We obtained a main legal event log of 116,275 cases, including 285 variants. We provide an example of the legal event log in Figure 1; cases from all the countries have a median duration of 18.6 weeks and a mean duration of 35.8 weeks. The initial process discovered from the event log of all tenders is described in Figure 2. The process model highlights the most common process behaviours; rectangles represent process activities, while edges represent pair-wise ordering relations among the activities. The darker a rectangle is, the more frequently the corresponding activity occurs in the event log. Similarly, the thicker an edge is, the more frequently it is observed in the event log that the source activity is eventually followed by the target activity. The exploration of the diagram and the log shows that some data need to be filtered out in order to get a better version of the process. For instance, we notice that some paths (e.g. from

[9] `https://sparknlp.org/docs/en/quickstart`
[10] `https://pypdf2.readthedocs.io/en/3.0.0`
[11] Source code publicly available: `https://github.com/roberto-nai/JUSMOD2023`

PUBLICATION to CONTRACT-END) are not possible according to domain experts and some cases with extremely high duration also, e.g., the tenders without a stop date (CONTRACT-END). The very high number of variants also includes combinations of activities that were judged to be incorrect (e.g. AWARD to PUBLICATION) or insignificant (e.g. only PUBLICATION and PARTICIPATION events). This issue is due to data quality problems, as the tender data were entered manually by the operators of the individual contracting authorities.

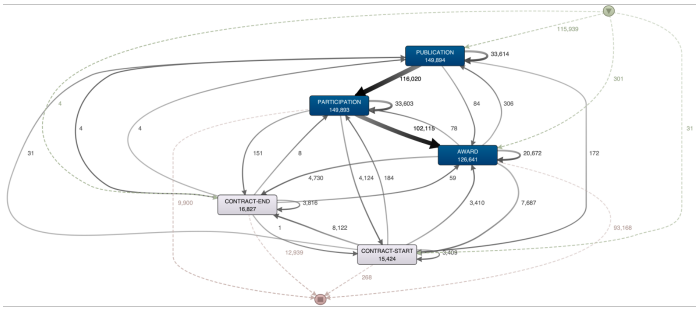

Fig. 2. Process discovery from raw TED data. Full size image available at `https://github.com/roberto-nai/JUSMOD2023`

Filtering the event log We opted to include only the processes starting after the initial date of this interval (1 January 2016) and concluding before the end (31 December 2022). In addition, to remove incomplete cases, a filter was also inserted to define the endpoints of the process: the first event must be PUBLICATION and the final event CONTRACT-END. Filtering the log we obtained a log with 6,502 cases (31,321 events), 5 variants, mean duration of 26.6 months, median case duration 18.3 months and a standard deviation of 16.5. In addition to the number of cases, the number of variants is also reduced compared to the initial number of cases, as the removed cases comprised various combinations of activities (e.g. about 60 per cent of the processes contained the variant of three activities PUBLICATION to PARTICIPATION to AWARD). The corresponding CSV and XES[12] filtered without cases that are not interesting were generated; the event log files are publicly available[13].

Event log analysis of three states We analyzed process variants by taking into account the performance of every country, expressed in terms of their duration. Table 1 summarizes the differences in event logs corresponding to French, Italian, and Spanish processes. Interestingly, we immediately noticed that there are

[12] XES is a standard format for process mining, `https://xes-standard.org/`
[13] `https://github.com/roberto-nai/JUSMOD2023`

significant differences in the number of procedures and the average duration. France contains a much larger number of cases, not proportionate to the larger population of that state. On the other hand, the median duration is almost 17 months, whereas the Italian case (with a quarter of the procedures) sees a longer duration of twice as much (about 37 months). The Spanish case has the lowest number of cases and a similar median duration to the French case.

Table 1. Comparison of three countries' event logs (Spain, France, Italy) by the number of Cases, Events, Variants, Mean and Median case duration (months), standard deviation (SD)

Country	**Cases**	**Events**	**Variants**	**Mean c.d.**	**Median c.d.**	**SD**
ESP	838	4,063	5	21	17.6	12.6
FRA	22,102	4,601	4	25.7	17.1	16.9
ITA	1,063	5,156	4	34.8	37.1	14.3

4.3 Process discovery results

It is possible to clearly visualise both similarities and differences in the three process diagrams for each country in Figure 3. The diagrams introduce the average duration between two activities in the weight of the arcs, and the size of the arcs easily detect bottlenecks. The average time from the notice of the tender to participation is about 36 days (min. 34, max. 39), the average time to determine the winner of the tender is about 45 days (min. 36, max. 54). The time to decide the winning tender is 19.3 months. In addition to the timing, loops in the Spanish process can be observed in the PUBLICATION and PARTICIPATION activities: according to the domain experts, this is due to the fact that the contracting authorities can re-publish a tender notice, so participation also takes place again on the basis of the new publication. A second observation is that AWARD activity does not always take place; according to domain experts, this is due to 'framework agreements' (multi-year contracts over several tenders, published but automatically awarded).

4.4 Event log enrichment

Starting from the 1,063 total cases for the ITA country we found, in the text of tenders, the presence of the bid opening section in 287 cases; in 17 cases, the section contained no text. In the remaining 270 cases, the three NLP techniques described above (Section 3.3) were applied and compared. Table 2 shows the results obtained. For this date recognition task, the RE method performed best in our experiment. As a demonstration, the dates extracted from the tender notices were associated with the new BID-OPENING event, which was then added to the ITA tender log.

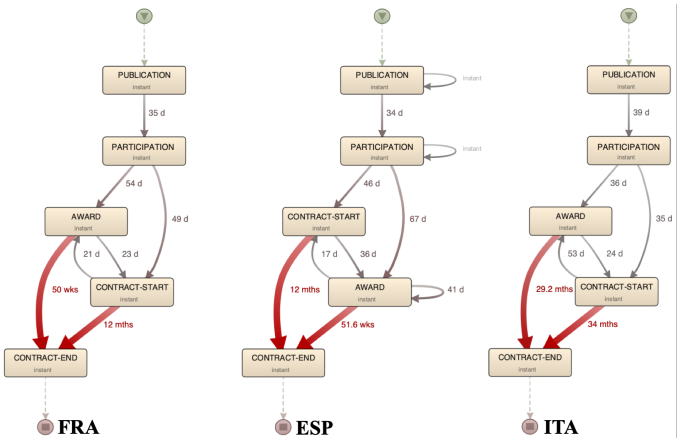

Fig. 3. The European public procurement processes of France, Spain, and Italy extracted from notices published in Tenders Electronic Daily (TED). Full size image available at `https://github.com/roberto-nai/JUSMOD2023`

Table 2. Comparison of the three NLP methods used to extract the new BID-OPENING event from PDF texts.

Method	Dates found (tot.)	Timing (min.sec)
RE	270	0.6
DateMatcher	223	1.10
RegexMatcher	209	1.11

To add a new event from dataset Figure 4 shows the new BID-OPENING event extracted from the tender notices. The new event shows how the opening of received bids takes place in a short time (transition from PARTICIPATION to BID-OPENING) while the main part of the time required for awarding the work to the economic operator is given by the decision-making process of the contracting authorities (transition from BID-OPENING to AWARD).

5 Conclusion

In this work, we applied PM techniques to a legal process to understand its execution in reality as well as to identify potential issues and inefficiencies. More precisely, we implemented a combination of process discovery, NLP, and variant analysis techniques. As far as RQ1 is concerned, the transformation of the CSVs has led to the creation of an event log. The consequent discovery techniques allowed to obtain relevant information on the main behaviour of the award process. With regard to RQ2, once the logs were filtered, it was possible to compare the country-based processing in terms of time performance. For RQ3, NLP tasks applied to the notice texts made possible to explore some information extraction techniques as well as to obtain a new event to add to the ITA event log, increas-

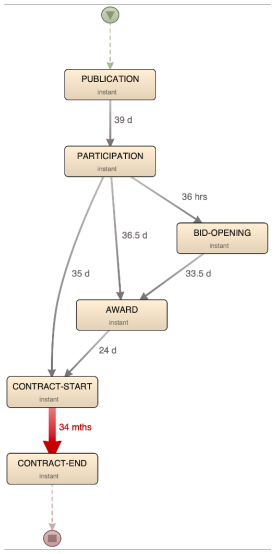

Fig. 4. The public procurement process of ITA, with the new event BID-OPENING extract from the texts. Full size image: `https://github.com/roberto-nai/JUSMOD2023`

ing knowledge about the process. In the future, we intend to further develop our investigation by exploring different techniques for process discovery and variant analysis (e.g. automated variant analysis), also apply it to other features in the log (amount, NUTS, etc.). Furthermore, we intend to further explore other NLP techniques to extract more data from the texts and enrich the process event log even more, also for other countries. Finally, we want to explore the application of prediction algorithms on track prefixes to determine in advance processes that are excessively long or that will not end (e.g. for an appeal to the administrative justice by economic operators).

References

1. van der Aa, H., Carmona, J., Leopold, H., Mendling, J., Padró, L.: Challenges and opportunities of applying natural language processing in business process management. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proc. of the 27th COLING. pp. 2791–2801. ACL (2018), `https://aclanthology.org/C18-1236/`
2. van der Aalst, W.M.P.: Process Mining - Data Science in Action. Springer (2016). https://doi.org/10.1007/978-3-662-49851-4
3. van der Aalst, W.M.P., Carmona, J. (eds.): Process Mining Handbook, LNBIP, vol. 448. Springer (2022). https://doi.org/10.1007/978-3-031-08848-3
4. Becker, S.O., Egger, P.H., Von Ehrlich, M.: Going nuts: The effect of eu structural funds on regional performance. J Public Econ **94**(9-10), 578–590 (2010)
5. Carmona, J., van Dongen, B.F., Solti, A., Weidlich, M.: Conformance Checking - Relating Processes and Models. Springer (2018). https://doi.org/10.1007/978-3-319-99414-7

6. Dumas, M., Rosa, M.L., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management. Springer (2018). https://doi.org/10.1007/978-3-662-56509-4
7. Geeganage, D.T.K., Wynn, M.T., ter Hofstede, A.H.: Text2el: Exploiting unstructured text for event log enrichment. In: 2022 16th Int. Conf. on SITIS. pp. 1–8 (2022). https://doi.org/10.1109/SITIS57111.2022.00010
8. Jorge Munoz-Gama et al.: Process mining for healthcare: Characteristics and challenges. J. Biomed. Informatics **127**, 103994 (2022). https://doi.org/10.1016/j.jbi.2022.103994
9. Maqbool, B., Azam, F., Anwar, M.W., Butt, W.H., Zeb, J., Zafar, I., Nazir, A.K., Umair, Z.: A comprehensive investigation of BPMN models generation from textual requirements. In: Kim, K.J., Baek, N. (eds.) ICISA. LNEE, vol. 514, pp. 543–557. Springer (2018). https://doi.org/10.1007/978-981-13-1056-0_54
10. Meo, R., Nai, R., Sulis, E.: Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next? In: Silvia Chiusano et al. (ed.) ADBIS 2022, Turin, Italy, September 5-8, 2022. LNCS, vol. 13389, pp. 25–34. Springer (2022). https://doi.org/10.1007/978-3-031-15740-0_3
11. Mohit, B.: Named entity recognition. In: Natural language processing of semitic languages, pp. 221–245. Springer (2014)
12. Nai, R., Sulis, E., Marengo, E., Vinai, M., Capecchi, S.: Process mining on students' web learning traces: A case study with an ethnographic analysis. In: Olga Viberg et al. (ed.) EC-TEL Int.Conf.proc. Lecture Notes in Computer Science, vol. 14200, pp. 599–604. Springer (2023). https://doi.org/10.1007/978-3-031-42682-7_48
13. Nai, R., Sulis, E., Meo, R.: Public procurement fraud detection and artificial intelligence techniques: a literature review. In: Danai Symeonidou et al. (ed.) 23rd EKAW Int. Conf. proc. CEUR, vol. 3256. CEUR-WS.org (2022), `https://ceur-ws.org/Vol-3256/km4law4.pdf`
14. Nai, R., Sulis, E., Pasteris, P., Giunta, M., Meo, R.: Exploitation and merge of information sources for public procurement improvement. In: Irena Koprinska et al. (ed.) Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Commun. Comput. Inf. Sci., vol. 1752, pp. 89–102. Springer (2022). https://doi.org/10.1007/978-3-031-23618-1_6
15. Prier, E., Prysmakova, P., McCue, C.P.: Analysing the european union's tenders electronic daily: Possibilities and pitfalls. IJPM **11**(6), 722–747 (2018)
16. Pufahl, L., Rehse, J.: Conformance checking with regulations - A research agenda. In: Koschmider, A., Michael, J. (eds.) 11th Int. Workshop on EMISA. vol. 2867, pp. 24–29. CEUR-WS.org (2021), `http://ceur-ws.org/Vol-2867/paper4.pdf`
17. Rabuzin, K., Modrušan, N., Križanić, S., Kelemen, R.: Process mining in public procurement in croatia. In: Lalic, B., Gracanin, D., Tasic, N., Simeunović, N. (eds.) Proceedings on 18th Int. Conf. on IS'20. pp. 473–480. Springer, Cham (2022)
18. Salloum, S., Dautov, R., Chen, X., Peng, P.X., Huang, J.Z.: Big data analytics on apache spark. Int J Data Sci Anal **1**, 145–164 (2016)
19. Sangil, M.J.: Heuristics-based process mining on extracted philippine public procurement event logs. In: 2020 7th Int. Conf. on BESC. pp. 1–4 (2020). https://doi.org/10.1109/BESC51023.2020.9348306
20. Thomas, A.: Natural Language Processing with Spark NLP: Learning to Understand Text at Scale. O'Reilly Media (2020)
21. Unger, A.J., Neto, J.F.d.S., Fantinato, M., Peres, S.M., Trecenti, J., Hirota, R.: Process mining-enabled jurimetrics: Analysis of a brazilian court's judicial performance in the business law processing. In: Proc. of 18th ICAIL. p. 240–244. ACM, NY, USA (2021). https://doi.org/10.1145/3462757.3466137