# Process Mining of Public Administration Operations from Big Data

Dmitry Mingazov[1,2,†], Fabio Celli[2,*,†]

[1]University of Camerino, via Madonna delle Carceri 7, Camerino, 62032, Italy
[2]R&D Gruppo Maggioli, via Bornaccino 101, Santarcangelo di Romagna, 47822, Italy

**Abstract**

In this paper we use Process Mining and unsupervised learning to extract Graphs from Big Data produced by Public Administration software logs. Starting from millions raw logs of a software used in many Italian municipalities, we group functions related to specific Public Administration operations - such as management of reversals, tax collection seizures, budget change - by means of clustering techniques. Then we apply Inductive Miner on clusters to extract process models and we visualize them in Business Process Models Notation, that represent generalized ways to perform specific operations and can be exploited for detailed process modeling, communication, and analysis of the workflows in the Public Administration. We argue that this work paves the way towards modeling Public Administration operations into Knowledge Graphs in a transparent way, suitable for the integration into ethical AI systems.

**Keywords**

Process Mining, Public Administration, Knowledge Graphs, Big Data

## 1. Introduction and Background

Public Administration (PA) increasingly relies on effective process management to ensure the successful execution of both administrative and front-end services to the citizens. The application of Artificial Intelligence (AI) to the PA is crucial for improving the efficiency and transparency of process management in the public sector. However, AI applications within the PA remain underdeveloped [1] for different reasons. These include data sparsity, lack of data interoperability [2], a general risk aversion in the public sector [3] and the legacy of outdated Information Technology systems that are hard to integrate with AI tools. Nevertheless, there is a huge effort of the scientific community to make advances and improvements into the PA sector. On the one hand there are top-down approaches with Knowledge Graphs (KGs). These represent entities, process steps and the relations between them in a machine-readable form. KGs can include complex knowledge about a domain and facilitate PAs to adopt a data-centic orientation and operation analytics [4]. On the other hand there are bottom-up approaches that try to extract patterns, rules and relations directly from data. Among these techniques, Process Mining [5], transparent Machine Learning and Association Rule Learning [6] are powerful tools for discovering, modeling and managing business processes in the PA. Many organizations are currently utilizing Process Mining to discover patterns in data, applying research and innovation actions to the business [7]. An analysis of 144 research papers in the business applications of Process Mining [8] revealed that most of the existing research focuses on extracting models within a single organization to improve a single business process. Research on using Process Mining across different systems or between organizations is still underdeveloped. Additionally, the current literature rarely explores how Process Mining can be applied to analyze physical services, like municipal operators working at the counter. Process Mining has the potential to offer valuable insights into customer processes, but to achieve this, researchers need to explore more complex use cases, and there is need for collaboration between academics and practitioners to obtain good results. Machine Learning in the public sector instead is mainly used for the automation of routine operations that have complicated elements, such as triaging phonecalls or correspondence to the right points of contact [9]. These algorithms are mainly supervised and trained for specific tasks but the advent of more powerful techniques with less transparent models, such as Deep Learning and Generative AI, increased the risk of bias and discrimination in using algorithms for taking decisions [10] and this is especially true in the PA [11]. Nevertheless, there are promising applications of transparent Process Mining [12] in the medical domain. This study utilizes Process Mining to extract Petri Nets and graphs in Business Process Model Notation (BPMN) from big data of many municipalities encompassing PA operations. BPMN excels at depicting the flow of activities within a process, it and has been ratified as ISO 19510 standard and also

*Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy*
*Corresponding author.
† These authors contributed equally.
dmitry.mingazov@maggioli.it (D. Mingazov);
fabio.celli@maggioli.it (F. Celli)
0000-0002-7309-5886 (F. Celli)

extended to cover some PA use cases [13]. It visually depicts the sequence of activities, decision points, and potential outcomes within a process and facilitates the collaboration between business analysts, process designers, and developers. Moreover, BPMN also provides a mapping with execution languages, particularly Business Process Execution Language (BPEL), thus it is possible to run automations and even build KGs from BPMN [14]. The paper is structured as follows: after a brief review of related works we introduce the data and the experiments, discuss the results, draw our conclusions and finally we trace our direction for future work.

### 1.1. Related Work

Recent attempts to apply Process Mining to big databases of logs from PA software revealed that this kind of data is very hard to process with existing techniques. Previous work of this kind counts 104 operators and 227.000 logs [15]. In particular these softwares are usually made of many different forms that allow the execution of nested operations or sub-parts of operations. In this scenario a form closure does not necessarily imply a parent relationship with the other open forms. Moreover, sometimes it happens that even if two forms are dependent, the closing date is incoherent, with a parent form closing before a child form. In fact, forms may remain opened for long even if they are not being used. The difficulties in the application of Process Mining to logs of PA data can be summarized in four problems [15]:

1. the impossibility to reduce multiple levels of interweaving to a simpler structure due to the need of the software to allow multiple nested operations in parallel;
2. the difficulty of making structural assumptions based on temporal relations;
3. the presence of loops and redundant activities, such as technical automated functions mixed with the actual operations;
4. the difficulty of labelling operations on the fly due to the potential incoherence between parent and child forms.

The presence of loops can be solved with correlation process mining [16], that is designed for logs in which events that belong to the same case are related to each other. Similar functions and similar control flows can be detected and grouped by coupling Process Mining with parametric dissimilarity measures and clustering algorithms like K-medoids [17]. However, it remains difficult to label operations and evaluate the quality of the labels, because clustering is an unsupervised Machine Learning technique. All these problems are current open challenges. The research in Knowledge Graphs is relatively less problematic. For example using data governance techniques on open data it is possible to build large semantic Knowledge Graphs that represent distributed data spaces for public e-procurement [18]. However, data heterogeneity within the PA presents a challenge for stakeholders, such as PA employees, developers and decision makers, in identifying relevant data standards, formats, and APIs for digitizing specific public services, especially those with few open data available. However, there are attempts to solve this issue with semantic modeling and linked open data principles [19], and link them to existing KGs. For example it is possible to enable the automated creation of human- and machine-readable descriptions of processes from data into ontologies, and link them to existing process descriptions of public services [20], such as legal ontologies. The gap between top-down and bottom-up approaches is still large. The main challenge in the bottom-up approach is the lack of semantics. In other words it is not possible to exactly know from software logs the semantics of the operation performed and its relation to the other operations. The main challenge with the Top-down approach instead is the heterogeneity of data. Ontologies and KGs encode the semantic relations between processes but lack the ability to link them to real processes of the PA. Bridging this gap would allow us to spot the inefficiencies in the PA and to have much more control on the entire administrative system.

## 2. Data Description

We collected logs generated by Sicraweb Evo, a software designed to perform many operations in Italian municipalities. This software is divided in a client side, i.e. a web application used by the municipality operators, and a server side, from which the logs are currently generated. The logging system was designed for debugging purposes and it does not yield direct information about the processes, as happens in similar software described in literature. Moreover, the quantity of logs is enormous, averaging at 7.7 million records per day from more than 2000 municipalities. For our experiments we random sampled 1 million logs from 15 different municipalities and more than 150 operators. To the best of our knowledge this is the first work that applies Process Mining on PA operations using such a large amount of data. The data is recorded as a sequence of REST calls to the server side of the application, where each call is a single activity, until recurrent patterns will be discovered and associated to higher level operations. Each REST call contains the following attribute fields:

- Activity: the atomic software function that is activated in the process;
- Resource: anonymized municipality and operator who used the software;

- Action order: a sequential number indicating the execution order of the activities;
- Relative time: progressive record of milliseconds starting at 0 with the first activity.

The presence of the Action order helps solving problem 2, making structural assumptions even when relative time is not consistent. However, case id and process id are inherently missing from data. The event logs used can be classified as ⋆⋆⋆ in the maturity level for Process Mining described in literature [21].

Process Mining algorithms operate on a set of cases, i.e. instances of processes. Since our dataset was lacking of case notations, we added them to the records. We assigned a case ID to each sequence of activities not interrupted by a change of client (municipality), date, operator or the opening of a new form. This approach was proven to work in a similar scenario [15].

# 3. Experiments and Discussion

Our contribution follows a bottom-up approach and presents two experiments. In the first experiment we want to understand how much the raw log data can be linked to operation labels. We assume the form titles as operation labels provided by the software designers, who are domain experts.We evaluate the relationship between operation labels and clustering by means of Homogeneity [22] and Silhouette metrics [23] [24]. Homogeneity measures how many clusters contain only logs which are members of a single operation, while Silhouette measures how similar are the logs in their own cluster compared to the other clusters. In the second experiment we apply Process Mining on clusters to extract Petri Nets and visualize them in BPMN. We use Replay Fitness [25] to evaluate the quality of the graphs extracted.

## 3.1. Clustering

Before applying any Process Mining algorithm to raw data, logs must be divided into chunks of homogeneous context. Following previous literature [17] we applied unsupervised clustering techniques, K-medoids and OPTICS for instance, to achieve that. We extracted features from the logs by using the frequency of specific activities. In this way we obtained a feature table, where rows represent case ids and the columns represent the frequency of activities. In order to reduce information sparseness, we applied Singular Value Decomposition and compressed the feature space from initial 1776 columns to two trials, with 100 and 50 columns respectively.

Results, reported in Table 1, show that K-medoid has higher Silhouette score, meaning that is able to aggregate more similar logs under operation labels. Homogeneity score is similar between the two, indicating that both algorithms are able to subsume the logs under the operations roughly the same way. A qualitative analysis revealed that OPTICS is able to manage noisy logs better than K-medoid, obtaining clearer graphs. The lower scores of OPTICS are possibly due to the fact that it tends to create a wastebasket cluster with noisy logs among other cleaner clusters, while K-medoid tends to aggregate noisy logs with others. Moreover, a manual check revealed that only 36.8% of the operations contains vertical functions from the same area. For example the management of reversals contains just functions from the financial area. The remaining 63.2% are operations that involve different areas. For example the management of purchase invoices contains functions used in the financial area as well as in the general affairs area. This indicates that the OPTICS algorithm may better reflect the actual percentage of homogeneous operations.

| algorithm | features | Silhouette | Homogeneity |
|---|---|---|---|
| K-medoid | 100 | 0.498 | 0.432 |
| OPTICS | 100 | 0.339 | 0.403 |
| K-medoid | 50 | 0.513 | 0.435 |
| OPTICS | 50 | 0.332 | 0.401 |

**Table 1**
Results of clustering experiments.

## 3.2. Process Mining

Each cluster of logs represent a supposed operation containing several variants. With the amount of data we processed we obtained more than 200 clusters with both algorithms. Some operations are represented by more than one cluster. There are by average 5.05 clusters per operation, with about 30 clusters that contain mainly technical and automatic functions, and cannot be mapped to any specific operation and can be discarded. Aiming at a representation of the software processes with high simplicity of understanding, We applied Inductive Miner to the clusters to obtain both Petri Nets and BPMNs, and ultimately chose BPMN to visualize our data. These represent generalized ways of performing operations. In order to make the process discovery more scalable, traces which shared the same set of activities, regardless of their edges, were grouped together and used as input for the discovery of BPMN. The whole discovery process was performed using custom Python scripts which made use of the PM4Py library [26]. We computed average Replay Fitness on 10 random clusters generated with both algorithms. The results with K-medoids is 0.976 and with OPTICS is 0.998, indicating that OPTICS captures information from all variants in a cleaner way, as emerged in the qualitative analysis. Figure 1 is a generalized BPMN graph of a purchase invoice management operation from 73 variants. The process can be represented by exclu-

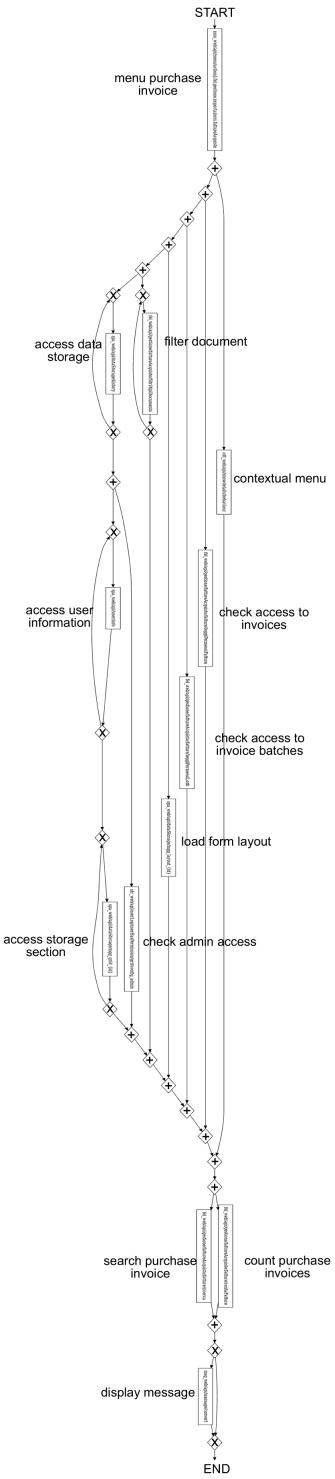


**Figure 1:** BMPN graph of purchase invoice management operation.

sive (x) and parallel (+) gateways. Despite BPMN models are not full KGs [27], they can serve as a ubiquitous visual tool across various disciplines, including software development, engineering design, and scientific experimentation. A great advantage of BPMN models is that it is possible to turn them into code and develop transparent automated processes from data with a bottom-up approach.

## 4. Conclusion and Future

We presented a method for the extraction of BPMN from big data using Process Mining and clustering techniques. The major contribution of this work to the scientific community is to apply these algorithms to big data in a real world scenario. We plan to evolve this work in three different ways: applying new Process Mining algorithms, enhancing inductive miner to extract configurable graphs and aggregate processes at a level above operations; test the development of automations by turning BPMN into code by means of AI tools; explore the integration of BPMN and KGs. The integration of BPMN and KGs holds significant promise for enhancing business process management. By combining the structured flow representation of BPMN with the rich semantic relationships captured in KGs, organizations can gain a deeper understanding of their processes and automate the management of PA processes based on a broader knowledge base. Future research can explore specific implementation frameworks and evaluate the impact of this integration on process efficiency and knowledge utilization within organizations.

## Acknowledgements

This work was supported by the European Commission grant 101120657: European Lighthouse to Manifest Trustworthy and Green AI - ENFIELD.

## References

[1] C. G. G Reddick, L. Anthopoulos, Information and communication technologies in public administration, 2015.

[2] G. Lodi, A. Maccioni, M. Scannapieco, M. Scanu, L. Tosco, Publishing official classifications in linked open data., in: SemStats@ ISWC, 2014, pp. 1–12.

[3] S. Nicholson-Crotty, J. Nicholson-Crotty, S. Fernandez, Performance and management in the public sector: Testing a model of relative risk aversion, Public Administration Review 77 (2017) 603–614.

[4] D. Zeginis, K. Tarabanis, An event-centric knowledge graph approach for public administration as

an enabler for data analytics, Computers 13 (2024) 17.

[5] S. Fioretto, Process mining solutions for public administration, in: European Conference on Advances in Databases and Information Systems, Springer, 2023, pp. 668–675.

[6] F. Guo, Research on public administration decision model based on big data association rules mining algorithm, in: 2023 International Conference on Networking, Informatics and Computing (ICNETIC), IEEE, 2023, pp. 544–549.

[7] C. dos Santos Garcia, A. Meincheim, E. R. F. Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos, E. E. Scalabrin, Process mining techniques and applications–a systematic mapping study, Expert Systems with Applications 133 (2019) 260–295.

[8] M. Thiede, D. Fuerstenau, A. P. Bezerra Barquet, How is process mining technology used by organizations? a systematic literature review of empirical studies, Business Process Management Journal 24 (2018) 900–922.

[9] M. Veale, I. Brass, Administration by algorithm? public management meets public sector machine learning, Public management meets public sector machine learning (2019).

[10] A. Páez, Negligent algorithmic discrimination, Law & Contemp. Probs. 84 (2021) 19.

[11] L. Cao, On machine learning and public administration, Frontiers in Management Science 1 (2022) 1–4.

[12] T. R. Neubauer, R. M. de Araujo, M. Fantinato, S. M. Peres, Transparency promoted by process mining: an exploratory study in a public health product management process, in: Anais do X Workshop de Computação Aplicada em Governo Eletrônico, SBC, 2022, pp. 37–48.

[13] V. Torres, P. Giner, B. Bonet, V. Pelechano, Adapting bpmn to public administration, in: International Workshop on Business Process Modeling Notation, Springer, 2010, pp. 114–120.

[14] S. Bachhofner, E. Kiesling, K. Revoredo, P. Waibel, A. Polleres, Automated process knowledge graph construction from bpmn models, in: International Conference on Database and Expert Systems Applications, Springer, 2022, pp. 32–47.

[15] F. Mouysset, C. Picard, C. Bortolaso, F. Migeon, M.-P. Gleizes, C. Maurel, M. Derras, Investigations of process mining methods to discover process models on a large public administration software, in: 37ème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2019), 2019, pp. 147–162.

[16] S. Pourmirza, R. Dijkman, P. Grefen, Correlation mining: mining process orchestrations without case identifiers, in: International Conference on Service-Oriented Computing, Springer, 2015, pp. 237–252.

[17] F. Corradini, C. Luciani, A. Morichetta, M. Piangerelli, A. Polini, Tlv-diss _ $\gamma$ $\gamma$: A dissimilarity measure for public administration process logs, in: Electronic Government: 20th IFIP WG 8.5 International Conference, EGOV 2021, Granada, Spain, September 7–9, 2021, Proceedings 20, Springer, 2021, pp. 301–314.

[18] C. Guasch, G. Lodi, S. V. Dooren, Semantic knowledge graphs for distributed data spaces: The public procurement pilot experience, in: International Semantic Web Conference, Springer, 2022, pp. 753–769.

[19] L. Asprino, E. Daga, A. Gangemi, P. Mulholland, Knowledge graph construction with a façade: a unified method to access heterogeneous data sources on the web, ACM Transactions on Internet Technology 23 (2023) 1–31.

[20] L. Feddoul, M. Raupach, F. Löffler, S. Babalou, J. Hoyer, M. Mauch, B. König-Ries, On which legal regulations is a public service based? fostering transparency in public administration by using knowledge graphs, Lecture Notes in Informatics (LNI) (2023).

[21] F. Daniel, K. Barkaoui, S. Dustdar, Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I, volume 99, Springer, 2012.

[22] A. Rosenberg, J. Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure, in: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 410–420.

[23] A. Struyf, M. Hubert, P. Rousseeuw, Clustering in an object-oriented environment, Journal of Statistical Software 1 (1997) 1–30.

[24] M. Shutaywi, N. N. Kachouie, Silhouette analysis for performance evaluation in machine learning with applications to clustering, Entropy 23 (2021) 759.

[25] V. Naderifar, S. Sahran, Z. Shukur, A review on conformance checking technique for the evaluation of process mining algorithms, TEM Journal 8 (2019) 1232.

[26] A. Berti, S. van Zelst, D. Schuster, Pm4py: a process mining library for python, Software Impacts 17 (2023) 100556.

[27] C. J. Turner, A. Tiwari, J. Mehnen, Mining process flowcharts from business data: An evolutionary approach, in: Proceedings of the 6th CIRP-Sponsored International Conference on Digital Enterprise Technology, Springer, 2010, pp. 1069–1087.