

# Activity Discovery Tool From Unstructured Data To Enhance Process Mining (Extended Abstract)

Marwa Elleuch<sup>1\*,†</sup>, Christophe Maillard<sup>1,†</sup>, Olivier Graille<sup>1,†</sup>, Sonia Laurent<sup>1,†</sup>,  
Oumaima Alaoui Ismaili<sup>1,†</sup> and Philippe Legay<sup>1,†</sup>

<sup>1</sup>Orange Labs, France

## Abstract

The free and unstructured textual records of process actors communications are nowadays not considered by the process mining tools. The confidentiality constraints of these records makes them difficult to be processed and integrated in process mining studies conducted on real data. This paper introduces the activity discovery tool which locally analysis, in unsupervised way, the communication records of a process actor (or a restricted set of process actors) to convert them into a structured log. This log could be shared to complete the partial view of process executions obtained from structured traces. We show, through a scenario example, how the results generated by this tool could enhance process mining.

## Keywords

Activity discovery, Unstructured textual records, Communication logs, Process mining

## 1. Introduction

Nowadays, process mining tools could be applied only on event logs having structured format. The free textual records that capture process actors interactions and communications were generally ignored if they are not converted into a structured format. However, such records are of big importance to enrich existing process knowledge or to discover new process fragments[1]. One of the main constraints for handling these unstructured records (e.g. emails, comments of incident tickets) is their confidentiality aspect. Taking the example of emails, process actors rarely agree to share the textual content of their emails to centralize their analysis. For some other types of free textual records, such as comments of incident tickets, the right of access and handling the records is generally restricted to a set of process actors. In fact, they are considered as sensitive data that could disclose the strategic aspect of an organism if they are largely shared. Therefore, it is not possible to process them outside the organism (as the case of the incident tickets) or the process actor machine (as the case of emails).

To handle these confidentiality restrictions (at individual or group of actors level), we propose in this paper ADT (the Activity Discovery Tool) that locally analysis the free textual commu-

---

*ICPM Doctoral Consortium and Demo Track 2023, ICPM 2023, Rome, Italy, September 23 - September 27, 2023*

\*Corresponding author.


†These authors contributed equally.

✉ marwa1.elleuch@orange.com (M. Elleuch); christophe.maillard@orange.com (C. Maillard);

olivier.graille@orange.com (O. Graille); sonia.laurent@orange.com (S. Laurent);

oumaima.alaouiismaili@orange.com (O. A. Ismaili); philippe.legay@orange.com (P. Legay)

ORCID 0000-0002-0877-7063 (M. Elleuch); 0009-0004-3028-3758 (O. A. Ismaili)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

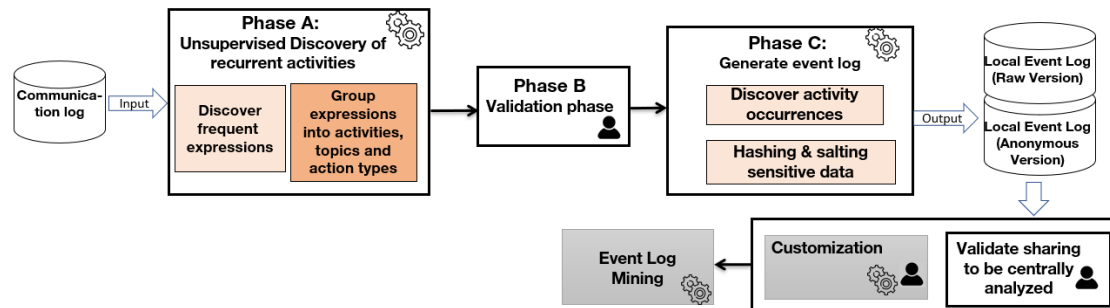
nication records of a process actor or a restricted group of actors. The tool implements and extends a recent work[2, 1]. It aims to reduce the textual records, in unsupervised way, into a structured event log reporting the relevant performed activities. These events are generated in the way that they could be shared for completing other traces of the same process (obtained from other information systems or other process actors) but without disclosing the confidential textual contents of the handled records. In what follows, we give an overview on the related work, describe the main functionalities of the tool, provide an example scenario related to the incident management process, discuss the maturity of the tool and conclude with future works.

## 2. Related Work

Some related works were mainly based on supervised approaches (e.g. [3]), which limits their potential to be applied in various scenarios. Tools that were designed in the same context allow employees at the most managing ongoing activities, e.g., by summarizing activities included in received emails [4] or displaying activity realization status [5]. A recent study [1] shows that the implemented solution in our tool, answers simultaneously several challenges comparing to existing works allowing richer event log that captures in addition to activity names; their speech acts, (ii) business data and (iii) several activities per textual segment.

## 3. Main Functionalities

ADT is an office application that allows process actors to analyse their communication records in order to reduce them into an event log. It ensures three main functionalities that we resume in Figure 1 (the functionalities colored in grey are to be ensured outside the tool):



**Figure 1:** Pipeline for event log mining using ADT

**A- Unsupervised discovery of recurrent activities:** It first discovers their recurrent activities by implementing and extending the approach proposed in [2]. Basically, it first discovers recurrent expressions that potentially reflects how business actors express their recurrent activities in their communication records. These expressions are then grouped into activities while considering: (i) rephrasing relations, and (ii) synonymy constraints to differentiate between those that are different and which could refer to contradictory actions. To facilitate their

exploration, the activities are then grouped into coarser topics and action types. Activities of the same action type (e.g. 'replace card', 'change fiber') share terms referring to the same coarser action (e.g. 'change', 'replace'). Activities grouped into the same topic (e.g. 'replace card', 'delete card') share terms referring to the same manipulated artifact (e.g. 'card').

**B- Validation phase:** It allows process actors to intervene after discovering activities to: (i) discard those that are judged confidential, and (ii) validate sharing the others.

**C- Generate anonymous event log:** The tool generates an event log to be shared according to the proposed structure in [1]. Each textual record is first reduced into the set of activity occurrences whose labels were validated (in terms of sharing). Each activity occurrence (i.e. event) is characterized by these attributes: activity name, activity speech act, business data, communication record attributes (i.e. ID, timestamp, sender, receivers and conversation ID), an action type and a topic. The tool offers the possibility to either access to such event log for further adaptation and customization (e.g. by business experts) or to its anonymous version. To obtain such anonymous version, sensitive data in each event (i.e. business data values, sender and receivers) are hashed (to guarantee that similar values could be mapped) and salted to complicate its cracking process. In this way, the textual content of the communication records are not shared. Only the relevant information w.r.t business processes are shared giving the possibility of being centrally analyzed and merged with (i) other event logs generated from the communication records of other process actors, or (ii) the structured part of the same records (e.g. IDs of incident tickets replacing the process instance information).

## 4. Scenario example

The scenario example is related to the incident managing process. We dispose the log capturing the comments exchanged, inside the incident tickets, between a restricted set of actor groups. This log is of two parts: (i) a structured part recording: the actor names sending comments, timestamps, human duration, and their ticket IDs and (ii) a non-structured part revealing the free textual content of the sent comments.

Using our tool, the comments of the concerned set of actors are analysed and reduced into the events recording the occurred activities. This log was then shared to be analysed by business experts in order to enrich the structured event log part and to inject it to [Celonis](#) as a process mining tool. We show in the [demo](#) how the additional attributes extracted from the unstructured log part enabled us to implement additional interfaces within Celonis for enriching: (i) the process actor perspective, and (ii) the filtering criteria to detect tickets containing incorrect activations of one actor.

1) Enrich the process actor perspective: At each actor activation, it becomes feasible to observe the detailed activities and generate a synthesis of the scope of the mentioned ones. This allows for the identification of instances where an action (i) was documented by an actor other than the one who performed it, or (ii) manipulating a material of a specific technical scope.

2) Enrich filtering criteria: Giving a process actor *expertGroup1* and other actors of different technical domains (i.e. A, B, C and D), the goal is to identify: (i) the tickets of incorrect activation of *expertGroup1* of domain A, C and D because they were resolved by actors of domain B, and (ii) the actors involved in such incorrect activation. Based only on the structured part of the tickets,

we could select those where *expertGroup1* was activated and domain B is the last activated compared to *expertGroup1* of other domains. However, the main constraints with such method, is that sometimes, actors of domain B are not explicitly activated in the tickets; they don't send comments, so they could not be detected from the set of senders. They are only reported in comments sent by other actors referring to their interventions with a corrective action. The [demo](#) shows how with the detected activities by ADT, it was possible to identify additional 50 tickets containing incorrect activations of *expertGroup1* (representing around 23% of the total tickets) . This helps us to: (i) identify more precisely the lost human time by *expertGroup1*, i.e. a total duration increased 2.5 times as the additional tickets contains longer tickets that could potentially correspond to anomalies of important calculated lost time, and (ii) implement precedence sequential constraints for detecting additional tickets where actors of domain B were involved in such incorrect activation (i.e. 49 tickets representing 34% of tickets validating such case).

## 5. Maturity and available resources

ADT is accessible in our organism for installation. The front-end is implemented with the Angular and Electron frameworks. The back-end is implemented in python. The implemented solution was validated in [1] using a public dataset of emails [Enron](#) where the performances are reported, and the obtained activities were publicly provided (i.e. see this [link](#)). We also conducted tests on other datasets like the incident ticket comments and the emails of the employees of our organism. We validated the results with business experts that reported that the major advantage of the tool is its ability to generate first results in unsupervised way (which means without human intervention) able to be interpretable and adapted to enrich other event logs. We provide the following elements:

- A [documentation](#) explaining how the tool is installed and run within our organism. However, we were not able to provide the access for external collaborators.
- A public [video](#) illustrating how our tool serves the described scenario example (Section 4).
- A [guide](#) to access the implemented Celonis interfaces.

## 6. Conclusion and future work

In this paper, we presented ADT wick analyses the free communication textual records of business actors to enrich event logs for process mining. We intend to leverage the studied use cases to communicate across all directions within our organism and visually demonstrate potential gains to enhance collective efficiency in other use cases. We aim to assess the extent to which these efforts are replicable to other processes and customize the developed tool to make it increasingly versatile whenever needed. In future works, we aim to cover various communication records types. Additionally, we aim to investigate the following points:

- Improve the anonymization functionality to consider the privacy risk of sharing sequence of events rather than only individual events [6]. This is by allowing users to check the

sequence of the occurred activities to edit confidential sub-sequences that does not seem sensitive when looking at individual activities.

- Extend the format of the generated event log to support recent format, mainly the Object-Centric Event Log (OCEL) [7].
- Study how the generative AI could enhance ADT performances. In fact, with the actual publicly available models, a large resources in terms of RAM and GPU (e.g. 80 GB for MOSAIC ML MPT30 B and at least 30 GB for LLAMA2 70B after quantization) is needed. This makes their integration in our tool as office application not feasible. Getting confidential data out of the user's machine to be processed in an external data center (as the case of chatGPT) is also unfeasible, as explained before.

## Acknowledgments

We would like to thank Alain Bouchard, David Menchi, Frédéric Bastard and Marjorie Deshayes: the experts in the studied process, for their invaluable assistance, the time they devoted to addressing our inquiries, for testing the tool we made available to them, and for placing their trust in us. This collaborative effort was instrumental in achieving promising results.

## References

- [1] M. Elleuch, O. Alaoui Ismaili, N. Laga, W. Gaaloul, Process fragments discovery from emails: Functional, data and behavioral perspectives discovery, *Information Systems (2023)* 102229.
- [2] M. Elleuch, O. Alaoui Ismaili, N. Laga, W. Gaaloul, B. Benatallah, Discovering activities from emails based on pattern discovery approach, in: *Business Process Management Forum: BPM Forum 2020*, Seville, Spain, September 13–18, 2020, *Proceedings 18*, Springer, 2020, pp. 88–104.
- [3] C. Kecht, A. Egger, W. Kratsch, M. Röglinger, Event log construction from customer service conversations using natural language inference, in: *2021 3rd International Conference on Process Mining (ICPM)*, IEEE, 2021, pp. 144–151.
- [4] S. Corston-Oliver, E. Ringger, M. Gamon, R. Campbell, Task-focused summarization of email, in: *Text Summarization Branches Out*, 2004, pp. 43–50.
- [5] M. Dredze, T. Lau, N. Kushmerick, Automatically classifying emails into activities, in: *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006, pp. 70–77.
- [6] G. Elkoumy, M. Dumas, Libra: High-utility anonymization of event logs for process mining via subsampling, in: *2022 4th International Conference on Process Mining (ICPM)*, IEEE, 2022, pp. 144–151.
- [7] A. F. Ghahfarokhi, G. Park, A. Berti, W. M. van der Aalst, Ocel: A standard for object-centric event logs, in: *European Conference on Advances in Databases and Information Systems*, Springer, 2021, pp. 169–175.