Supporting software evolution actions with process mining

Daniel Calegari[1], **Andrea Delgado**[1]

[1]Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay

{dcalegar,adelgado}@fing.edu.uy

Abstract. *There are several forces driving software evolution. One is the business process (BP) misalignment, i.e., when the behavior of the information systems supporting the BPs, or its users, is not aligned with the intended behavior of the BPs identified during the requirement engineering phase. Process Mining (PM) is an essential strategy for BP alignment evaluation. Nevertheless, PM initiatives do not usually focus on connecting BP alignment misfits with concrete software requirements for software evolution. This paper provides initial insights into how PM can support software evolution actions by considering research questions posed during a PM initiative. We exemplify this idea by analyzing administrative procedures within an Electronic Document Management System, providing action guides for its evolution obtained from the PM initiative.*

1. Introduction

Business Processes (BPs) are at the center of organizations' daily operation, supported by a combination of traditional Information Systems (IS) and Process-Aware Information Systems (PAIS) [Dumas et al. 2005]. Many different internal and environmental forces require the software to be corrected, adapted, and improved, guiding its evolution [Sommerville 2016]. One is BP misalignment, i.e., when the actual behavior of the information systems supporting the BPs, or its users, is not aligned with the intended behavior of the BPs identified during the requirement engineering phase.

BPs life-cycle [Weske 2019] is closely linked with the software engineering life-cycle since there must be an evaluation phase measuring business alignment [Aversano et al. 2012, Rabelo et al. 2019] pushing process improvement and the corresponding software evolution. Alignment between the BPs and the supporting software systems requires comparing the actual behavior of an information system or its users with the intended behavior of the BPs identified during the requirements phase [van der Aalst 2005]. BPs could be not entirely aligned, e.g., their execution does not comply with performance requirements, there were environmental changes (e.g., policies and laws) that are not considered, or there is evidence of unexpected behaviors caused by a flexible execution of the BP or an incorrect interpretation of the process during their implementation. Depending on the source of the misalignment, the BPs need to be redesigned, and/or the supporting system needs to evolve.

Process Mining (PM) ([van der Aalst 2016]) is an essential strategy for BPs alignment evaluation. It allows for analyzing the records (logs) of events associated with the execution of processes in information systems. There are several techniques to discover corresponding business process models (process discovery) from the recorded events and verify the correspondence of the enacted BPs concerning the defined one (conformance checking). It also allows obtaining measures such as the duration of the processes, bottlenecks, or the underuse of resources, among others.

There are also methodologies, e.g., [Eck, van et al. 2015, Delgado et al. 2021], to guide the execution of PM initiatives (projects), defining activities, roles, and artifacts. They usually start with a planning stage for setting up the initiative with the definition of research questions that guide the mining and analysis activities. These methodologies also identify an evaluation stage that relates analysis findings to improvement opportunities. Although many possible use cases exist for PM [Milani et al. 2022], they focus on identifying BP alignment misfits. They do not focus on connecting problems that arise during the execution of the PM methodology with concrete requirements of the software systems supporting the BPs that must be addressed during software evolution. Moreover, since these research questions are separate from the requirement engineering stage, some business requirements for evaluation are only sometimes considered during software construction. Even if there is a perfect fit in BPs alignment, how data is recorded may obstruct or prevent answering the research questions. For example, the supporting systems register the activities performed but not the responsible users; thus, the process can be (mostly) enacted as expected. However, it is necessary to include data about who performs the activities to perform resource analysis.

This paper explores how PM can support software evolution actions such that improvements from these actions complement BP alignment needs. We define a general schema that connects research questions posed during a PM initiative with software evolution actions. We explore the feasibility of the schema by performing a PM initiative to an Electronic Document Management System (EDMS) [ISO 2016] providing such action guides for software evolution. In particular, we perform an in-depth analysis of open-access information on administrative procedures from our university. As a complementary objective of this initiative, we identify challenges for applying PM in EDMS.

The rest of the paper is structured as follows. In Section 2, we analyze some related work. In Section 3, we present the approach we followed. In Section 4, we present the PM initiative, and in Section 5, we discuss the initiative concerning the general approach and identify some challenges in applying PM for EDMS. Finally, in Section 6, we provide conclusions and an outline of future work.

2. Related Work

PM has many applications to software engineering [Keith and Vega 2017]. Most of them are focused on what is called Software Process Intelligence (SPI), which is the application of mining and analysis to software processes. The information needed for applying PM comes from software repositories such as version control systems, bug trackers, and mail archives. Using this information, the data can be extracted, combined [Poncin et al. 2011], and integrated with the perceptions of stakeholders [Vavpotic et al. 2022] to determine deviations and take corrective actions or provide improvement recommendations for the whole process. Generally, it is possible to study how to use PM to monitor the development process to demonstrate adherence to methodologies, rules, regulations, guidelines, or best practices [Bala and Mendling 2018]. Moreover, it is possible to apply the same idea to concrete software engineering disciplines, e.g., testing [Hernandez-Resendiz et al. 2023] and maintenance [Gupta et al. 2017].

The alignment between business requirements and software systems is also an active research line [Aversano et al. 2012, Rabelo et al. 2019]. In [Aversano et al. 2016], the authors present a general approach for monitoring and managing the alignment level. It comprises a modeling stage for BPs and software systems, an evaluation phase for measuring the alignment level between them, and a final step for identifying and executing the evolution actions. This approach is based on the existence of the previous modeling of the software systems and a mapping between software and business entities, which can be restrictive. In [van der Aalst 2005], PM is used for measuring business alignment based on predefined process models and discovered models that are compared using delta and conformance analysis, similar to the works about PM application to software engineering described before.

Works applying PM techniques to EDMS focus on extracting information from the system's database rather than identifying challenges. In [Berti et al. 2021], the authors propose an approach to guide and ease the extraction of event logs from SAP ERP, which, in a broader sense, can be seen as an EDMS. In [Osman and Ghiran 2019], the authors exemplify the application of PM as part of Industry 4.0 by analyzing an event log describing a document management process. In [Repta et al. 2018], the authors analyze processes related to e-government systems, focusing on the discovery and extraction of document flows. Finally, in [Markowski and Przybylek 2016], the authors present an example of the application of PM to analyze the loss of documents in a document flow.

3. Connecting a PM initiative with software evolution actions

Any PM initiative tends to analyze the actual execution of BPs within traditional IS or PAIS, supporting them for the organization's daily operation. We claim that from these initiatives, it is possible to discover BP alignment misfits and determine concrete actions for software evolution. This section presents the general idea of linking a PM initiative to identifying software evolution actions and a proposal to guide such identification.

3.1. PM initiatives for software evolution

Although there are many PM methodologies to guide a PM initiative, we take PRICED [Delgado et al. 2021] as a reference since we have applied it in several PM efforts before and in the EDMS example we present here. It provides guidance and support for organizational data science projects, i.e., from a traditional process mining project to a more complex project requiring integrating process data, and organizational data [Calegari et al. 2021]. From a very abstract perspective, as depicted in Figure 1, the PRICED methodology defines four phases. The **Enactment** phase involves the actual execution of processes within the supporting software systems. The **Data** phase involves determining the research questions that will lead the mining effort and extracting, integrating, and loading data to build event logs which are the primary source of information to analyze. An event log consists of cases (executions) of a process. For each case, there is a partially ordered collection of events that describe such execution, each with possibly many attributes describing the event and its environment (e.g., resources involved, timestamps, etc.) The **Mining/Analysis** phase focuses on executing the analysis and providing many outputs, such as process and analytic models, which are evaluated to provide findings for the organization. Finally, the **Improvement** phase corresponds to the organization's improvement efforts after the analysis.

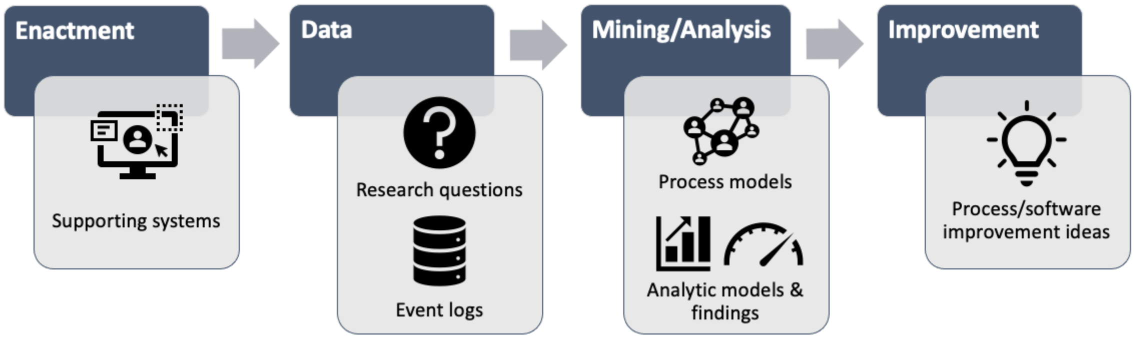


Figure 1. An overview of the PRICED methodology

Within the Improvement phase, there are two different kinds of improvement ideas to consider. On the one hand, process improvement ideas come from identifying BP alignment misfits, e.g., some activities can be improved by assigning more resources, or the process can be uniformly performed by avoiding undesirable paths. On the other hand, software evolution actions can be taken indirectly from process improvement ideas, i.e., the software system needs to evolve to better align with the business processes it supports.

In this paper, as a second kind of software improvement idea, we connect problems that arise during the execution of the PM methodology with concrete software requirements that must be addressed during the evolution of the supporting systems. A PM initiative could be used to ensure that the software provides the organization with the answers it needs for evaluating its operations.

3.2. Identifying software evolution actions

Based on the motivation presented above, we defined an approach to connect PM initiatives with software evolution actions as another result of the PM effort. As depicted in Figure 2, we consider three dimensions: the purposes guiding a PM initiative, the perspectives addressed by the research questions, and the existing challenges and guidelines for building an event log.

The **Purpose dimension** was taken from [Milani et al. 2022], in which PM use cases are classified, as well as the business questions they might address. The authors define the following categories, each with specific questions to be addressed: *Transparency* focuses on discovering process perspectives: process models, resource interactions, decision rules, etc.; *Efficiency* focuses on performance analysis; *Quality* focuses on process variants and deviation analysis; *Compliance* focuses on conformance checking and compliance monitoring; and *Agility* focuses on predictive monitoring, concept drift, etc.

The **Perspective dimension** was taken from [van der Aalst 2016], in which the following perspectives are defined: *Control-flow* focuses on finding a good characterization of all possible paths within a process; *Organizational* focuses on the resources enacting the process and their relations; *Case* focuses on the data that characterizes process instances, i.e., organizational data related to each event; and *Time* focuses on analyzing the timing and frequency of events, e.g., on discovering bottlenecks.

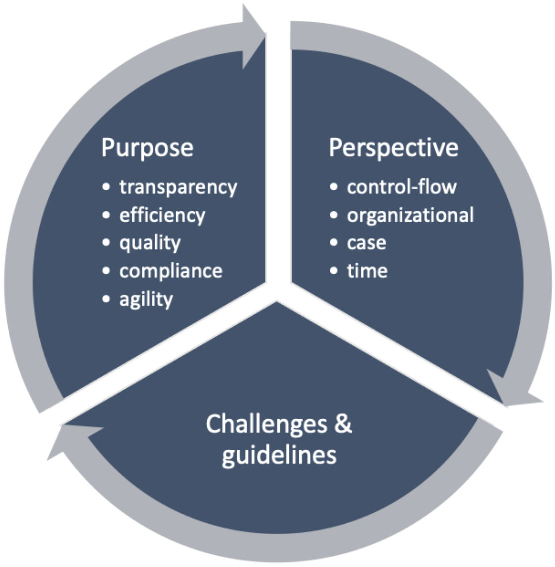


Figure 2. Dimensions considered within a PM initiative

There are also **Challenges & Guidelines** that apply to an event log [van der Aalst 2016]. Essential aspects to consider are: *events* need to be grouped and (partially) ordered per case. They also need to be all at the same level of granularity; Every event needs at least two *attributes*: a name and a timestamp that partially orders them within a case. There can be other attributes of interest, e.g., the resource causing the event or any other type of data (costs, risk, age, etc.); *attribute values* must have clear semantics and be structured, precise, and stable within a log.

The three dimensions depicted in Figure 2 could be overlapped sometimes (e.g., the efficiency purpose is related to the time perspective) and are non-exhaustive. Nevertheless, they provide an initial approach to identifying software improvement (evolution) actions from a PM initiative. Figure 3 depicts our proposed general schema. Once the initiative's **purpose** is defined, many research **questions** must be addressed, and each question involves a particular **perspective** related to a set of guidelines. During the PM initiative, some **challenges** could arise related to the perspective involved, thus inducing some guidelines to consider. These guidelines can be mapped to specific software evolution **actions** by looking at the former software requirements and how they were addressed within the software system. Sometimes different challenges can be connected to the same guideline, and different guidelines could be related to the same improvement action.

For example, a transparency initiative could be focused on the organizational perspective. In such a case, as described in [Milani et al. 2022], specific research questions can be addressed, e.g., what are the relationships among the resources involved in a process? This research question leads to specific log requirements (guidelines), e.g., events must have an attribute describing the resource involved with the event, which also needs to be appropriately connected with the organizational structure. During the PM initiative, we can find some events that need more information about the people performing them or the role people have within the organization. It identifies how the organizational structure is managed and which software requirements involve associating these events with concrete people/roles allowed to perform certain activities in the BP.

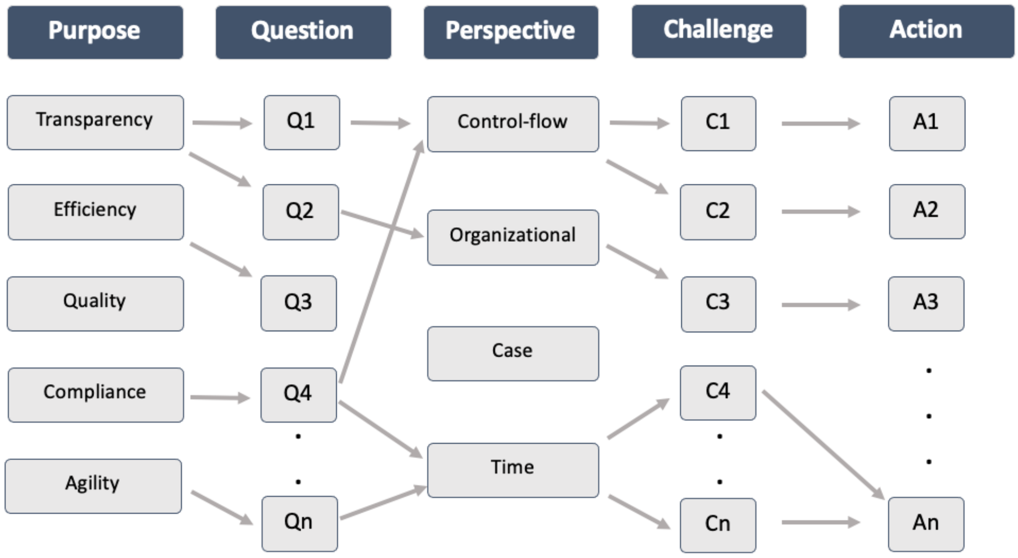


Figure 3. Schema for identifying improvement actions from a PM initiative

4. PM initiative example: EXPE+

To assess the general schema depicted in Figure 3, we performed an exploratory PM initiative following the PRICED methodology for analyzing the EXPE+[1] system. In what follows, we present the system and the main results of the Data and the Mining/Analysis phases of the PRICED methodology application.

4.1. EXPE+

The EXPE+ system is an Electronic Document Management System (EDMS) used at Universidad de la República, Uruguay's main public university. EXPE+ manages the electronic creation, storage, and control of documents, supporting the organization's processes. It is used to carry out every administrative procedure helping to organize and exchange data in a distributed environment with more than six thousand administrative people and a hundred administrative areas within their twenty decentralized schools.

EXPE+ defines how records need to be expressed to trace an administrative procedure and what information can be accessed as public. Every administrative procedure has a physical document attached to new information at each step. The system registers primary data used to carry out the administrative procedure, e.g., the step that has to be done, the person and organizational section that performs it, and when it is done. The procedures follow flexible but optional guidelines, i.e., a predefined process may need to be carried out as intended. It could generate deviations in the processes and the impossibility of objectively analyzing their execution.

An open-access web application enables tracking of the stage of the process in which an administrative procedure is. The application shows the cover of an administrative procedure, e.g., the one depicted in Figure 4, which provides the primary information about it, e.g., the procedure ID, the kind of procedure, and a summary of it, and the data of each step that has been enacted, as well as the active step.

[1]EXPE+. `http://www.expe.edu.uy/`

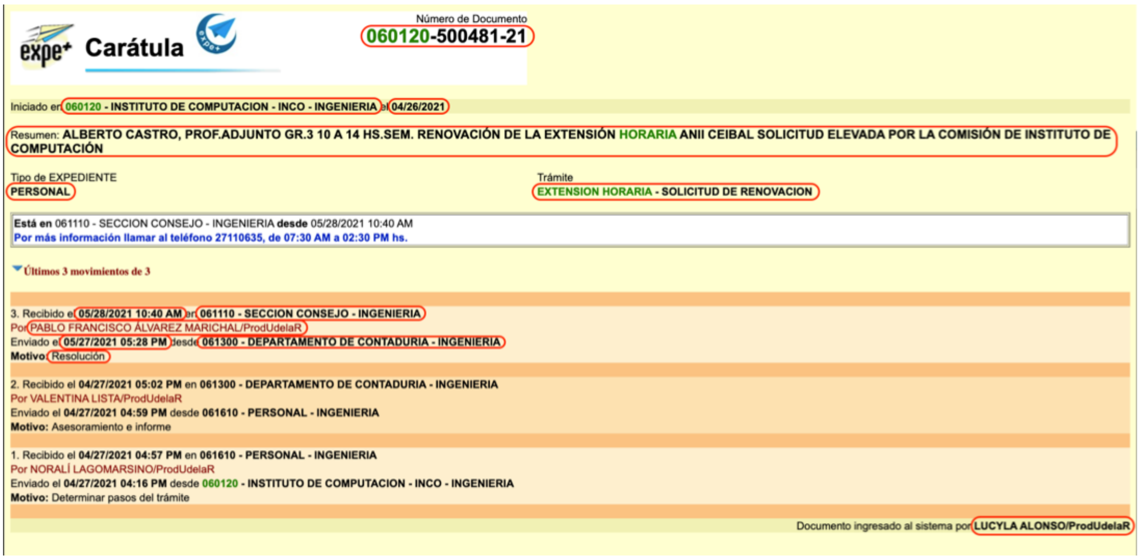


Figure 4. EXPE+: exampling cover of an administrative procedure

4.2. Data phase

We selected well-known processes for us: the "Extension of working hours" and "Reduction of working hours" for teachers within the Computer Science Institute (InCo). The processes are commonly used to extend/reduce working hours, in the first case mainly to participate in research projects or to teach extra courses. The university has a theoretical high-level base procedure specified in the text for those processes. These routes should be followed within the EXPE+ system, and one of the critical interests of the analysis is to be able to contrast the actual process execution with the defined procedures.

- Procedure: *"Extension of working hours"*
  - *Step 1:* Organizational unit request with the consent of the interested party
  - *Step 2:* Human Resources: position report
  - *Step 3:* Accounting: availability report
  - *Step 4:* Faculty Council: resolution
  - *Step 5:* Human Resources: Registration, notification, and archiving
  - *Step 6:* Secretary: Archive
- Procedure: *"Reduction of working hours"*
  - *Step 1:* Interested party request with the endorsement of the requesting organizational unit
  - *Step 2:* Human Resources: position report and preventive reserve
  - *Step 3:* Accounting: preventive confirmation
  - *Step 4:* Faculty Council: resolution
  - *Step 5:* Human Resources: Registration, notification, and archiving
  - *Step 6:* Secretary: Archive

We carry out a process mining initiative answering basic questions of any initiative [van der Aalst 2015], which is focused on detecting and analyzing deviations regarding the control-flow and throughput time. These basic questions are:

(RQ1) *Are the processes executing as expected regarding the defined base procedures?*
(RQ2) *Which is the average/minimum/maximum throughput time of the processes?*

The website of the EXPE+ system was identified as the primary data source since the data published there is the public one corresponding to the steps (activities) carried out within the procedures, and it is directly connected to its database. The supporting database provides no additional information for the study. We used web scraping to extract the data from the pages and store it in a database for further manipulation. In this way, we could evaluate the web application and perform the PM initiative without adding the risk of delays in access to information. We recovered processes from 2000 to 2021, totaling 2072 cases containing 12513 events. We automatically generated the event log as a CSV sheet which can be used directly as input in PM tools such as Disco[2].

Mapping events as activities was indirect since the EXPE+ names for the activities are unrelated to the specific process. They are generic for all processes in the system and refer to the movement's objective from one organizational unit to the next, i.e., what the receiving organizational unit should do. For example, when the aim is to resolve, the activity following the transition will be named "Resolution", or when the objective is to analyze and report regarding the procedure's content, the activity's name following the transition will be "Advice and report". Names are selected from a predefined list of options.Using only the activity names to identify the events leads to erroneous loops. Although the names are repeated, they are not the same activity as the origin, and the target organizational unit is different; hence the correct path in the process should not be a loop but moving forward. After analyzing different combinations for the identification of activities, we identified the name activity + target organizational unit (i.e., the one executing the activity) as the one that leads to the expected paths.

We analyzed the event log concerning quality characteristics and improved it accordingly. Each attribute was scanned to check whether it did not present a null value. In particular, we found 22 cases with null values in the sending date, which were solved using the same date as the reception date for the next movement. Regarding the names of the activities, although they are currently available in a list to choose from, some previous records had the names in uppercase, lowercase, or misspelled. To solve this, all the names were standardized to use uppercases. We also checked the timestamps of the registered dates, which in the EXPE+ extend up to minutes. It is not the best granularity to allow exact calculation of waiting and processing times. Still, in this context, it should be sufficient.

4.3. Mining/Analysis phase

We analyzed and filtered the data in the log to get a consistent one to apply the PM tasks. Table 1 presents descriptive statistics regarding the raw and filtered data.

The "Extension of working hours" log contains 1766 cases and 10785 events from 10/10/2000 to 19/11/2021, with a median duration of 90.8 days and a mean of 142.1 days (20.3 weeks) and 140 activities. As expected, most cases have 5-6 events (1112 representing 63% of the cases), with a maximum of 19 events and a minimum of 2 events per case. After filtering incomplete cases (i.e., ongoing and canceled), we get 817 cases (46%) and 4380 (40%) of the events, with a median duration of 88.1 days and a mean of 106.4 days (15.2 weeks).

[2]Fluxicon Disco. https://fluxicon.com/disco/

Table 1. Descriptive statistics of the raw and filtered event logs

Name	#traces raw	#traces filtered	#events raw	#events filtered	#steps raw	#steps filtered
Extension of working hours	1766	817	10785	4380	140	21
Reduction of working hours	201	94	1238	510	111	22

The “Reduction of working hours” process event log contains 201 cases with 1238 events, from 02/03/2001 to 15/11/2021, with a median duration of 84.8 days and a mean duration of 142.1 days (20.3 weeks), and activities. Also, as expected, most cases have 5-6 events (118 cases representing 59% of the cases), with a maximum of 13 events and a minimum of 2 events per case. After filtering incomplete cases (i.e., ongoing and canceled), we get 94 cases (46%) and 510 events (41%), with a median duration of 83.9 days and a mean of 90.9 days.

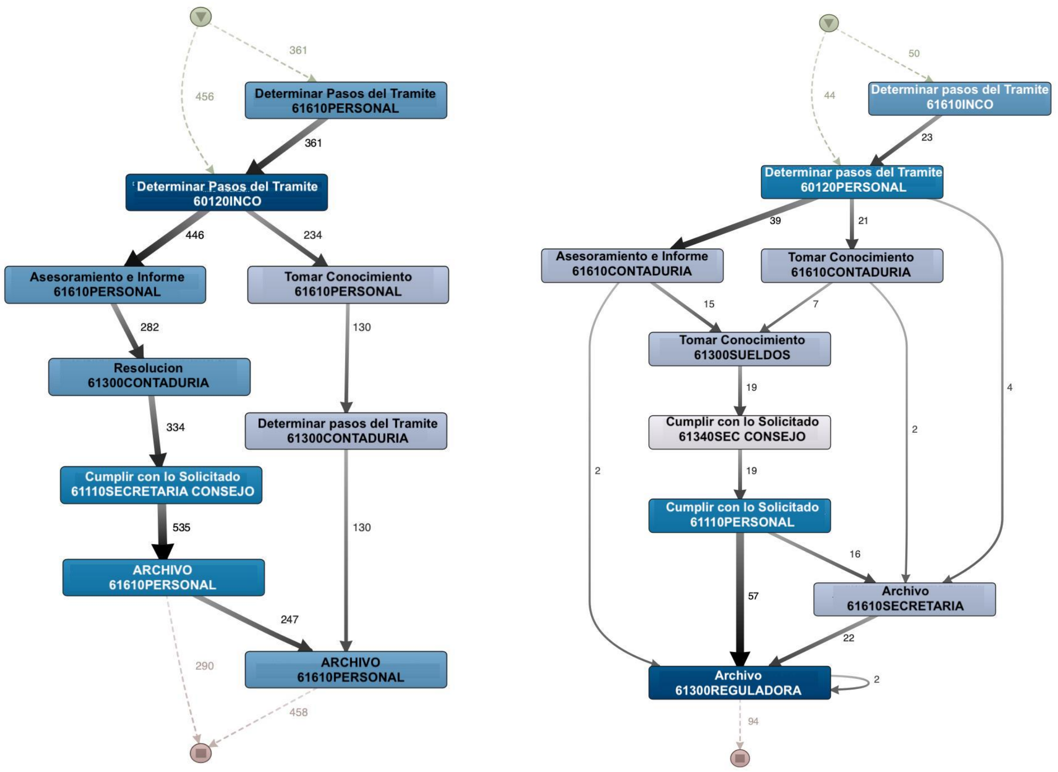
Figure 5 depicts the general paths of the filtered event logs. It can be seen that the two start activities selected correspond to the organizational units InCo (Instituto de Computación-InCo-Ingeniería) and Human Resources (HR, Personal). In both processes, the ones initiated in HR correspond to expiration notices of an already previous Extension granted. These are sent to the InCo to inform if the renewal should be processed. The ones passed through the organizational unit Secretary of the Faculty Council was renewed.

To answer the first question of interest, whether the processes are executing as expected when comparing the process models’ general paths with the procedures defined, it can be easily seen that the real execution of the processes added several activities to the ones planned. Mainly, there are several variants in which cases go through no needed activities and are returned to where they were before to continue within the correct paths. It is caused by participant mistakes since there is a free route option in which an underlying process model does not predefine the activities. Although the most frequent path is similar to the one defined for the procedure, it only represents around 7% of the cases.

We detected five organizational units enacting the process not described in the procedure defined. It probably happens since the ones named in the procedure are the high-level ones, i.e., Accounting, but not the sub-units that are part of them, e.g., Salaries, or the Secretary of the Faculty Council, when only the Faculty Council is named.

To answer the second question regarding the throughput time of the processes (using the filtered log), both processes take several months, but their duration is around the expected times. The Extension process has a median duration of 88.1 days and a mean of 106.4 days (15.2 weeks) with a maximum of 3 years and 265 days and a minimum of 9 days 23 hours. The Reduction process has a median duration of 83.9 days and a mean of 90.9 days, with a maximum of 252 days and a minimum of 19 days and 7 hours.

Reviewing the maximum durations for the unfiltered log, we found that in most cases with long duration, the final archiving activity was missing and added several years later, particularly in 2017, probably due to reviews of the system data. The cases with minimum duration were ongoing cases that we filtered out when adding the endpoint filter to keep only complete cases. Both observations also hold for the Reduction process.



(a) Extension of working hours **(b) Reduction of working hours**

Figure 5. Process models general paths from the filtered event logs

5. Discussion & Challenges

Using the PM initiative, we can now discuss how PM can help support software evolution as well as identify challenges for applying PM in EDMS.

5.1. PM for software evolution

Figure 6 summarizes the application of the schema described in Section 3. The initiative had transparency, quality, compliance, and efficiency purposes. From the two main research questions, some other specific ones arise, according to [Milani et al. 2022], involving the control-flow, organizational, and time perspectives, e.g.,

The alternatives can be accumulative. For example, action (A1) involves three accumulative possibilities. As a minimum requirement, activity names must be refined (A1.a). Moreover, the list of predefined activity names in EXPE+ could be extended. A filtering option could be implemented to provide only the list of activities directly associated with the corresponding process (A1.b). The specific activity names for each process can be identified from the theoretical description of the process or the business people involved in each unit. A third and more advanced alternative, while keeping the free routing capabilities of the EDMS, is to incorporate functionality to suggest the following activity to be performed (A1.c), considering the prescribed process or even a software component for predictive analysis that can also be based on PM strategies. Moreover, the alternatives can be opposite or considered exclusive. For example, action (A2) involves

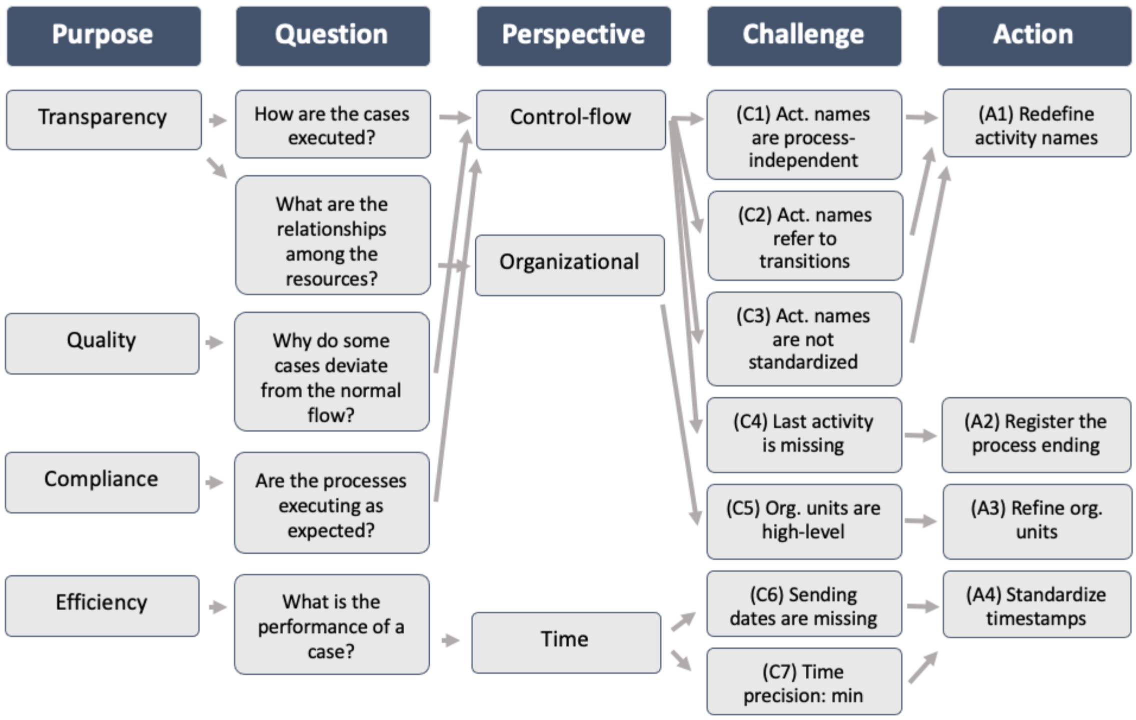


Figure 6. Application of the schema to the PM initiative

two potentially exclusive possibilities. The first (A2.a) consists of locating the concrete requirement that ends the process and changing their implementation for registering the last activity. The second one (A2.b) involves keeping the implementation as is and adding an automatic mechanism to identify "suspended" cases and mark them as ended.

Actions could also imply changing the architecture, e.g., adding new or third-party software components, such as an LDAP (A3.a) or a logging mechanism (A4.b). They could also be considered recommendations and not mandatory requirements. For example, action (A4) implies refining the precision of timestamps which is optional since the way timestamps are registered is sufficient for analyzing the selected procedures.

From the organization's perspective, the EXPE+ software system complies with their former requirements, e.g., it allows managing administrative processes, provides customization capabilities for adding new users and roles, etc. We are not evaluating the functional alignment level between the BP and the supporting software systems. On the contrary, we are adding new requirements about how to register data (e.g., activity names), technologies to incorporate (e.g., LDAP), and additional functionalities (e.g., listing potential routing destinations). The potential actions to perform could vary depending on the software system supporting the business process. Moreover, no BP reengineering is needed from a business perspective. Nevertheless, we can identify ideas that lead to updating or refining the processes and their available information, e.g., the procedures could be updated to reflect the paths and participants in more detail.

The initiative was carried out with missing or incorrect data, as usual when working with traditional systems. We assumed that building the log is performed considering the best practices, so if something obstructs the initiative, it can be linked to some software weaknesses. In this sense, we can detect software improvement ideas.

Table 2. Challenges identified within the PM initiative EXPE+

Id	Challenge	Description
(C1)	Activity names are process-independent	Names are generic for all processes in the system.
(C2)	Activity names refer to transitions	Names refer to the objective of the transition from one unit to the next.
(C3)	Activity names are not standardized	Although they are available in a list to choose from, some records had the names in uppercase, lowercase, or misspelled.
(C4)	Last activity is missing	The end archiving activity was missing and added several years later.
(C5)	Organizational units are high-level	Organizational units referred are the high-level ones, but not the sub-units that are part of them.
(C6)	Sending dates are missing	There are null values in the sending date.
(C7)	Time precision: minutes	Timestamps extend up to minutes

Table 3. Software evolution actions identified within the PM initiative EXPE+

Id	Action	Description
(A1)	Redefine activity names (C1, C2, C3)	(a) Rename activities in a standardized process-specific way; (b) extend the list of predefined activity names and filter them w.r.t. each process; (c) suggest the following activity to be performed.
(A2)	Register the process ending (C4)	(a) Locate the requirement ending the process and registering the activity there; (b) periodically check and finish "suspended" cases that are waiting to be completed.
(A3)	Refine organizational units (C5)	(a) Include sub-units, and relate them to people, e.g., use an LDAP to store the organization information; (b) list only appropriate sub-units to be selected as the routing destination of an activity.
(A4)	Standardize timestamps (C6, C7)	(a) Change the precision of timestamps to include seconds; (b) log timestamps mandatory for every system's action.

5.2. Challenges for PM in EDMS

EDMS are generally integrated by several software modules, including a workflow for routing administrative procedures and a web interface that allows publishing public information about the administrative procedure to ease inquiries from interested parties. Workflow technologies have advanced in the last decades to complete BPMS solutions. , including introducing the BPMN 2.0 standard for enacting BPs. However, these systems mostly use limited capacities for process enactment and allow working with free routes, even defining ad-hoc paths within the process based on specifying the movement and the target organizational unit. It leads to several problems in the control-flow, such as errors when routing procedures to an organizational unit that is not the correct one, or selecting the wrong type of movement, adding unnecessary complexity and duration to the process. It could be helpful to use a predefined most frequent path as the preferred one for execution, leaving the free routing option for exceptional cases.

We can discuss the experience from the perspective of [Burattin 2015] about the most common problems regarding the applicability of PM in real-world environments. Although our findings cannot be generalized to other EDMS, it could be helpful to consider them as potential scenarios in similar initiatives.

- **P-01 Incompleteness of information**. There were several issues regarding the representation of elements on the website. The system is from 2000, and there were no easy ways to identify the elements within the web pages. Also, the back-end system for handling the procedures changed in 2020. Additionally, some cases had extra events not listed on the main page, for which other links had to be followed and parsed to get all the events for the cases.
- **P-02 Exploiting as much information as possible**. The EXPE+ database contains only a little more information than the one accessed through the website. For example, there is no information about each step's start and end timestamps, only the timestamp of when the procedure is sent to another administrative unit. Moreover, there are no attached documents with concrete information on the administrative procedure, approvals, etc., since they are stored on paper. However, this information could be digitally stored using standardized formats, not just PDFs. Explaining this information in a broader context could be possible, e.g., integrating process and organizational data using the PRICED methodology.
- **P-03 Difficulties in using tools and algorithms**. It could be interesting to validate the comprehension of the results within the same organization and analyze the possibility that non-expert users perform some process mining analysis.
- **P-04 Results interpretation**. Following the line above, we should validate it. The processes carried out in EXPE+ are linear, and there is a training of the experts in carrying out the processes (as part of their test competitions), so the interpretation of the models would not be a challenge.
- **P-05 Computational power and storage capacity required**. The scrapping process and the storage of logs were fine in EXPE+. The processes are usually short, and less information is stored. However, as we see in e-gov environments, the logs are larger and more complicated in larger organizations.

6. Conclusions

This paper explored how PM can support software evolution by connecting the research questions posed during a PM initiative with software evolution actions. We also essayed the general schema through the analysis of administrative procedures registered within the EXPE+ system of our University. We need further assessment for validating and identifying limitations of the general schema since we just performed a promising but initial experience. To improve the schema, we need to strengthen the connection between the challenges and the evolution actions, e.g., by expressing the actions in the form of concrete software requirements and using them for performing feasibility and cost analysis. We can also apply the idea to software testing, in which artificially created data could address an idealized PM initiative.

As a complementary objective of the PM initiative, we also identified challenges for applying PM in EDMS. We discussed some findings, such as the problems gathering information from basic (and old) web technologies and the free routing that allows defining where the procedure should go next. Applying PM techniques to EDMS is crucial in checking the operation of the administrative procedures with the predefined or desired one. Since the EXPE+ is being partially replaced by a new solution looking for complete digitalization of the administrative procedures, our analysis provides valuable input for implementing the new system to enable the analysis of processes and improve the University's operations. As ongoing work, we are currently validating the results with business experts to determine the contribution the inputs can generate.

7. Acknowledgements

We want to thank the undergraduate students who worked on the practical application: Ignacio Jorge and Nicolás Menna.

References

Aversano, L., Grasso, C., and Tortorella, M. (2012). A literature review of business/it alignment strategies. *Procedia Technology*, 5:462–474. 4th Conf. of ENTERprise Information Systems, aligning technology, organizations and people (CENTERIS).

Aversano, L., Grasso, C., and Tortorella, M. (2016). Managing the alignment between business processes and software systems. *Information and Software Technology*, 72:171–188.

Bala, S. and Mendling, J. (2018). Monitoring the software development process with process mining. In *Business Modeling and Software Design*, pages 432–442. Springer.

Berti, A., Park, G., Rafiei, M., and van der Aalst, W. M. P. (2021). An event data extraction approach from SAP ERP for process mining. In *Process Mining Workshops - ICPM 2021 Intl. Workshops, Revised Selected Papers*, volume 433 of *LNBIP*, pages 255–267. Springer.

Burattin, A. (2015). Obstacles to applying process mining in practice. In *Process Mining Techniques in Business Environments: Theoretical Aspects, Algorithms, Techniques and Open Challenges in Process Mining*, pages 59–63. Springer.

Calegari, D., Delgado, A., Artus, A., and Borges, A. (2021). Integration of business process and organizational data for evidence-based business intelligence. *CLEI Electron. J.*, 24(2).

Delgado, A., Calegari, D., Marotta, A., González, L., and Tansini, L. (2021). A methodology for organizational data science towards evidence-based process improvement. In *Software Technologies - 16th Intl. Conf., ICSOFT*, volume 1622 of *Communications in Computer and Information Science*, pages 41–66. Springer.

Dumas, M., van der Aalst, W. M., and ter Hofstede, A. H. (2005). *Process-Aware Information Systems: Bridging People and Software through Process Technology*. John Wiley & Sons, Inc.

Eck, van, M., Lu, X., Leemans, S., and Aalst, van der, W. (2015). PM2 : a process mining project methodology. In *Advanced Inf. Systems Engineering: 27th Intl. Conf., CAiSE 2015*, LNCS, pages 297–313. Springer.

Gupta, M., Serebrenik, A., and Jalote, P. (2017). Improving software maintenance using process mining and predictive analytics. In *IEEE Intl. Conf. on Software Maintenance and Evolution (ICSME)*, pages 681–686.

Hernandez-Resendiz, J. D., Ramirez-Alcocer, U. M., and Tello-Leal, E. (2023). *An Approach Based on Process Mining Techniques to Support Software Development*, pages 25–49. Springer.

ISO (2016). ISO 15489-1:2016 information and documentation — records management. Technical report, ISO/TC 46/SC 11.

Keith, B. and Vega, V. (2017). Process mining applications in software engineering. In *Trends and Applications in Software Engineering*, pages 47–56. Springer.

Markowski, P. and Przybylek, M. R. (2016). Process mining methodology in industrial environment: Document flow analysis. In *Proc. of the 2016 Federated Conf. on Computer Science and Information Systems (FedCSIS)*, volume 8, pages 1175–1178. IEEE.

Milani, F., Lashkevich, K., Maggi, F. M., and Francescomarino, C. D. (2022). Process mining: A guide for practitioners. In *Research Challenges in Information Science - 16th Intl. Conf., RCIS, Proc.*, volume 446 of *LNBIP*, pages 265–282. Springer.

Osman, C. and Ghiran, A. (2019). When industry 4.0 meets process mining. In *Knowledge-Based and Intelligent Information & Engineering Systems: Proc. of the 23rd Intl. Conf. KES-2019*, volume 159 of *Procedia Computer Science*, pages 2130–2136. Elsevier.

Poncin, W., Serebrenik, A., and Brand, M. v. d. (2011). Process mining software repositories. In *15th European Conf. on Software Maintenance and Reengineering*, pages 5–14.

Rabelo, L. C., Habba, M., Fredj, M., and Benabdellah Chaouni, S. (2019). Alignment between business requirement, business process, and software system: A systematic literature review. *Journal of Engineering*, 2019:6918105.

Repta, D., Moisescu, M. A., Nae, A. C., Sacala, I. S., and Dumitrache, I. (2018). Towards document flow discovery in e-government systems. In *2018 Intl. Symp. on Electronics and Telecomms. (ISETC)*, pages 1–4.

Sommerville, I. (2016). *Software engineering, 10th Edition.* International computer science series. Pearson.

van der Aalst, W. M. P. (2005). Business alignment: using process mining as a tool for delta analysis and conformance testing. *Requirements Engineering*, 10(3):198–211.

van der Aalst, W. M. P. (2015). *Extracting Event Data from Databases to Unleash Process Mining*, pages 105–128. Springer.

van der Aalst, W. M. P. (2016). *Process Mining - Data Science in Action, 2nd Ed.* Springer.

Vavpotic, D., Bala, S., Mendling, J., and Hovelja, T. (2022). Software process evaluation from user perceptions and log data. *J. Softw. Evol. Process.*, 34(4).

Weske, M. (2019). *Business Process Management - Concepts, Languages, Architectures, 3rd Ed.* Springer.