A Methodology for Integrated Process and Data Mining and Analysis towards Evidence-based Process Improvement

Andrea Delgado, Daniel Calegari, Adriana Marotta, Laura González and Libertad Tansini
Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, 11300, Uruguay

Keywords: Process Mining, Data Mining, Data Science, Methodology, Organizational Improvement, Business Intelligence.

Abstract: The socio-technical system supporting an organization's daily operations is becoming more complex, with distributed infrastructures integrating heterogeneous technologies enacting business processes and connecting devices, people, and data. This situation promotes large amounts of data in heterogeneous sources, both from their business processes and organizational data. Obtaining valuable information and knowledge from this is a challenge to make evidence-based improvements. Process mining and data mining techniques are very well known and have been widely used for many decades now. However, although there are a few methodologies to guide mining efforts, there are still elements that have to be defined and carried out project by project, without much guidance. In previous works, we have presented the PRICED framework, which defines a general strategy supporting mining efforts to provide organizations with evidence-based business intelligence. In this paper, we refine such ideas by presenting a concrete methodology. It defines phases, disciplines, activities, roles, and artifacts needed to provide guidance and support to navigate from getting the execution data, through its integration and quality assessment, to mining and analyzing it to find improvement opportunities.

1 INTRODUCTION

Organizations face many challenges within their complex socio-technical systems composed of distributed infrastructures with heterogeneous technologies enacting business processes, connecting devices, people, and data. A combination of traditional information systems (IS) and Process-Aware Information System (PAIS) (Dumas et al., 2005) usually manage structured and unstructured data. Data science (IEEE, 2020; van der Aalst, 2016) emerged as an interdisciplinary discipline responding to the problem of management, analysis, and discovery of information in large volumes of data. It is fundamental within the aforementioned organizational context to provide organizations with the evidence-based business intelligence necessary to improve their daily operation.

An organizational data science project usually involves applying data mining (Sumathi and Sivanandam, 2006), and/or process mining (van der Aalst, 2016) techniques, among others. There are methodologies guiding both kind of projects, e.g., PM^2 (Eck, van et al., 2015) for process mining, and CRISP-DM (Shearer, 2000), and SEMMA (Mariscal et al., 2010) for data mining. These methodologies provide different guidance levels and consider these two kinds of a project as separate efforts. It happens because there is usually a compartmentalized vision on the process and organizational data, e.g., process data is managed within a Business Process Management Systems (BPMS) (Chang, 2016) with a focus on the control flow execution of the process. In contrast, organizational data is stored in distributed heterogeneous external databases ,not completely linked to the BPMS.

In (Delgado et al., 2020) we proposed an integrated framework for organizational data science called PRICED (for Process and Data sCience for oRganIzational improvEment). It supports business process improvement by integrating process and organizational data into a unified view, allowing the application of process and data mining techniques over the same integrated data set. It also considers data quality and process compliance assessments. This framework's main objective is to help reduce the effort to identify and apply techniques, methodologies, and tools in isolation, integrating them in one place.

In this paper, we present the methodological dimension of the PRICED framework, by means of

Delgado, A., Calegari, D., Marotta, A., González, L. and Tansini, L.
A Methodology for Integrated Process and Data Mining and Analysis towards Evidence-based Process Improvement.
DOI: 10.5220/0010607004260437
In *Proceedings of the 16th International Conference on Software Technologies (ICSOFT 2021)*, pages 426-437
ISBN: 978-989-758-523-4

a concrete methodology defining phases, disciplines, activities, roles, and artifacts needed to provide guidance and support for organizational data science projects. It considers the extraction of execution data, its integration, quality assessment, mining and analysis, and evaluation of the results to find improvement opportunities within an organization. We also provide an example of the application of the methodology as proof of concept.

The rest of the paper is structured as follows. In Section 2 we introduce the methodology with its static and dynamic views, and in Section 3 we describe an example of application. Then, in Section 4 we present related work. Finally, in Section 5 we provide conclusions and an outline of future work.

2 METHODOLOGY DEFINITION

In (Delgado et al., 2020) we devised a methodology composed of a static and a dynamic view without delving into details as we present in this section. The **static view** defines the different elements involved within the methodology, i.e., phases, disciplines, activities, roles, and artifacts. It helps to understand *what* needs to be done (artifacts), *how* it should be done (activities), and by *whom* (roles and responsibilities). The **dynamic view** describes a lifecycle guiding the efforts from getting the execution data to mining and evaluating the results to find improvement opportunities. In other words, it defines *when* the activities that must be performed.

2.1 Static View

We present the static view in what follows, defining disciplines with their activities, roles, and artifacts, without the dynamic view's temporal perspective.

2.1.1 Disciplines & Activities

Disciplines are usually used to group related activities regarding the topic they deal with, e.g., data quality assessment. We define five disciplines to tackle the different issues, with associated activities to guide the work to be carried out.

Process & Data Extraction and Integration (PDE). This discipline groups activities that deal with the identification, definition of goals, and extraction of process and organizational data from associated sources and its integration within a unified metamodel (Delgado and Calegari, 2020).

PDE1 – Select Business Processes. To identify and select business processes from the organization that will be the object of mining efforts to identify improvement opportunities. To define the mining/analysis effort goals, including the selection of execution measures when applicable.

PDE2 – Define Mining/Analysis Goals. To define the purposes of the mining/analysis efforts for the selected business processes and integrated process and organizational data, such as the need to know process variants that behave differently regarding the data they manage, the process model that better explains the process data, participants and roles involved in types of traces or managing specific types of data, among others. Also, execution measures such as duration of traces and/or activities and/or compliance requirements such as message interaction order in choreographies or tasks execution patterns between different process participants in collaborative processes can be defined/selected.

PDE3 – Identify Process and Data Sources. To identify the sources of process and organizational data that must be integrated to serve as the complete mining effort's first input. It includes evaluating and analyzing the availability of elements needed to access and obtain data from the corresponding sources (i.e., BPMS process engine, organizational databases, and associated history logs).

PDE4 – ETL Process and Organizational Data. To carry out the ETL process to extract process data from the BPMS process engine and heterogeneous organizational databases and corresponding history logs to the metamodel, we have defined (Delgado and Calegari, 2020). The metamodel includes four quadrants for process definition, process instances (i.e., cases), data definition, and data instances. As shown in Figure 1, we envision a general mechanism to extract data from heterogeneous databases at two levels: i) the process level, from different BPMS and corresponding process engines databases (i.e., Activiti BPMS with PostgreSQL, Bonita BPMS with MySQL, etc.); ii) organizational data level, from different and heterogeneous databases (relational or NoSQL, i.e., PostgreSQL, MySQL, MongoDB, Cassandra, Neo4j, etc.). It is based on extending a previous definition of a Generic API for BPMS (Delgado et al., 2016) and a new Generic API for databases (SQL/NoSQL)(Hecht and Jablonski, 2011; Khasawneh et al., 2020), allowing us to decouple the ETL process from a specific implementation of the sources. We are

currently defining this ETL process.

PDE5 – Integrate Process and OrganizationalData. To execute matching algorithms over the data loaded in the metamodel, find and define relationships between process instance variables (in the process instances quadrant) and organizational data attributes (in the process instances quadrant). Several options can be used to discover these relationships. We have implemented a basic algorithm (Delgado and Calegari, 2020) based on values and timestamps to detect which data register corresponds to which activity execution in the process, as part of the framework.

Process & Data Quality (PDQ). This discipline groups activities that deal with the selection, evaluation, and improvement (cleaning) of quality characteristics of process and organizational data from the integrated data (i.e., integrated metamodel and generated extended log). (Bose et al., 2013) identifies four main categories for quality issues in event logs: missing data, incorrect data, inaccurate data, and irrelevant data. We have defined a Business Process and Organizational Data Quality Model (BPODQM) in which specific dimensions, factors, and metrics for the integrated data from process and organizational databases are provided. It is based on previous quality models we have defined for other contexts (Cristalli et al., 2018)(Valverde et al., 2014), and on (Verhulst, 2016).

PDQ1 – Specify Data Quality Model. To instantiate the Business Process and Organizational Data Quality Model (BPODQM), select which quality characteristics will be evaluated over which data and how the evaluation is done. A quality model defines which quality dimensions and factors are considered, to which data they apply and how they are measured. The BPODQM defines dimensions, factors, and metrics specific to the context of process logs and associated organizational data, but not necessarily all these elements must be present in every particular case. Also, the selected metrics may be adapted to the particular needs and available tools for processing data. Dimensions included in the BPODQM are *Accuracy*, *Consistency*, *Completeness*, *Uniqueness*, *Freshness*, *Credibility*, and *Security*.

PDQ2 – Evaluate Quality Characteristics. To evaluate the selected quality characteristics over the integrated process and organizational data, detecting quality problems that should be resolved before the mining/analysis effort. To do this, the specified data quality model metrics are measured over the extended event log (or the integrated metamodel), and results are obtained for each one that gives insight regarding the quality of the dataset.

PDQ3 – Improve Quality Characteristics. To take the necessary corrective actions to eliminate the detected quality problems, cleaning the event log and associated organizational data. It can include removing data, i.e., unwanted outliers, duplicates, null values, correcting data according to a specific domain of possible values, etc.

Process & Data Preparation (PDP). This discipline group activities dealing with the preparation of the integrated data to be used as input for the mining/analysis effort. It includes taking data to the format that will allow mining it (i.e., extended event log) or performing the analysis (i.e., data warehouse). We have defined three extensions to the event log format for i) including corresponding organizational data in events; ii) including participants in events for collaborative processes; and iii) including data regarding message interaction participants for choreographies.

PDP1 – Build Extended Event Logs. To automatically generate the extended log from the integrated metamodel to be used as input for the mining/analysis effort. It includes gathering all integrated process and organizational data for each corresponding event when it applies, the involved participants in collaborations, and messages interactions in choreographies. We have defined two extensions for the eXtensible Event Stream (XES)[1] following the definitions of the standard: i) one to include the event type (task, message, service task, etc.) and a list of variables and entities associated with the event, and for each entity, a list of corresponding attributes with their types; ii) another that also includes the event type and two views on collaboratives BPs, wherein case of collaborations the participant that executed the event is included (as another level over role and user), and in case of choreographies the from and to a participant that sends or received the message interaction is included.

PDP2 – Build Integrated Data Warehouse. To generate the integrated data warehouse from the integrated metamodel, to be used as input for the analysis effort. We defined dimensions directly related to the metamodel quadrants, i.e., process-definition, process-instance, data-definition, and data-instance, adding a user dimension, a time dimension, and an entity relations dimension to capture entities references. It is based solely on

[1] https://xes-standard.org/start

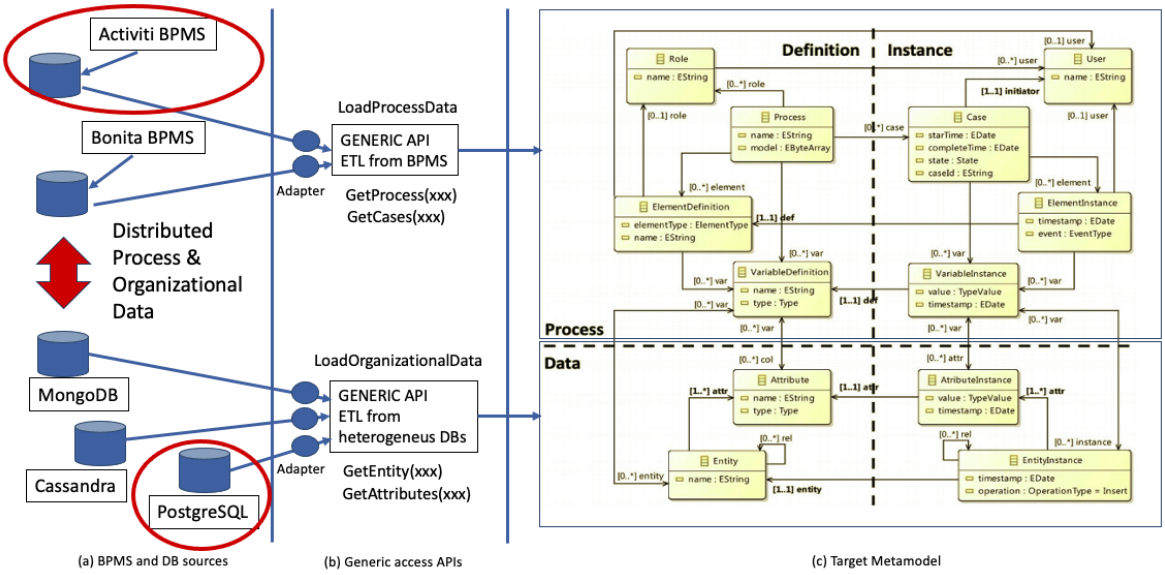


Figure 1: Mechanism for the ETL process and organizational data activity.

the relationships between process and organizational data that we previously discovered in the metamodel using matching algorithms. The fact table relates the dimensions mentioned before. We include process duration and element duration to analyze execution times for both process and elements, and we also included the value of attributes. The data warehouse allows crossing processes and organizational data to provide an integrated view of the actual execution of BPs.

PDP3 - Filter Event Log and Data. To filter different levels in the extended event log based on process or organizational data to be able to perform additional perspective mining over the data, e.g., to partition the log in process variants with similar behavior based on control flow or on the type of organizational data they manage, or by applying compliance rules, or selecting cases based on duration, among others.

Process & Data Mining and Analysis (PDMA). This discipline groups activities that deal with selecting, executing, and evaluating approaches and tools for the mining/analysis effort. We also provide an extensive catalog of existing techniques and algorithms of both process and data mining approaches and existing tools implementing them and our definitions and tools to support process and data mining of integrated data. It helps organizations using the methodology to find all the information and guidance they need in one place, to carry out the mining/analysis effort, easing its adoption.

PDMA1 - Select Mining/Analysis Approach. To select the mining and/or analysis approach to apply to the data, i.e., discovering process models (based on algorithms such as inductive miner, heuristic miner, or BPMN miner, among others), conformance and/or enhancement of process models for process mining approaches, and/or descriptive (clustering, decision trees, association rules) or predictive (classification, regression) for data mining approaches, crossing data from the business process perspective with the organizational data perspective. An example is the clustering of traces based on organizational data or clustering of data based on the control flow of traces, depending on the goals and data under analysis. Also, compliance requirements and execution measures can be selected as the desired approach to applying to the data. We provide a catalog of existing techniques and algorithms with a summary and corresponding links for each one.

PDMA2 - Select Mining/Analysis Tools. To select the mining tool to be used corresponding to the chosen approach, since different tools and/or plug-ins implement different algorithms. Also, for analysis, the tool depends on the approach selected, i.e., the data warehouse can be used to cross-process and organizational data, or the execution measures can be evaluated in a specific tool. We provide a catalog with links to existing tools and the support they provide for different approaches.

PDMA3 - Execute Mining/Analysis Approach. To carry out the selected mining/analysis approaches in the selected tools over the integrated process and organizational data, including execution measures analysis and/or compliance requirements evaluation, when defined. It includes dealing with data input issues or tool execution problems (i.e., significant execution times) that would need to return to previous activities to correct the data's problems or change the approach and/or tool selected.

PDMA4 - Evaluate Mining/Analysis Results. To evaluate the results of the mining/analysis effort from different perspectives, including the answers to goals and information needs to be defined by the business area, and more technical elements such as the correctness of results (i.e., measures such as fitness or recall, precision, overfitting, and underfitting), assessing of statistical significance, and other elements to evaluate the technical soundness of the results obtained. The business evaluation of mining/analysis results will lead to valuable information and knowledge on the organization's actual execution of business processes, identifying improvements opportunities to be carried out to generate a new version of the process.

Process & Data Compliance (PDC). This discipline groups activities that deal with selecting, specifying, and evaluating compliance requirements focusing on collaborative and choreography processes. Although it corresponds to another type of mining/analysis of processes and organizational data, we work on this topic as an independent discipline due to its specific characteristics. We have defined a Business Process Compliance Requirements Model (BPCRM) based on specified compliance requirements metamodel from previous works in another context, and on (Turetken et al., 2012) and (Knuplesch and Reichert, 2017). We have also defined a compliance requirements modeling language (González and Delgado, 2021) to specify this type of requirement.

PDC1 - Identify Compliance Requirements. To instantiate the Business Process Compliance Requirements Model (BPCRM) to select specific dimensions and corresponding factors to evaluate compliance requirements for the process selected for the mining/analysis effort. It includes collaborative and choreography processes, which are the focus of the compliance model. The BPCRM, as the BPODQM quality model, defines specific dimensions, factors, controls, and metrics to evaluate compliance requirements over BPS. Dimensions included in the BPCRM are Control flow, Interaction, Time, Resources, and Data. An example of an Interaction factor is Send/Receive Messages, and control is M occurs, e.g., message M is exchanged from sender S to receiver R, specified within the choreography or over the message itself, and measured accordingly. The compliance requirements modeling language (González and Delgado, 2021) is used for specifying process compliance requirements over the process to be evaluated.

PDC 2 - Evaluate Compliance Requirements. To evaluate the results of the compliance requirements specified over the process within the extended event log, including process and organizational data, to analyze violations in traces that do not comply with the requirements specified. We define a post mortem compliance evaluation, and we are working on a ProM plug-in to automate the analysis. Compliance requirements evaluation will lead to getting valuable information and knowledge on the actual execution of BPs, focusing on collaborations and choreographies, detecting violations to norms and business rules that should be corrected in a new version of the process.

2.1.2 Roles & Artifacts

Figure 2 shows the disciplines and activities presented, and for each activity, the roles involved and the input and output artifacts used and generated by the activity, respectively.

2.2 Dynamic View

We have refined the phases within the framework in (Delgado et al., 2020), including more specific sub-phases and adding elements. The dynamic view phases are Enactment, Data Phase, and Mining Phase, which we have updated to Mining/Analysis Phase. To carry out the improvement effort over the process, we integrate an existing Improvement phase of a previous methodology we have defined (Delgado et al., 2014).

Figure 3 presents a summary of the dynamic view of the methodology, showing for each Phase and corresponding sub-phase, the activities that are performed, and their order, i.e., previous activities.

We will present the flow of execution within the Phases, i.e., the dynamic view of the methodology in the example, on a step-by-step basis.

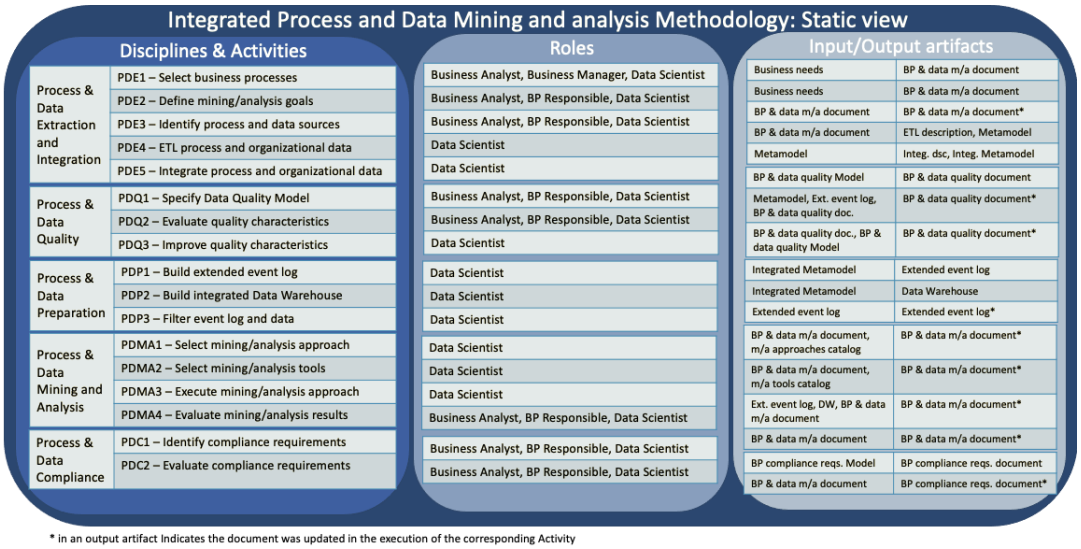


Figure 2: Summary of the Static view of the methodology.

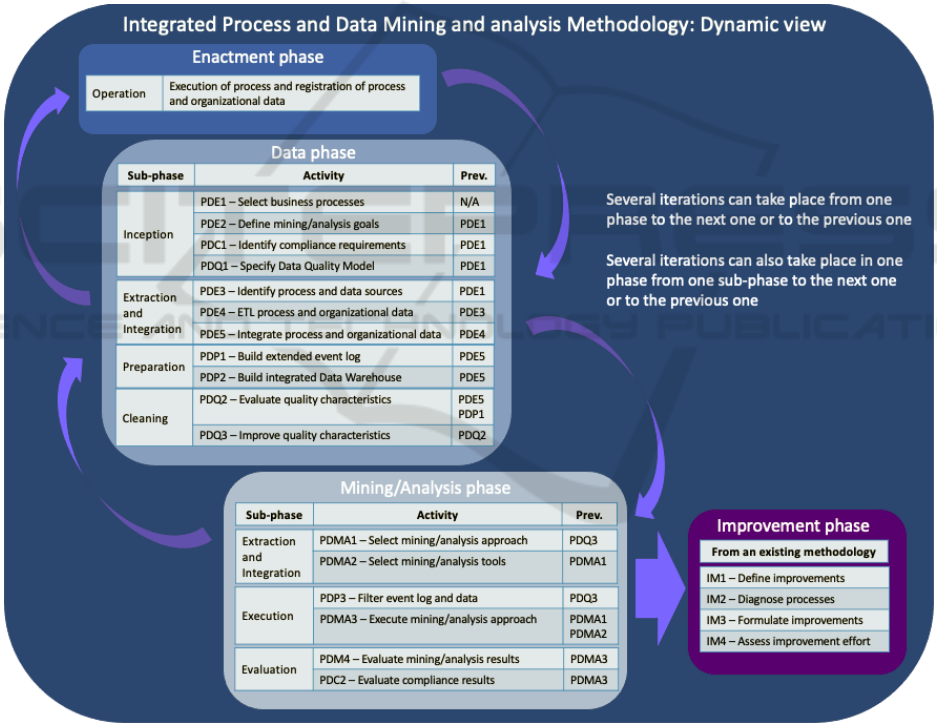
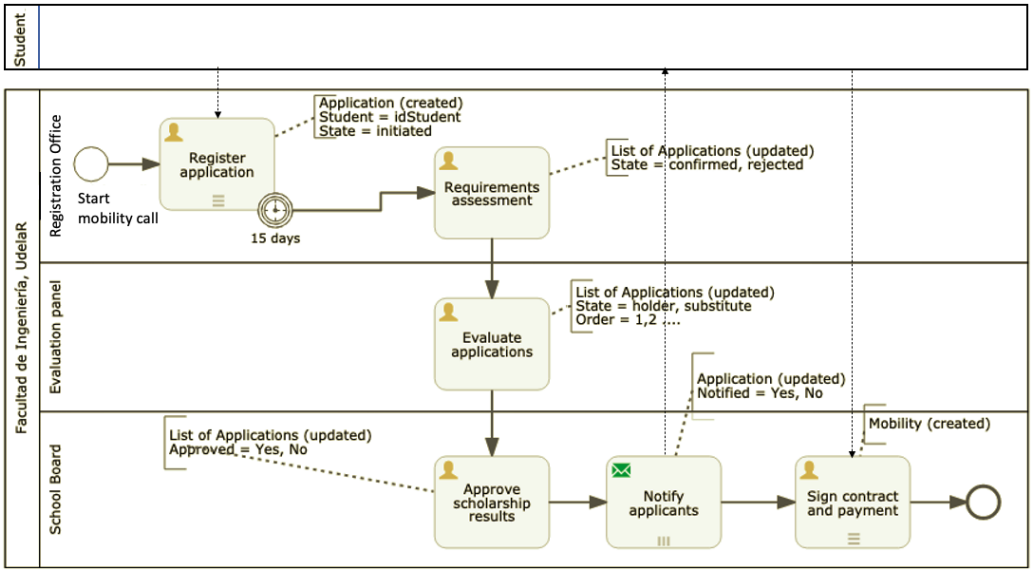


Figure 3: Summary of the Dynamic view of the methodology.

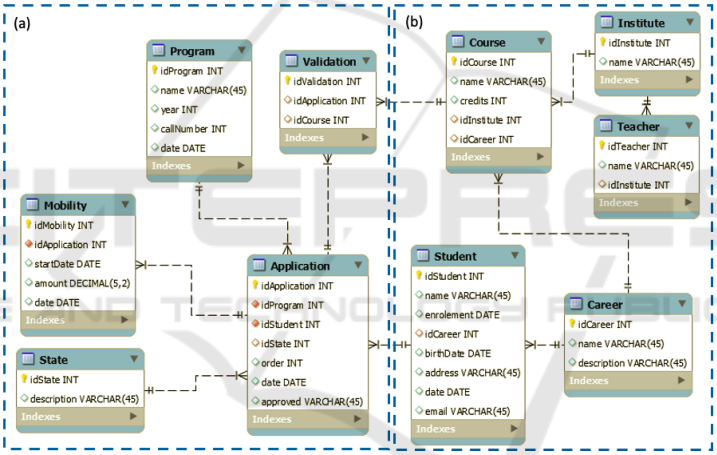
3 EXAMPLE OF APPLICATION

This section presents an example of applying the methodology that we have carried out on a real business process regarding our university. The "Students Mobility" business process has been introduced in (Delgado and Calegari, 2020) and corresponds to the application for students' scholarships to take courses at other universities. Figure 4a shows a simplified business process model using BPMN 2.0, and Figure 4b an excerpt of the organizational data model extended from (Delgado and Calegari, 2020).

The process depicted in Figure 4a begins when a new mobility call is defined and the period for receiving student's applications is opened. Students present their applications with the required documen-



(a) Students Mobility business process from (Delgado and Calegari, 2020).



(b) Extended data model for the Students Mobility business process.

Figure 4: Students Mobility proof of concept.

tation within the Registration Office. After 15 days, the period is closed, and all submitted applications go through an assessment to see if they comply with the call, and those complying go through an evaluation by an Evaluation panel, where applications are ranked, and scholarships are assigned. Finally, the School board approves the assignments, notify applicants about the results, and ask the selected ones to sign a contract for the scholarship and get paid.

The data model shown in Figure 4b presents, in the left side (a), specific tables to support the "Students mobility" process, i.e., the mobility `Program`, `Application` (with reference to the `Student`) and `Validation` (with reference to `Course`) tables, as well as the `Mobility` table to register the scholarships that were assigned. The `State` table registers the states that the application goes through the process control flow. In the right side (b), there are tables containing organization's master data, i.e., `Student` that apply to the call, their `Career` and `Course` to validate the courses selected which are associated to an `Institute` and with a `Teacher` responsible of it.

This process was implemented and executed in Activiti 6.0 BPMS[2] community edition using a PostgreSQL[3] database for the organizational data. For the analysis, we applied process and data mining techniques using Disco[4] and ProM[5], and built a data ware-

[2]https://www.activiti.org/
[3]https://www.postgresql.org/
[4]https://fluxicon.com/disco/
[5]https://www.promtools.org/

house using Pentaho Platform[6].

3.1 Execution of the Methodology

Since the methodology covers any mining/analysis effort, some activities may not apply to specific scenarios. In this case, we describe the activities we performed for each phase defined in Section 2 and justify those activities that were not considered.

3.1.1 Enactment Phase

The Enactment Phase does not have any concrete activity within the methodology. It consists of the organization's actual operation, where processes are executed, and process and organizational data are registered in their corresponding databases. In Figure 4, comments in the "Student Mobility" show when an activity access the data model to insert, query or modify data, e.g., within the "Register Application" task, the `Application` table is accessed to create a new application for a specific student (registered in table `Student`), with `State` "Initiated".

3.1.2 Data Phase

The Data Phase is essential for the mining/analysis efforts since the final outputs of this phase are the integrated process and organizational data, improved, cleaned, and with a minimum quality level to be used as a valuable input for the Mining/Analysis Phase.

Inception. In this sub-phase, we define the basis for the mining/analysis efforts.

PDE1 – Select Business Processes. We select the "Student mobility" process already introduced.

PDE2 – Define Mining/Analysis Goals. Business people (e.g., the process owner) define several business questions about the domain with a mixed perspective of data and processes, such as:

- Which organizational data were managed by cases that took the longest to execute?
- Which organizational data are involved in cases where no successful results were obtained?
- Which cases in the successful path are related to specific organizational data ?
- Which users are involved in the cases that took the longest to execute or to the ones that correspond to the successful path?
- Are there paths defined in the process model that are never executed in the actual operation?

[6]https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform.html

PDC1 – Identify Compliance Requirements. We did not perform this activity since there were no compliance requirements defined for the process.

PDQ1 – Specify Data Quality Model. We selected basic quality characteristics from the BPODQM model, to be checked over the integrated data:

- Dimension: *Accuracy*, Factor: *Syntactic accuracy*, Metric: *Format*
- Dimension: *Completeness*, Factor: *Density*, Metric: *Not null*
- Dimension: *Uniqueness*, Factor: *Duplication-free*, Metrics: *Duplicate attribute/event*

Extraction and Integration. In the Extraction and Integration sub-phase, we perform activities for extracting and loading process and organizational data into the metamodel and integrating data by finding the corresponding relationships between events (i.e., activities) and organizational data that they handled.

PDE3 – Identify Process and Data Sources. With the information of the "Students mobility" process technical infrastructure, we identify the BPMS process engine database and the organizational database and corresponding access data (i.e., machine and SID) and permits. As it is common practice in the configuration of databases, it should have been configured to allow historical logging, which we use to get all organizational data related to the process execution under evaluation in the defined period.

PDE4 – ETL Process and Organizational Data. In Figure 1, we describe the process for performing this activity. We used two databases in this proof of concept (within the ellipsis on the figure's left side): the Activiti BPMS engine database and a relational PostgreSQL database for the organizational data. We also implemented the metamodel in a PostgreSQL database for simplicity.

PDE5 – Integrate Process and Organizational Data. After the process and organizational data are loaded into the metamodel, we executed the matching algorithm to find the relations between the metamodel's process-instance and data-instance quadrants. Our basic data matching algorithm is based on discovering matches between variables (from the process-instance quadrant) and attributes instances (from the data-instance quadrant) by searching similar values within a configurable period near the start and complete events timestamps. The initial definitions for integrating process and organizational data can be seen in (Delgado and Calegari, 2020).

Preparation. In this sub-phase, we focus on putting the data in a suitable format to use as input for the mining/analysis effort.

PDP1 – Build Extended Event Logs. We automated this activity with a model-to-text transformation from the integrated metamodel to the extended event log, which includes the organizational data related to each process event. (i.e., activity).

PDP2 – Build Integrated Data Warehouse. We defined a generic data warehouse that has no domain-specific elements regarding the process or organization involved. We also automated the loading process from the integrated metamodel. The data warehouse has a star schema representing the four metamodel quadrants as dimensions and other dimensions such as users and time. We also define several measures regarding duration and values in the fact table.

Cleaning. In this sub-phase, we performed the following activities.

PDQ2 – Evaluate Quality Characteristics. We checked some of the primary factors selected, such as date format and not null for timestamps, not null, and no duplicates for event names.

PDQ3 – Improve Quality Characteristics. We found some inconsistencies in the date format for timestamps that were corrected, no nulls were found, and some duplicates on event names were corrected based on domain information.

3.1.3 Mining/Analysis Phase

The Mining/Analysis Phase is the actual core of the mining/analysis effort, where an integrated view on process and data mining is applied. Approaches and tools are selected, and the integrated data is analyzed to discover valuable information on process execution and improvement opportunities.

Inception. In this sub-phase, we select approaches and tools for the mining/analysis effort.

PDMA1 – Select Mining/Analysis Approach. As an analysis approach, we used the data warehouse to answer some of the questions included in the mining/analysis effort goals. We also use process and data mining approaches over the extended event log to provide another view of the integrated data.

PDMA2 – Select Mining/Analysis Tools. We selected the Pentaho platform to implement the data warehouse and Disco and ProM to analyze the extended log. The same data was loaded in every tool, i.e., integrated process and organizational data from the metamodel. However, as the analysis focus is different, it allows us to analyze data from different perspectives, providing a complete view on process execution.

Execution. In this sub-phase, we inspected and filtered the extended event log and data and executed the mining/analysis activities.

PDP3 – Filter Event Log and Data. We inspected the extended event log to analyze the process cases, the organizational data that was integrated with their data, and different variants of the process. Figure 5 shows Disco the frequency of selected elements in the extended event log: a) entities and b) corresponding attributes from the organizational data; and c) associated process variables. In Figure 5 a), it can be seen that organizational tables: `Application`, `Program`, and `Validation` are present in the extended event log, which were defined in the data model presented in Figure 4b.

PDMA3 – Execute Mining/Analysis Approach. Regarding process mining, we used the extended event log we generated as input to discover the process model in Disco and with the BPMN miner plug-in in ProM, to analyze the execution against the defined model. Figure 5 d) shows the model discovered in ProM, and Figure 5 e) shows the model discovered in Disco. It can be seen that the activities are not completely corresponding to the model presented in 4a. We also worked with the data warehouse, crossing data from different dimensions to answer the questions defined. For example: which courses and from which careers have been involved in cases that took more than 15 days to complete? (in the example, 15 days equals 200.000 milliseconds). We filtered data by the relation validation-course, which defines the courses included in the applications with the case id and the corresponding attributes. As rows, we included attributes from dimensions “Entityrelation”, “ProcessInstance”, “DataDefinition” and “DataInstance”. We selected the “Process duration” measure and filtered it by duration over 200.000 milliseconds. Figure 6 shows the results in our Pentaho implementation.

Evaluation. In this sub-phase, we perform the activities to evaluate mining/analysis results obtained from the execution of approaches using the selected tools.

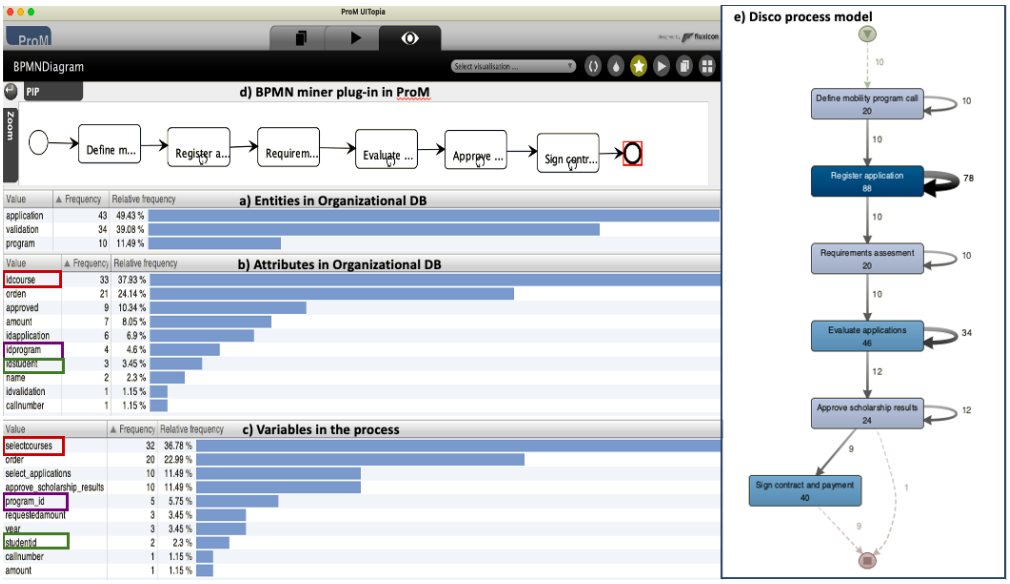


Figure 5: Extended event log analysis: a) entities; b) attributes; c) process variables; d) ProM model; and e) Disco model.

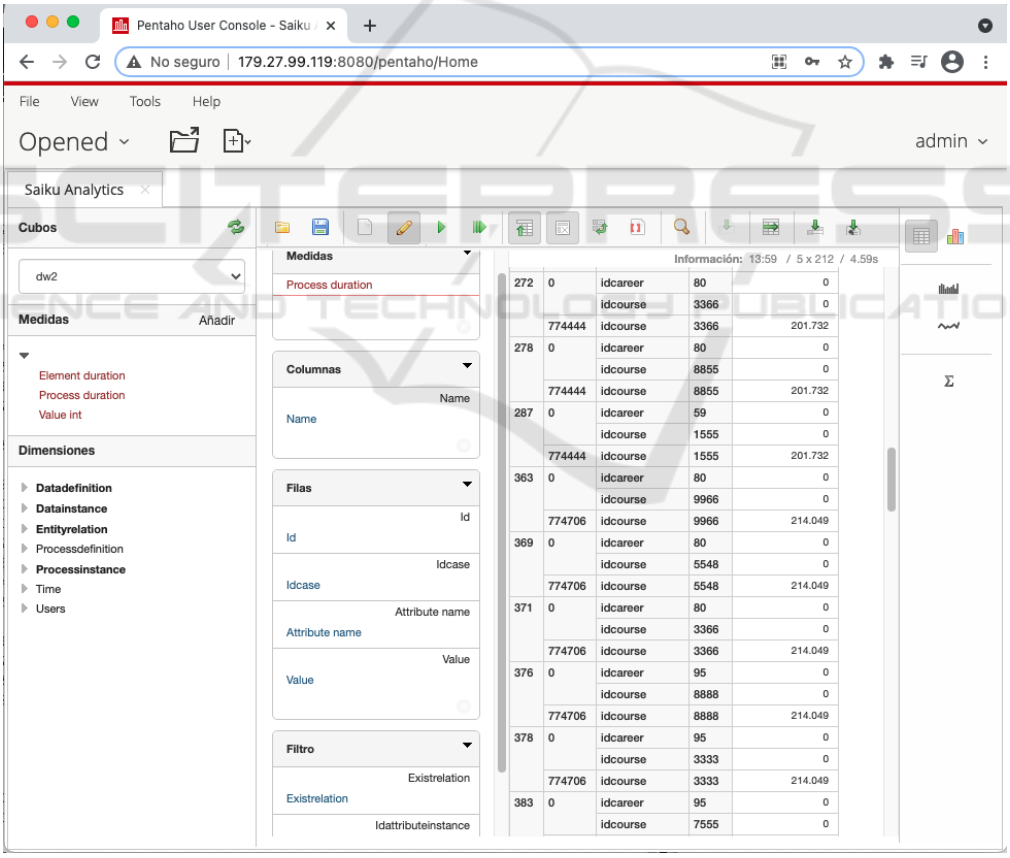


Figure 6: Data warehouse result for courses and careers involved in cases that took more than 15 days to complete.

PDMA4 – Evaluate Mining/Analysis Results. Regarding the process models discovered by ProM and Disco, although this process is elementary, several issues were detected. For instance, the activity ”Notify applicants” was not present in neither of the models, pointing to an implementation problem. Concerning the data warehouse and the example question we showed,

career with id 80 presented the most cases with process duration over the limit defined, leading to an analysis of the type of courses that students select, which can be the cause of the delays.

PDC2 - Evaluate Compliance Results. We omitted this activity since there were no compliance requirements defined for this particular process.

Improvements regarding issues discovered were not performed since new iterations over the data need to be done to obtain a deeper analysis of the results.

4 RELATED WORK

The classical data-centric analysis is most commonly guided by methodologies such as CRISP-DM (Shearer, 2000), KDD (Brachman and Anand, 1996), and SEMMA (Mariscal et al., 2010). Still, neither of them includes detailed guidelines on identifying and incorporating data that is useful to analyze organizations' processes and improve them. CRISP-DM stands for Cross-Industry Standard Process for Data Mining. This methodology was initially developed in IBM for Data Mining tasks, it is a cyclic model defining stages: Business understanding, Data understanding, Data Preparation, Modeling, Evaluation, and Deployment. KDD or Knowledge Discovery in Databases, which consists of five stages: Selection, Preprocessing, Transformation, Data Mining, and Interpretation/Evaluation. SEMMA is an acronym that stands for Sample, Explore, Modify, Model, and Access, which are its stages. A wide range of algorithms or methods are used (Gupta and Chandra, 2020).

In (Eck, van et al., 2015) the authors propose PM^2, a methodology to guide the execution of process mining projects. It consists of six stages with their corresponding activities. This methodology is consistent and complementary with ours. Planning, extraction, and data processing stages are considered within the data phase of our methodology. They also consider enriched event logs with external data, but they neither pay special attention to organizational data nor to related problems as quality assessment. Mining & analysis and evaluation stages are also considered within the Mining/Analysis phase, but in this case, they provide deeper information that ours can use. Finally, process improvement stage is considered by integrating an Improvement phase from the BPCIP methodology (Delgado et al., 2014).

Although there are many data quality proposals on data quality methodologies and frameworks, such as (Batini and Scannapieco, 2016) and (Tepandi et al., 2017), to the best of our knowledge, none of them are focused on integrated process and organizational data quality management for process mining activities. In our work, we select and adapt to our needs the main tasks of existing approaches, obtaining the three proposed tasks (definition of data quality model, evaluation, and improvement of the quality characteristics).

Various approaches propose activities for business process compliance (Hashmi et al., 2018). The COMPAS project defines a life cycle with four phases (e.g. evaluation) (Birukou et al., 2010). The C^3 Pro Project defines a design-time methodology for compliance of collaborative workflows (Knuplesch et al., 2013). The MaRCo Project defines activities for compliance management (Kharbili et al., 2011) (e.g. modeling, checking, analysis, enactment). However, these proposals neither consider these activities in the context of an integrated methodology nor leverage process and data mining for compliance control and analysis.

5 CONCLUSIONS

We have presented the PRICED methodology to carry out process and data mining and analysis efforts over integrated process data and organizational data. Key elements of our proposal include: integrated process and organizational data, i.e., from process engines and distributed organizational DBs, loaded in an integrated metamodel; quality assessment over the integrated process and organizational data; extended event logs and a data warehouse to be used for mining/analysis over the integrated data; integrated process and data mining/analysis approaches to provide a complete view of the organization's actual operation.

We are applying the methodology over more complex processes to strengthens the capabilities of the approach. We believe the methodology is a valuable tool to guide the mining/analysis efforts in organizations towards evidence-based process improvement, with a complete and integrated view on data.

ACKNOWLEDGEMENTS

Supported by project "Minería de procesos y datos para la mejora de procesos en las organizaciones" funded by Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay.

REFERENCES

Batini, C. and Scannapieco, M. (2016). *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer.

Birukou, A., D'Andrea, V., Leymann, F., Serafinski, J., Silveira, P., Strauch, S., and Tluczek, M. (2010). An integrated solution for runtime compliance governance in SOA. In *Service-Oriented Computing*, pages 122–136. Springer.

Bose, R. P. J. C., Mans, R. S., and van der Aalst, W. M. P. (2013). Wanna improve process mining results? In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 127–134.

Brachman, R. J. and Anand, T. (1996). The process of knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*, pages 37–57. AAAI/MIT Press.

Chang, J. (2016). *Business Process Management Systems: Strategy and Implementation*. CRC Press.

Cristalli, E., Serra, F., and Marotta, A. (2018). Data quality evaluation in document oriented data stores. In *Advances in Conceptual Modeling - ER 2018 Workshops*, volume 11158 of *LNCS*, pages 309–318. Springer.

Delgado, A. and Calegari, D. (2020). Towards a unified vision of business process and organizational data. In *XLVI Latin American Computing Conference, CLEI 2020*, page To appear. IEEE.

Delgado, A., Calegari, D., and Arrigoni, A. (2016). Towards a generic BPMS user portal definition for the execution of business processes. In *XLII Latin American Computer Conference - Selected Papers, CLEI 2016*, volume 329 of *ENTCS*, pages 39–59. Elsevier.

Delgado, A., Marotta, A., González, L., Tansini, L., and Calegari, D. (2020). Towards a data science framework integrating process and data mining for organizational improvement. In *15th Intl. Conf. on Software Technologies, ICSOFT 2020*, pages 492–500. ScitePress.

Delgado, A., Weber, B., Ruiz, F., de Guzmán, I. G. R., and Piattini, M. (2014). An integrated approach based on execution measures for the continuous improvement of business processes realized by services. *Information and Software Technology*, 56(2):134–162.

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.

González, L. and Delgado, A. (2021). Towards compliance requirements modeling and evaluation of e-government inter-organizational collaborative business processes. In *54th Hawaii Intl. Conf. on System Sciences, HICSS 2021*, pages 1–10. ScholarSpace.

Gupta, M. and Chandra, P. (2020). A comprehensive survey of data mining. *Int. Journal of Inf. Technology*.

Hashmi, M., Governatori, G., Lam, H.-P., and Wynn, M. T. (2018). Are we done with business process compliance: state of the art and challenges ahead. *Knowledge and Information Systems*, 57(1):79–133.

Hecht, R. and Jablonski, S. (2011). Nosql evaluation: A use case oriented survey. In *2011 Intl. Conf. on Cloud and Service Computing*, pages 336–341.

IEEE (2020). Task Force on Data Science and Advanced Analytics. http://www.dsaa.co/.

Kharbili, M. E., Ma, Q., Kelsen, P., and Pulvermueller, E. (2011). CoReL: Policy-based and model-driven regulatory compliance management. In *IEEE 15th Int. Enterprise Dist. Object Computing Conf.* IEEE.

Khasawneh, T. N., AL-Sahlee, M. H., and Safia, A. A. (2020). Sql, newsql, and nosql databases: A comparative survey. In *2020 11th Intl. Conf. on Information and Communication Systems (ICICS)*, pages 013–021.

Knuplesch, D. and Reichert, M. (2017). A visual language for modeling multiple perspectives of business process compliance rules. *Software & Systems Modeling*, 16(3):715–736.

Knuplesch, D., Reichert, M., Ly, L. T., Kumar, A., and Rinderle-Ma, S. (2013). Visual modeling of business process compliance rules with the support of multiple perspectives. In *Conceptual Modeling*, pages 106–120. Springer.

Mariscal, G., Marbán, O., and Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*, 25(2):137–166.

Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4).

Sumathi, S. and Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*, volume 29 of *Studies in Computational Intelligence*. Springer.

Tepandi, J., Lauk, M., Linros, J., Raspel, P., Piho, G., Pappel, I., and Draheim, D. (2017). The Data Quality Framework for the Estonian Public Sector and Its Evaluation. In *Trans on Large-Scale Data- and Knowl-Cent Sys XXXV*, LNCS, pages 1–26. Springer.

Turetken, O., Elgammal, A., van den Heuvel, W., and Papazoglou, M. P. (2012). Capturing compliance requirements: A pattern-based approach. *IEEE Software*, 29(3):28–36.

Valverde, M. C., Vallespir, D., Marotta, A., and Panach, J. I. (2014). Applying a data quality model to experiments in software engineering. In *Advances in Conceptual Modeling - ER 2014 Workshops*, volume 8823 of *LNCS*, pages 168–177. Springer.

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

Verhulst, R. (2016). Evaluating quality of event data within event logs:an extensible framework. Master's thesis, Eindhoven University of Technology.