A Methodology for Organizational Data Science 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,
11300 Montevideo, Uruguay
{adelgado,dcalegar,amarotta,lauragon,libertad}@fing.edu.uy

Abstract. Organizational data science projects provide organizations with evidence-based business intelligence to improve their business processes (BPs). They require methodological guidance and tool support to deal with the complexity of the socio-technical system that supports the organization's daily operations. This system is usually composed of distributed infrastructures integrating heterogeneous technologies enacting BPs and connecting devices, people, and data. Obtaining knowledge from this context is challenging since it requires a unified view capturing all the pieces of data consistently for applying both process mining and data mining techniques to get a complete understanding of the BPs execution. We have presented the PRICED framework in previous works, which defines a general strategy for performing data science projects. In this paper, we propose a methodology with phases, disciplines, activities, roles, and artifacts, providing 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.

Keywords: Process mining · Data mining · Data science · Methodology · Organizational improvement · Business intelligence

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) [17] usually managing structured and unstructured data. The complexity of this socio-technical system composed of distributed infrastructures with heterogeneous technologies enacting business processes, connecting devices, people, and data, adds many challenges for organizations. Obtaining valuable information and knowledge from this context is challenging. It requires a unified view capturing all the pieces of data consistently for applying both process mining [1] and data mining [32] techniques to get a complete understanding of the business process execution.

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.

H.-G. Fill et al. (Eds.): ICSOFT 2021, CCIS 1622, pp. 41–66, 2022.
https://doi.org/10.1007/978-3-031-11513-4_3

Organizational data science projects provide organizations with evidence-based business intelligence to improve their business processes. Data science [1,23] emerged as an interdisciplinary discipline responding to the problem of management, analysis, and discovery of information in large volumes of data. Data science projects require methodological guidance and tool support to deal with the complexity of such socio-technical systems. There are methodologies guiding both kind of projects, e.g., PM^2 [18] for process mining, and CRISP-DM [31], and SEMMA [29] for data mining. However, they consider them separate initiatives due to a compartmentalized vision of the process and organizational data. Process data is usually managed within a Business Process Management Systems (BPMS) [9]. In contrast, organizational data is stored in distributed heterogeneous databases, not wholly linked to the BPMS.

In [15] we proposed the PRICED framework (for Process and Data sCience for oRganIzational improvEment) guiding organizational data science projects to find improvement opportunities within an organization. It involves methodologies, techniques, and tools to provide organizations with key elements to analyze their processes and organizational data in an integrated manner. It considers three main aspects: integrating process and organizational data into a unified view [8] for applying process and data mining techniques over the same data set [2,12], corresponding data quality assessment [4], and evaluating compliance requirements for business processes [19]. In [14], we introduced a concrete methodology defining phases, disciplines, activities, roles, and artifacts to provide guidance and support for concrete projects. The methodology covers the extraction of systems execution data and its integration and quality assessment to evaluate the results of mining and analysis techniques to find improvement opportunities. We also provide an example of the application of the methodology as proof of concept, and in [12] we applied it in the context of E-government.

In this paper, we provide a substantially extended and thoroughly revised version of [14]. We extend the work mentioned above by providing:

1. a description of two models that are part of the conceptual dimension that supports the methodology: the Business Process and Organizational Data Quality Model (BPODQM) [4], and the Business Process Compliance Requirements Model (BPCRM) [19] (Sect. 3);
2. a detailed description on how process and data mining techniques can be applied, from the integration of process and organizational data to its combined application based on developed tools (Sect. 3);
3. an extension of the application of the methodology presented, including the integrated process and data mining analysis and evaluation view, and a new example with focus on compliance requirements evaluation (Sect. 4).

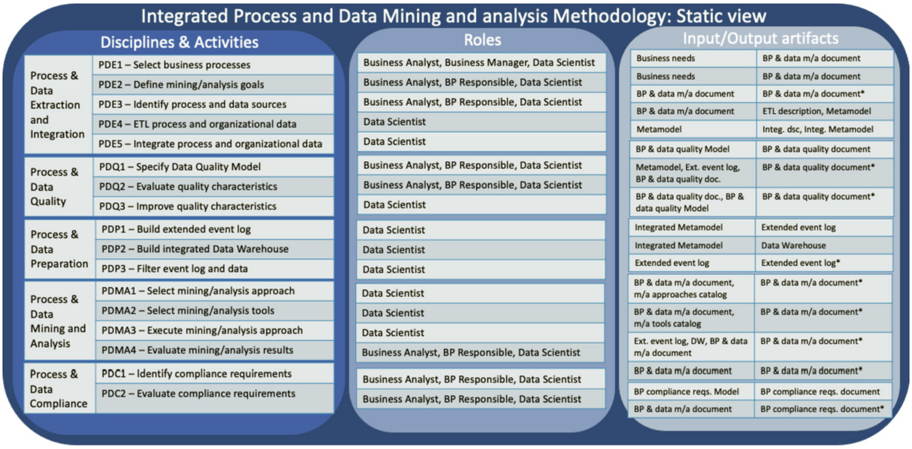
The rest of the paper is structured as follows. In Sect. 2 we introduce the methodology by presenting its static and dynamic views. In Sect. 3 we provide a deeper description of the conceptual, technical, and tool dimensions supporting the methodology. In Sect. 4 we describe examples of application. In Sect. 5 we present methodological approaches related to our proposal. Finally, in Sect. 6 we provide conclusions and an outline of future work.

2 Methodological Dimension of the PRICED Framework

In [14, 15] we introduced the methodological dimension of the PRICED framework, composed of a static and a dynamic view. 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. In what follows, we present both views, as done in [14].

2.1 Static View

Figure 1 summarizes the static view that is presented in detail next. It shows the disciplines and their activities, and, for each activity, the roles involved and the input and output artifacts used and generated by the activity, respectively.



Integrated Process and Data Mining and analysis Methodology: Static view

Disciplines & Activities		Roles	Input/Output artifacts	
Process & Data Extraction and Integration	PDE1 – Select business processes	Business Analyst, Business Manager, Data Scientist	Business needs	BP & data m/a document
	PDE2 – Define mining/analysis goals	Business Analyst, BP Responsible, Data Scientist	Business needs	BP & data m/a document
	PDE3 – Identify process and data sources	Business Analyst, BP Responsible, Data Scientist	BP & data m/a document	BP & data m/a document*
	PDE4 – ETL process and organizational data	Data Scientist	BP & data m/a document	ETL description, Metamodel
	PDE5 – Integrate process and organizational data	Data Scientist	Metamodel	Integ. dsc, Integ. Metamodel
Process & Data Quality	PDQ1 – Specify Data Quality Model	Business Analyst, BP Responsible, Data Scientist	BP & data quality Model	BP & data quality document
	PDQ2 – Evaluate quality characteristics	Business Analyst, BP Responsible, Data Scientist	Metamodel, Ext. event log, BP & data quality doc.	BP & data quality document*
	PDQ3 – Improve quality characteristics	Data Scientist	BP & data quality doc., BP & data quality Model	BP & data quality document*
Process & Data Preparation	PDP1 – Build extended event log	Data Scientist	Integrated Metamodel	Extended event log
	PDP2 – Build integrated Data Warehouse	Data Scientist	Integrated Metamodel	Data Warehouse
	PDP3 – Filter event log and data	Data Scientist	Extended event log	Extended event log*
Process & Data Mining and Analysis	PDMA1 – Select mining/analysis approach	Data Scientist	BP & data m/a document, m/a approaches catalog	BP & data m/a document*
	PDMA2 – Select mining/analysis tools	Data Scientist	BP & data m/a document, m/a tools catalog	BP & data m/a document*
	PDMA3 – Execute mining/analysis approach	Data Scientist	Ext. event log, DW, BP & data m/a document	BP & data m/a document*
	PDMA4 – Evaluate mining/analysis results	Business Analyst, BP Responsible, Data Scientist	BP & data m/a document	BP & data m/a document*
Process & Data Compliance	PDC1 – Identify compliance requirements	Business Analyst, BP Responsible, Data Scientist	BP compliance reqs. Model	BP compliance reqs. document
	PDC2 – Evaluate compliance requirements	Business Analyst, BP Responsible, Data Scientist	BP & data m/a document	BP compliance reqs. document*

* in an output artifact Indicates the document was updated in the execution of the corresponding Activity

Fig. 1. Summary of the static view of the methodology (from [14]).

Disciplines and Activities. Disciplines are usually used for grouping 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 and 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 [11].

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 mining effort's 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 with their 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 [11]. The metamodel includes four quadrants: process definition, process instances (i.e., cases), data definition, and data instances.

PDE5 - Integrate Process and Organizational Data. 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 implemented a basic algorithm [11] based on values and timestamps.

Process and Data Quality (PDQ). This discipline groups activities that deal with the selection, evaluation, and improvement (cleaning) of quality characteristics of the integrated data (i.e., integrated metamodel and generated extended log). In [6] the authors identify 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) [4] in which specific dimensions, factors, and metrics for the integrated data from process and organizational databases are provided (c.f. Sect. 3). It is based on previous quality models we have defined for other contexts [10, 34], and on [35].

PDQ1 - Specify Data Quality Model. To instantiate the 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, which data they apply and how they are measured. The dimensions, factors, and metrics defined in BPODQM are 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.

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). 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 and 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 (i.e., extended event log) or performing the analysis (i.e., data warehouse). We have defined two extensions to the event log format for i) including corresponding organizational data in events; ii) including participants in events and messages exchanged for collaborative processes and including data regarding message interaction participants for choreographies.

PDP1 - Build Extended Event Logs. To automatically generate the extended log from the integrated metamodel 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 exchanged, and messages interactions in choreographies. We have defined two extensions for the eXtensible Event Stream (XES) [24] following the definitions of the standard (c.f. Sect. 3).

PDP2 - Build Integrated Data Warehouse. To generate the integrated data warehouse from the integrated metamodel, 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 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 BPs execution.

PDP3 - Filter Event Log and Data. To filter the extended event log 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 and Data Mining and Analysis (PDMA). This discipline groups activities that select, execute, and evaluate approaches and tools for the mining/analysis effort. We also provide a catalog of existing techniques and algorithms of process and data mining approaches and existing tools implementing them, and new definitions and tools to support integrated analysis. It helps organizations use 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 (c.f. Sect. 3). 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 of tools and the support they provide.

PDMA3 - Execute Mining/Analysis Approach. To carry out the selected mining/analysis approaches in the selected tools over the integrated data, including execution measures analysis and compliance requirements evaluation. It includes dealing with data input issues and 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 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 and Data Compliance (PDC). This discipline groups activities that deal with the identification and evaluation, business process compliance requirements. We have defined a Business Process Compliance Requirements Model (BPCRM) [20] in which specific dimensions, factors and controls for collaborative BPs are defined (c.f. Sect. 3). It is mainly based on the compliance perspectives proposed in [27] as well as on the pattern vision presented in [30].

PDC1 - Identify Compliance Requirements. To instantiate the BPCRM to select specific dimensions, factors, and corresponding controls 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, and controls to evaluate compliance requirements over collaborative BPS. The compliance requirements modeling language [19] is used for specifying process compliance requirements over the process to be evaluated.

PDC2 - 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 over the extended event logs from BPs execution. Compliance requirements evaluation will get 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.

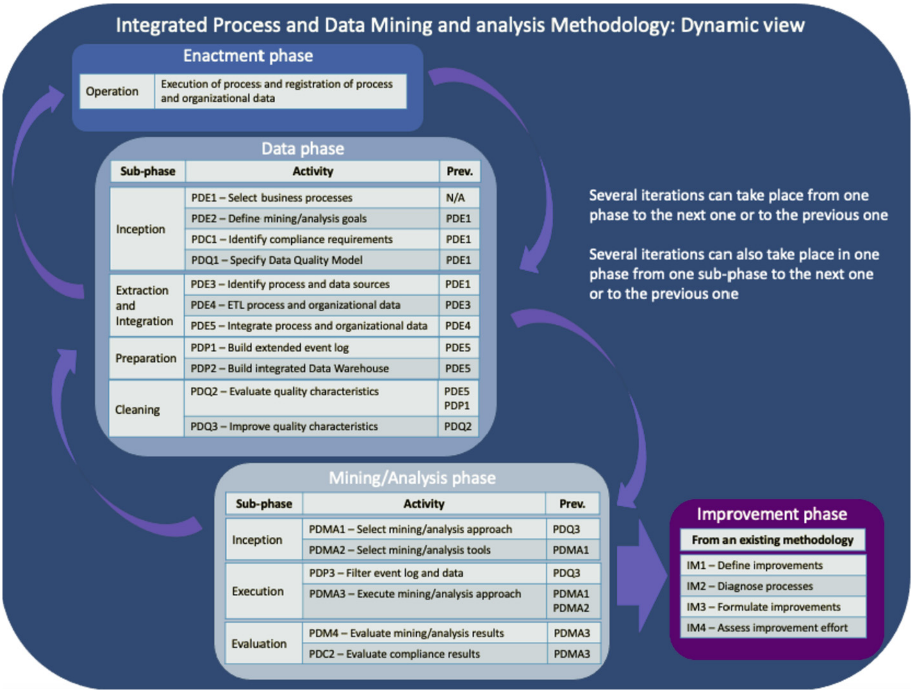
Roles and Artifacts. There are four roles within the methodology. The *Business Manager* supervises and leads a company's operations and employees. Since it is interested in improving business processes, it selects the business processes that will be analyzed. From there, the *BP Responsible* (also known as Process Owner) is in charge since it is responsible for managing such process from end-to-end. In this context, it participates in providing domain information and requirements, e.g., providing access to data sources, defining analysis goals, and also on the evaluation activities of the methodology. The *Business Analyst* also participates in the same activities as the BP Responsible, bridging the gaps between IT and the business. Finally, the *Data Scientist* represents the more technical role responsible for making value out of data, from getting and integrating the source information to analyzing it.

Concerning the artifacts, the primary artifacts of the methodology are the integrated metamodel that integrates process and organizational data, the extended event log and the data warehouse used for the analysis, and the data quality and compliance requirements models that are refined for each specific process. Also, there are other documents describing business needs, business process and data mining and analysis, and tools catalog, among others.

2.2 Dynamic View

Figure 2 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. The dynamic view is composed of three iterative phases: *Enactment*, *Data*, and *Mining/Analysis*. The Enactment phase corresponds to the actual execution of processes from which data is registered. The Data phase involves the inception, extraction, integration, preparation, and cleaning of data. Finally, the Mining/Analysis phase considers the selection and execution of the mining/analysis approaches and the evaluation of their results.

We also integrated an existing Improvement phase from the Business Process Continuous Improvement Process (BPCIP, [16]) methodology to carry out the improvement


Integrated Process and Data Mining and analysis Methodology: Dynamic view
Enactment phase
Data phase
Mining/Analysis phase
Improvement phase
Several iterations can take place from one phase to the next one or to the previous one
Several iterations can also take place in one phase from one sub-phase to the next one or to the previous one

Operation	Execution of process and registration of process and organizational data

Sub-phase	Activity	Prev.
Inception	PDE1 – Select business processes	N/A
	PDE2 – Define mining/analysis goals	PDE1
	PDC1 – Identify compliance requirements	PDE1
	PDQ1 – Specify Data Quality Model	PDE1
Extraction and Integration	PDE3 – Identify process and data sources	PDE1
	PDE4 – ETL process and organizational data	PDE3
	PDE5 – Integrate process and organizational data	PDE4
Preparation	PDP1 – Build extended event log	PDE5
	PDP2 – Build integrated Data Warehouse	PDE5
Cleaning	PDQ2 – Evaluate quality characteristics	PDE5 PDP1
	PDQ3 – Improve quality characteristics	PDQ2

Sub-phase	Activity	Prev.
Inception	PDMA1 – Select mining/analysis approach	PDQ3
	PDMA2 – Select mining/analysis tools	PDMA1
Execution	PDP3 – Filter event log and data	PDQ3
	PDMA3 – Execute mining/analysis approach	PDMA1 PDMA2
Evaluation	PDM4 – Evaluate mining/analysis results	PDMA3
	PDC2 – Evaluate compliance results	PDMA3

From an existing methodology
IM1 – Define improvements
IM2 – Diagnose processes
IM3 – Formulate improvements
IM4 – Assess improvement effort

Fig. 2. Summary of the dynamic view of the methodology (from [14]).

effort over the selected processes. This phase consists of defining the specific improvements that are going to be integrated into the improvement phase of the BP lifecycle, a diagnosis of the maturity of the BP process involved to assess the appropriateness of such improvement, a refinement of the improvements that need to be done, and the final assessment of such improvement effort.

3 PRICED Dimensions Supporting the Methodology

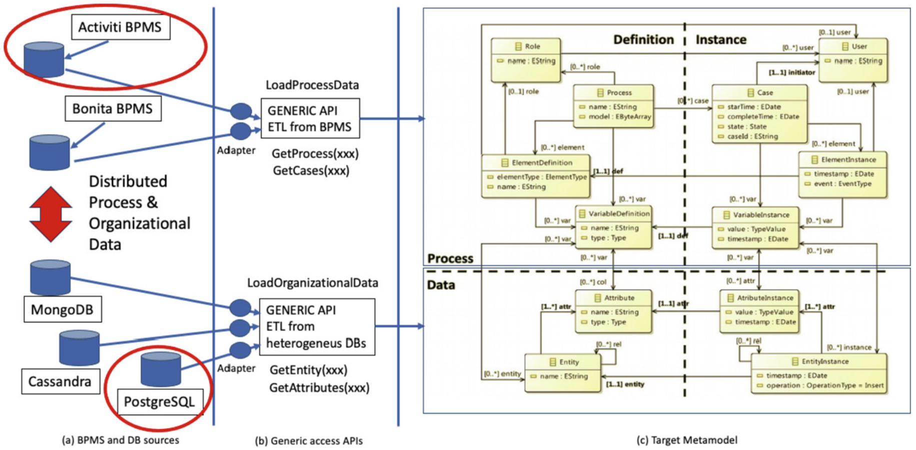
The conceptual dimension of the PRICED framework defines concepts for process and data mining, data quality, and process compliance that support the methodological dimension presented in the last section. Also, the methodology requires the definition of technical and tool dimensions, techniques, algorithms, and tools for its concrete application.

In what follows, we firstly present the general approach for process and organizational data integration, including the extensions for event logs we have defined to deal with integrated process and organizational data and collaborative BPs. Then, we present two main concepts of the conceptual dimension: the Business Process and Organizational Data Quality Model (BPODQM) [4], and the Business Process Compliance Requirements Model (BPCRM) [19], which allow us to select quality characteristics and compliance requirements to be evaluated over the extended event logs. Finally, we describe the approach for integrated process and data mining techniques over the integrated data.

3.1 Process and Data Integration Approach

During the Data phase of the methodology, we extract process and organizational data and integrate it into a unified view. Data is structured based on a generic metamodel called Business Process and Organizational Data Integrated Metamodel (BPODIM), and an algorithm matches process and organizational data exploiting their data values, and timestamps [8].

As shown in Fig. 3, 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.). We are currently defining this ETL process. It is based on extending a previous definition of a Generic API for BPMS [13] and a new Generic API for databases (SQL/NoSQL) [22,26], allowing us to decouple the ETL process from a specific implementation of the sources.



(a) BPMS and DB sources (b) Generic access APIs (c) Target Metamodel

Fig. 3. ETL for process and organizational data (from [14]).

Once the data is integrated within a database whose schema is based on the BPODIM metamodel, it is prepared to be used within the mining/analysis phase. For this, we build a generic data warehouse [2] and extended event logs based on the eXtensible Event Stream (XES) standard [24]. An XES log represents events grouped in traces (cases) for a given process. They are used as input for applying integrated process and data mining techniques, as is described in Sect. 3.4. XES provides an extension mechanism for defining new attributes to events, e.g., organizational, representing roles, and time, representing timestamps. We have defined two extensions to deal with organizational data and collaborative BPs, not just process orchestrations as usual.

The Organizational Data extension [4] defines string attributes representing organizational data associated with each event. For each event, we describe the list of variables and entities, which contains a list of the attributes related to the event. Variables correspond to process variables handled by an event, i.e., an activity within the BPMS execution (top-right quadrant of the BPODIM metamodel). Entities, and their corresponding attributes, correspond to the organizational data registered in the organizational database. They are linked to the variables through the matching algorithm (bottom-left and right quadrants). For each element in the list, we register its value and its type. In the case of attributes that matched a specific variable, we register a reference to such variable. The Collaborative BPs extension [20] define string attributes to identify the participants associated with the events, in two scenarios: the owner of the event within a collaboration between two or more participants and the sender/receiver for message elements, and within a choreography which is focused on the interchange of messages, only the sender/receiver for message elements. We also represent the type of element in both extensions, e.g., user task, service task, send or receive message task, etc.

We automatize all the processes from the data extraction to the generation of the extended event logs and data warehouse, following a model-driven approach. In particular, we have defined a chain of model transformations that takes the information within the database registering the metamodel information and generating a model conforming to the BPODIM metamodel, and then an Acceleo model-to-text transformation for generating the XES file.

3.2 Business Process and Organizational Data Quality Model

As said before, we defined the BPODQM data quality model to manage data quality issues in log data, first evaluating and then cleaning. It is based on previous quality models we have defined for other contexts [10,34], and on [35]. This model comprises all the quality aspects that should be considered, how these aspects should be measured, and the elements of the log data corresponding to process events and the organizational databases, over which the quality aspects apply. These quality aspects are organized in quality dimensions, which in turn are composed of quality factors. One or more metrics are defined for each quality factor, which specifies how the factor is measured. Each metric is defined for a certain data granularity, which is the data unit whose quality will be measured and to which the quality measures will be associated.

Considering the log data, whose quality should be measured, and its format, specific granularities are defined as follows: **attribute value**, which is the particular value of an attribute, **attribute**, which refers to the set of values corresponding to the same key, **event**, which involves all data included in an event data, and **log**, which is used for metrics that refer to the whole log.

The data quality dimensions and factors included in BPODQM are presented in the following. A more detailed description of the metrics can be found in [4].:

- *Accuracy* dimension, which is related to the correctness of the data with respect to a referential value. The quality factors that compose this dimension are *syntactic accuracy*, *semantic accuracy* and *precision*.
- *Consistency* dimension, which addresses the problem of consistency between data. The quality factors corresponding to this dimension are *domain consistency*, *inter-element consistency* and *intra-element consistency*, the first one representing consistency of a data value concerning a particular domain, and the second and third ones representing consistency between two data values of the same data element, and two data values of different elements, respectively.
- *Completeness* dimension, which refers to the absence of data that should be present. Two factors are defined for this dimension: *coverage* and *density*. The first one explores what portion of the real-world entities are represented in the data. The second one focuses on how many data values that should be present are not, for example, appearing as NULL values.
- *Uniqueness* dimension, which addresses the problem of duplicate data. The quality factors considered in this dimension are *duplication free* and *contradiction free*, each one evaluating if the data is not duplicated and, in the case, it is duplicated, if it has no contradictions, respectively.
- *Freshness* dimension, which is related to the consistency of the log data timestamps.
- *Credibility* dimension, which is composed of two factors: *provenance* and *trustworthiness*. The first one refers to the credibility of the responsible of the log data and the event origin, and the reproducibility of a log, and the second one is related to the believability of data.
- *Security* dimension, which is composed by three factors: *user permissions*, *encrypted data*, and *anonymity*, each one addressing the problems of user rights, data encryption and data anonymization, respectively.

We have developed a ProM plug-in that uses the extended event log with integrated process and organizational data as input to support the automated evaluation of event log data quality with the BPODQM (Sect. 4).

3.3 Business Process Compliance Requirements Model

The Bussiness Process Compliance Requirements Model (BPCRM) aims to provide a library of built-in compliance elements in order to facilitate the specification and validation of compliance requirements over collaborative BPs The model comprises a set of more than seventy predefined compliance controls, which are organized in five dimensions and twenty-one factors. These elements are mainly based on the compliance persepectives proposed in [27] as well as on the pattern vision presented in [30].

The set of generic compliance controls apply to both the collaboration and choreography views of collaborative BPs. In addition, they can be instantiated over a concrete process in order to specify particular compliance requirements, and used as input to evaluate violations with process mining. Therefore, the proposed model constitutes a catalogue of compliance controls (patterns), which can be used for two purposes: the specification of compliance requirements and the validation of compliance rules.

Next, the compliance dimensions and factors that conform the BPCRM and examples of compliance factors for each dimension are presented. For a complete description of the model and its components refer to [20].

- *Control Flow* dimension deals with compliance aspects related to the occurrence and order of tasks as well as their flow [28]. This dimension has eleven controls which are organized into five factors: Tasks, Sequence Flow, Parallel Flow, Exclusive Flow and Alternative Flow. For example, one of the compliance controls within this dimension enables the specification of requirements such as *"if activity A is not present, then activity B must not be present"*.
- *Interaction* dimension deals with compliance aspects related to message exchanges between participants as well as their flow [28]. This dimension has eleven controls which are organized into two factors: Send/Receive Messages and Message Flow. For example, one of the compliance controls within this dimension enables the specification of requirements such as *"if message M is exchanged, then message N must not be exchanged, and vice versa"*.
- *Time* dimension deals with compliance aspects related to points in time as well as time intervals and conditions [28]. This dimension has twelve controls which are organized into three factors: Point in Time, Interval and Duration. For example, one of the compliance controls within this dimension enables the specification of requirements such as *"if activity A occurs then activity B must occur within interval I"*.
- *Resources* dimension deals with compliance aspects related to the resources used in processes as well as their relations [28]. This dimension comprises controls which are organized into seven factors: Roles, Staff Members, Groups, Organizational Units, Participants, Resource Relations, and Performer Relations. For example, one of the compliance controls within this dimension enables the specification of requirements such as *"if activity A is performed by user U and activity B is performed by user V, then U and V are assigned to organizational unit O"*.
- *Data* dimension deals with compliance aspects related to data elements used in processes as well as their relations and flows [28]. This dimension has twenty controls which are organized into four factors: Data Objects, Data Containers, Data Relations and Data Flow. For example, one of the compliance controls within this dimension enables the specification of requirements such as *"data object DO written by activity A must be contained in message M"*.

We have developed a ProM plug-in that uses the extended event log for collaborative BPs as input, to support the automated evaluation of compliance requirements over the event log data with the BPCRM (Sect. 4).

3.4 Integrated Process and Data Mining Approach

The integrated process and data mining approach we have defined operates over the Organizational Data extension for the event logs. Organizational data is included in the corresponding event as described above. We apply data mining techniques over organizational data from the events to view the process traces that manipulated such data.

We use process mining techniques over process data to discover traces with different behavior and relate it to the data they manage.

For example, in the Loan request process from a bank, clients can submit their request, including identification data and the requested amount. The process registers these data in an external organizational database where loan requests are maintained, apart from the process data. Traditionally data is analyzed without linking it to the process, and the process is analyzed without connecting it to the data it managed. For example, with data mining, patterns regarding the loan request data can be discovered, relating different attributes, but not with the process execution that managed the data.

With our integrated approach, apart from grouping traces regarding control flow behavior (i.e., process variants), we can group them by values of the organizational data. For example, regarding the result of the loan request: was it approved or rejected? Who managed the approval? or the ranks of the amount requested. Then we can analyze each group of traces to find common elements that could have led to one or the other outcome using the control flow behavior, i.e., discovering the process for each group. Without including organizational data in the event log, this type of analysis is not possible. Also, we can analyze each process variant based on the behavior it groups, i.e., which activities are executed and in what order, and analyze the organizational data related to this specific type of path over the process to discover common data elements that are related with the variant.

We have developed a ProM plug-in that uses the extended event log with integrated process and organizational data as input and implements the integrated process and data mining approach. It provides the most common data mining techniques for analysis: decision trees, clustering, and association rules, as well as the process mining techniques that are already provided in the framework (Sect. 4).

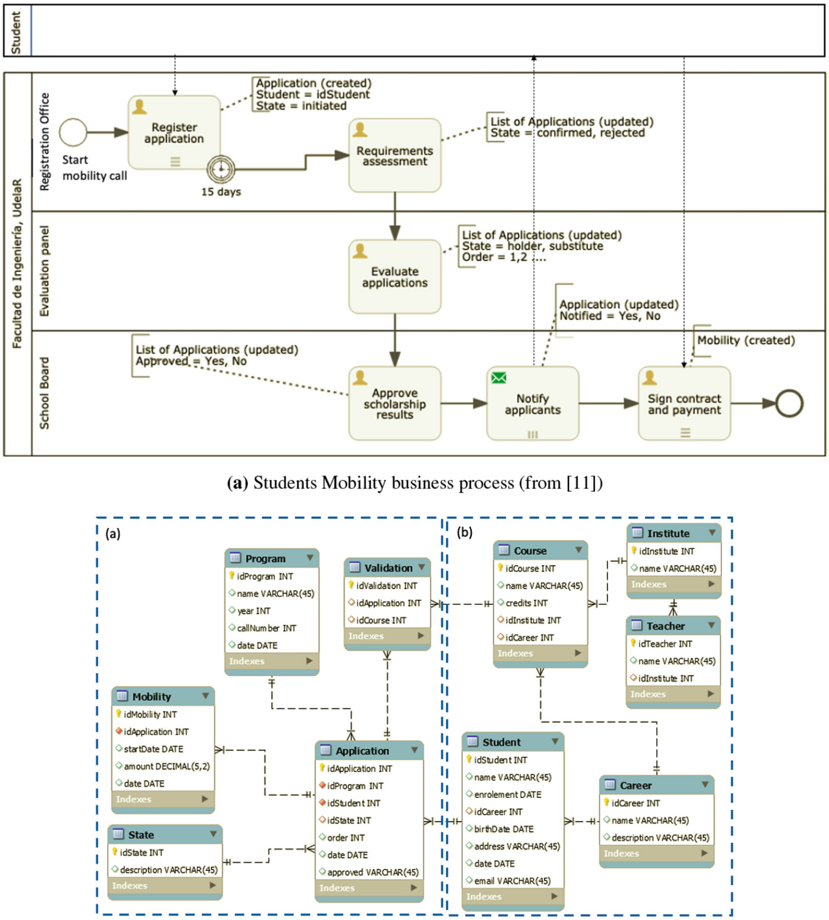
4 Applications of the Methodology

This section presents two examples of applying the methodology on actual BPs regarding our university and e-Government processes from the Uruguayan digital services. The "Students Mobility" BP, has been introduced in [11] and corresponds to the application for students' scholarships to take courses at other universities. The "Passport request" BP has been introduced in [19] and corresponds to the collaborative BP for requesting a passport by a citizen. In the first case, we present a step-by-step application of the methodology showing the integrated process and organizational data approach, data matching, quality evaluation, process mining tools, and data warehouse for analysis, but with no compliance requirements evaluation. In the second case, we focus on the compliance evaluation approach, showing the use of the compliance requirements specification, execution, and evaluation.

4.1 Students Mobility BP with Organizational Data Extension

The simplified BPMN 2.0 process depicted in Fig. 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 documentation 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. Those complying go through an evaluation panel evaluation, where applications are ranked and scholarships are assigned. Finally, the School board approves the assignments, notifies applicants about the results, and asks the selected ones to sign a contract for the scholarship and get paid.



(a) Students Mobility business process (from [11])

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

Fig. 4. Students mobility proof of concept (from [14]).

The data model shown in Fig. 4b presents an excerpt of the organizational data model extended from [11]. In the left side (a), there are 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[1] community edition using a PostgreSQL[2] database for the organizational data. We applied process and data mining techniques using Disco[3] and ProM[4], and built a data warehouse using Pentaho Platform[5].

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 Sect. 2.

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 Fig. 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 with `State` "Initiated".

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?

[1] https://www.activiti.org/.

[2] https://www.postgresql.org/.

[3] https://fluxicon.com/disco/.

[4] https://www.promtools.org/.

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

- Which users are involved in the cases that took the longest to execute or the ones that correspond to the successful path?
- Are there paths defined in the process model that are never executed in the actual operation?

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 Fig. 3, 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.

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 data can be seen in [11].

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

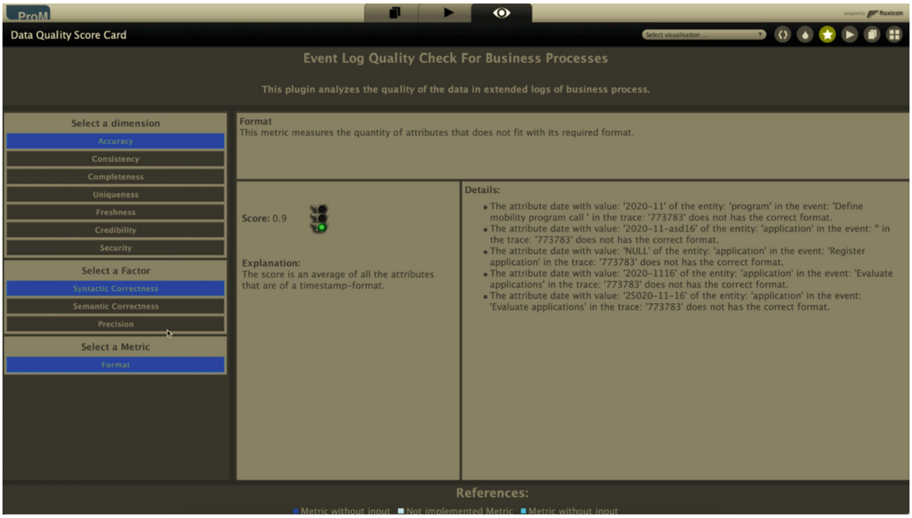
PDP1 - Build Extended Event Logs. We automated this activity with a model-to-text transformation from the integrated metamodel to the extended event log, including the organizational data related to each process event.

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 others 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, not null for timestamps, not null, and no duplicates for event names. To do so, we used the ProM plug-in we have developed that automatically analyzes the extended event log with integrated data evaluating quality issues as defined in the BPOQM model. In Fig. 5 we present an example of the results of the analysis for Dimension Accuracy, Factor Syntactic accuracy, and Metric Format applied to date.

PDQ3 - Improve Quality Characteristics. As it can be seen in Fig. 5 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.



Event Log Quality Check For Business Processes

This plugin analyzes the quality of the data in extended logs of business process.

Format
This metric measures the quantity of attributes that does not fit with its required format.

Score: 0.9

Explanation:
The score is an average of all the attributes that are of a timestamp-format.

Details:

- The attribute date with value: '2020-11' of the entity: 'program' in the event: 'Define mobility program call ' in the trace: '773783' does not has the correct format.
- The attribute date with value: '2020-11-asd16' of the entity: 'application' in the event: '' in the trace: '773783' does not has the correct format.
- The attribute date with value: 'NULL' of the entity: 'application' in the event: 'Register application' in the trace: '773783' does not has the correct format.
- The attribute date with value: '2020-1116' of the entity: 'application' in the event: 'Evaluate applications' in the trace: '773783' does not has the correct format.
- The attribute date with value: '2S020-11-16' of the entity: 'application' in the event: 'Evaluate applications' in the trace: '773783' does not has the correct format.

Fig. 5. ProM quality plug-in for extended event logs with integrated data.

Mining/Analysis Phase. The Mining/Analysis Phase is the core of the mining/analysis effort, where an integrated view of 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. In addition, we also used our approach for integrated process and data mining over process and organizational integrated data.

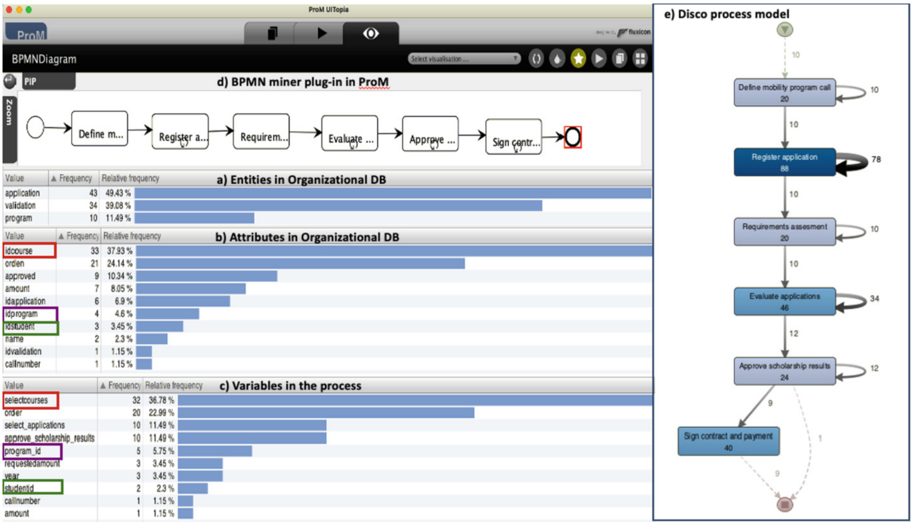
PDMA2 - Select Mining/Analysis Tools. We selected the Pentaho platform to implement the data warehouse and the mining tools Disco and ProM to analyze the extended log, including our ProM plugin for integrated process and data mining for the extended log. The same data was loaded in every tool, i.e., 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 process variants. Figure 6 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 Fig. 6 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 Fig. 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 6 d) shows the model discovered in ProM, and Fig. 6 e) shows the model discovered in Disco. Activities do not completely correspond to the model presented in 4a. We also worked with the data warehouse, crossing data from different dimensions to answer the questions defined, e.g., 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 7 shows the results in Pentaho.

Regarding the integrated process and data mining approach that is implemented in our ProM plug-in, we analyzed the extended event log based on organizational data to know the cases that were associated with these data, for example, cases that have scholarships approved and rejected, cases that manage different ranks of amounts for scholarships, teachers that were involved in evaluating the scholarships, etc. We can then analyze the resulting cases to see whether there is a different or specific behavior associated with the organizational data. In Fig. 8 we present an example of the results for clustering cases based on approved and rejected scholarships. It can be seen that when selecting one case in the cluster on the left panel, on the main panel, the path of the case over the process model is highlighted.



a) Entities in Organizational DB

Value	▲ Frequency	Relative frequency
application	43	49.43 %
validation	34	39.08 %
program	10	11.49 %

b) Attributes in Organizational DB

Value	▲ Frequency	Relative frequency
idcourse	33	37.93 %
orden	21	24.14 %
approved	9	10.34 %
amount	7	8.05 %
idapplication	6	6.9 %
idprogram	4	4.6 %
idstudent	3	3.45 %
name	2	2.3 %
idvalidation	1	1.15 %
callnumber	1	1.15 %

c) Variables in the process

Value	▲ Frequency	Relative frequency
selectcourses	32	36.78 %
order	20	22.99 %
select_applications	10	11.49 %
approve_scholarship_results	10	11.49 %
program_id	5	5.75 %
requestedamount	3	3.45 %
year	3	3.45 %
studentid	2	2.3 %
callnumber	1	1.15 %
amount	1	1.15 %

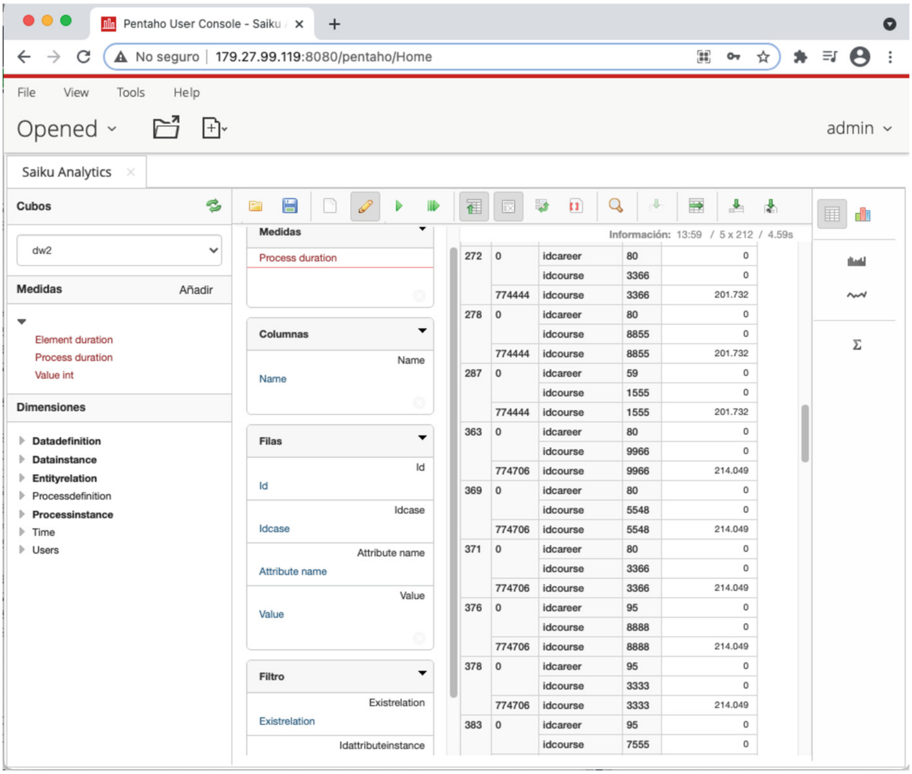
Fig. 6. Extended event log analysis: a) entities; b) attributes; c) process variables; d) ProM model; and e) Disco model.

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

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 absent in both models, pointing to an implementation problem. Concerning the data warehouse and the example question, a career with id 80 presented the most cases with process duration over the defined limit, leading to an analysis of the type of courses that students select, which can cause the delays. The integrated analysis over the extended log also gave us insight into the execution of the process and the relation with organizational data, particularly for the approved and rejected results for scholarships.

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.


Pentaho User Console - Saiku
No seguro | 179.27.99.119:8080/pentaho/Home
File View Tools Help
Opened
admin
Saiku Analytics
Cubos
dw2
Añadir

Medidas

Element duration
Process duration
Value int

Dimensiones

- Datadefinition
- Datainstance
- Entityrelation
- Processdefinition
- Processinstance
- Time
- Users

Medidas: Process duration

Columnas: Name — Name

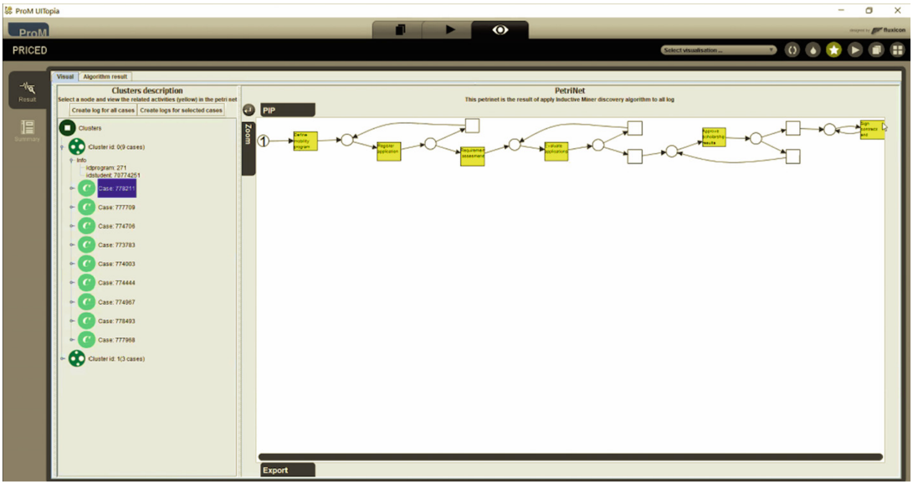
Filas: Id — Id; Idcase — Idcase; Attribute name — Attribute name; Value — Value

Filtro: Existrelation — Existrelation; Idattributeinstance

Información: 13:59 / 5 x 212 / 4.59s

272	0	idcareer	80	0
		idcourse	3366	0
	774444	idcourse	3366	201.732
278	0	idcareer	80	0
		idcourse	8855	0
	774444	idcourse	8855	201.732
287	0	idcareer	59	0
		idcourse	1555	0
	774444	idcourse	1555	201.732
363	0	idcareer	80	0
		idcourse	9966	0
	774706	idcourse	9966	214.049
369	0	idcareer	80	0
		idcourse	5548	0
	774706	idcourse	5548	214.049
371	0	idcareer	80	0
		idcourse	3366	0
	774706	idcourse	3366	214.049
376	0	idcareer	95	0
		idcourse	8888	0
	774706	idcourse	8888	214.049
378	0	idcareer	95	0
		idcourse	3333	0
	774706	idcourse	3333	214.049
383	0	idcareer	95	0
		idcourse	7555	0

Fig. 7. Data warehouse result for courses and careers involved in cases that took more than 15 days to complete.



Clusters description

Select a node and view the related activities (yellow) in the petri net

- Clusters
  - Cluster id: 0(9 cases)
    - Info
      - idprogram: 271
      - idstudent: 70774251
    - Case: 778211
    - Case: 777709
    - Case: 774706
    - Case: 773783
    - Case: 774003
    - Case: 774444
    - Case: 774967
    - Case: 778493
    - Case: 777968
  - Cluster id: 1(3 cases)

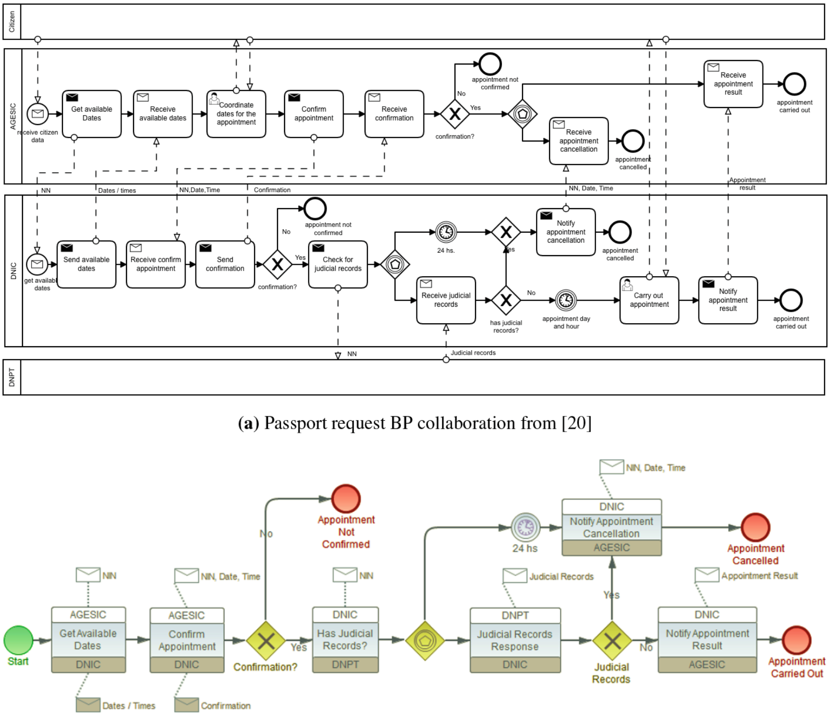
PetriNet

This petrinet is the result of apply Inductive Miner discovery algorithm to all log

Fig. 8. ProM plug-in for integrated process and data mining over integrated data.

4.2 Passport Request BP with Collaborative Extension

The Passport request BP allows a citizen to request a passport interacting with several e-Government organizations. In the first place, the e-Government National Agency (AGESIC) receives the request and interacts with the National Identification Agency (DNIC) to schedule a meeting for issuing the passport. The DNIC interacts with the National Police office (DNPT) to check the Judicial record's background of the citizen. If there is none, the meeting is carried out, and the passport can be issued or not, depending on the defined criteria. If the citizen has judicial records or the response is not received within 24 h, the meeting is canceled. Figure 9a shows the collaborative BP [20] using BPMN 2.0, and Fig. 9b its choreography [19].



(a) Passport request BP collaboration from [20]

(b) Collaborative Passport request BP choreography view from [19]

Fig. 9. Passport request proof of concept.

Execution of the Methodology. In this case, we focus only on the activities we performed for identifying, executing, and evaluating compliance requirements. The rest of the activities for each phase defined in Sect. 2 are the same as in the previous example, i.e., selecting BPs, evaluating data quality, etc.

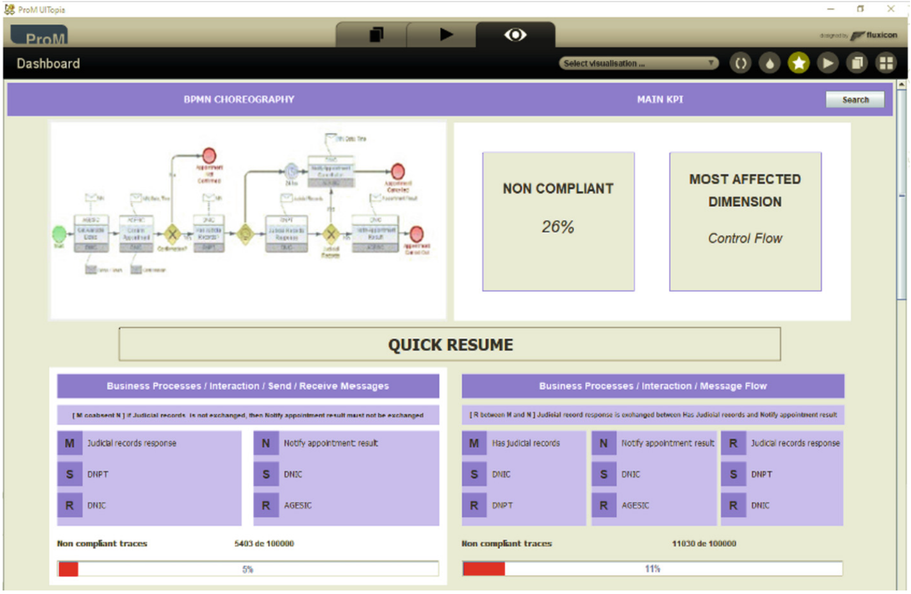
PDC1 - Identify Compliance Requirements. We selected compliance requirements from the BPCRM model to be evaluated over the choreography:

- Dimension: *Interaction*, Factor: *Send/Receive Messages*, Control: *M coabsent N*
- Dimension: *Interaction*, Factor: *Message flow*, Control: *R between M and N*

The first control *M coabsent N* is instantiated over the choreograpy as: If *Judicial records response* is not exchanged, then *Notify appointment result* must not be exchanged, and the second control *R between M and N* is instantiated as: *Judicial records response* is exchanged between *Has judicial records* and *Notify appointment result*.

PDMA3 - Execute Mining/Analysis Approach. The compliance analysis over the extended collaborative event log is implemented in our ProM plug-in, taking as input the compliance requirements for the process, i.e., the instantiation of controls for the specific messages, tasks, etc., and the extended event log for the collaborative BP (collaboration, choreography). In Fig. 10 we present an example of the results. Non-compliant traces are shown in the summary panel with the number and percentage of trace violations. Different control results for the choreography can be seen in [20].

PDC2 - Evaluate Compliance Results. Several traces presented violations regarding the two selected controls. In the first case, a message appeared in some traces where it should not occur since the first message was not present. In the second case, a message was not exchanged in the correct order. It requires looking deeper into the violating traces to gain insight into the causes.



NON COMPLIANT

26%

MOST AFFECTED DIMENSION

Control Flow

Business Processes / Interaction / Send / Receive Messages

[M coabsent N] if Judicial records is not exchanged, then Notify appointment result must not be exchanged

M	Judicial records response	N	Notify appointment result
S	DNPT	S	DNIC
R	DNIC	R	AGESIC

Non compliant traces 5403 de 100000

5%

Business Processes / Interaction / Message Flow

[R between M and N] Judicial record response is exchanged between Has Judicial records and Notify appointment result

M	Has Judicial records	N	Notify appointment result	R	Judicial records response
S	DNIC	S	DNIC	S	DNPT
R	DNPT	R	AGESIC	R	DNIC

Non compliant traces 11030 de 100000

11%

Fig. 10. ProM plug-in for compliance requirements evaluation choreography view.

5 Related Work

CRISP-DM [31], KDD [7], and SEMMA [29] are the most common methodologies for performing classical data-centric analysis. None of them include detailed guidelines on identifying and incorporating data useful to analyze organizations' processes and improve them. CRISP-DM was initially developed in IBM for data mining tasks, and it is used for a wide variety of projects. It consists of a cyclic model with the following defining stages that can be reversed: Business understanding, Data understanding, Data Preparation, Modeling, Evaluation, and Deployment. KDD is a method to guide specialists in extracting patterns and required information from data. It consists of five stages: Selection, Preprocessing, Transformation, Data Mining, and Interpretation/Evaluation. Finally, SEMMA is also a cyclic method that does not focus as heavily on data-specific stages. In this case, a wide range of algorithms or methods are used.

From the business process perspective, in [18], the authors propose PM^2, a methodology to guide the execution of process mining projects with different goal levels. It consists of six stages with their corresponding activities: *planning*, for setting up the project and defining the research questions; *extraction*, for extracting data and process models; *data processing*, for creating appropriate event logs; *mining & analysis*, for applying process mining techniques; *evaluation*, for relating the analysis findings to improvement ideas; and *process improvement & support*, for modifying the actual process execution. 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 related problems as quality assessments. 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, the process improvement stage is considered by integrating an Improvement phase from the BPCIP methodology [16].

Although there are many data quality proposals on data quality methodologies and frameworks, e.g., [3,33], 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 the main tasks of existing approaches to our needs, 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 [21]. The COMPAS project defines a life cycle with four phases (e.g., evaluation) [5]. The C^3 Pro Project describes a design-time methodology for compliance of collaborative workflows [28]. The MaRCo Project defines activities for compliance management [25] (modeling, checking, analysis, enactment). However, they neither consider these activities in the context of an integrated methodology nor leverage process and data mining for compliance control and analysis.

6 Conclusions

We have presented the PRICED methodology to carry out process and data mining and analysis efforts over integrated process data and organizational data. The static view of the methodology includes the definition of disciplines, tasks, roles, and artifacts, and the dynamic view comprises phases and sub-phases to guide the work within the framework. Key elements of our proposal include: (i) a metamodel-based integration of process and organizational data from process engines and distributed organizational DBs; (ii) a quality model for quality assessment over the integrated data; (iii) a compliance requirements model for compliance assessment over collaborative BPs; (iv) extended event logs and a data warehouse to be used for mining/analysis over the integrated data; (v) and integrated process and data mining/analysis approaches over the integrated data to provide a complete view of the organization's actual operation.

Also, we have provided two applications of the methodology. The first one focused on integrated process and organizational data, and the second focused on collaborative BPs. Both applications allowed us to show the utility of the elements defined in the methodology.

We believe it is a valuable tool to guide organizations' mining/analysis efforts towards evidence-based process improvement, with a complete and integrated data view. Nevertheless, we are still improving the whole framework, applying it over more complex processes and heterogeneous organizational data to assess its capabilities. We are also performing further analysis over the integrated data, with different process and data mining approaches.

Acknowledgement. We would like to thank students: Alexis Artus, Andrs Borges, Federico Prez, Francisco Betancor, Fabin Gambetta, Juan Canaparo, Martn Rubio, for their work in the PRICED framework and prototypes.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, 2nd Edn. Springer, Berlin (2016). https://doi.org/10.1007/978-3-662-49851-4
2. Artus, A., Borges, A., Calegari, D., Delgado, A.: Integrated process data and organizational data analysis for business process improvement. In: Golfarelli, M., Wrembel, R., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DaWaK 2021. LNCS, vol. 12925, pp. 207–215. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86534-4_19
3. Batini, C., Scannapieco, M.: Data and Information Quality. DSA, Springer, Cham (2016). https://doi.org/10.1007/978-3-319-24106-7
4. Betancor, F., Pérez, F., Marotta, A., Delgado, A.: Business process and organizational data quality model (BPODQM) for integrated process and data mining. In: Paiva, A.C.R., Cavalli, A.R., Ventura Martins, P., Pérez-Castillo, R. (eds.) QUATIC 2021. CCIS, vol. 1439, pp. 431–445. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85347-1_31
5. Birukou, A., D'Andrea, V., Leymann, F., Serafinski, J., Silveira, P., Strauch, S., Tluczek, M.: An integrated solution for runtime compliance governance in SOA. In: Maglio, P.P., Weske, M., Yang, J., Fantinato, M. (eds.) ICSOC 2010. LNCS, vol. 6470, pp. 122–136. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17358-5_9
6. Bose, R.P.J.C., Mans, R.S., van der Aalst, W.M.P.: Wanna improve process mining results? In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 127–134 (2013)
7. Brachman, R.J., Anand, T.: The process of knowledge discovery in databases. In: Advances in Knowledge Discovery and Data Mining, pp. 37–57. MIT Press, Cambridge (1996)

8. Calegari, D., Delgado, A., Artus, A., Borges, A.: Integration of business process and organizational data for evidence-based business intelligence. CLEI Electron. J. **24**(2), 7:1-7:19 (2021)
9. Chang, J.: Business Process Management Systems: Strategy and Implementation. CRC Press, Boca Raton (2016)
10. Cristalli, E., Serra, F., Marotta, A.: Data quality evaluation in document oriented data stores. In: Woo, C., Lu, J., Li, Z., Ling, T.W., Li, G., Lee, M.L. (eds.) ER 2018. LNCS, vol. 11158, pp. 309–318. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01391-2_35
11. Delgado, A., Calegari, D.: Towards a unified vision of business process and organizational data. In: XLVI Latin American Computing Conference (CLEI), pp. 108–117. IEEE (2020)
12. Delgado, A., Calegari, D.: Discovery and analysis of e-government business processes with process mining: a case study. In: 55th Hawaii International Conference on System Sciences, (HICSS) (2022)
13. Delgado, A., Calegari D., Arrigoni A.: Towards a generic BPMS user portal definition for the execution of business processes. In: XLII Latin American Computer Conference - Selected Papers, CLEI 2016 Selected Papers, Valparaiso, Chile, 10–14 October 2016, pp. 39–59. Elsevier (2016)
14. Delgado, A., Calegari, D., Marotta, A., González, L., Tansini, L.: A methodology for integrated process and data mining and analysis towards evidence-based process improvement. In: Proceedings of the 16th International Conference on Software Technologies (ICSOFT), pp. 426–437. ScitePress (2021)
15. Delgado, A., Marotta, A., González, L., Tansini, L., Calegari, D.: Towards a data science framework integrating process and data mining for organizational improvement. In: 15th International Conference on Software Technologies (ICSOFT), pp. 492–500. ScitePress (2020)
16. Delgado, A., Weber, B., Ruiz, F., de Guzmán, I.G.R., Piattini, M.: An integrated approach based on execution measures for the continuous improvement of business processes realized by services. Inf. Softw. Technol. **56**(2), 134–162 (2014)
17. Dumas, M., van der Aalst, W.M., ter Hofstede, A.H.: Process-Aware Information Systems: Bridging People and Software through Process Technology. Wiley, Hoboken (2005)
18. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.P.: PM^2: a process mining project methodology. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) CAiSE 2015. LNCS, vol. 9097, pp. 297–313. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19069-3_19
19. González, L., Delgado, A.: Towards compliance requirements modeling and evaluation of e-government inter-organizational collaborative business processes. In: 54th Hawaii International Conference on System Sciences, (HICSS), pp. 1–10. ScholarSpace (2021)
20. González, L., Delgado, A.: Compliance requirements model for collaborative business process and evaluation with process mining. In: XLVII Latin American Computing Conference (CLEI) (2021)
21. Hashmi, M., Governatori, G., Lam, H.P., Wynn, M.T.: Are we done with business process compliance: state of the art and challenges ahead. Knowl. Inf. Syst. **57**(1), 79–133 (2018)
22. Hecht, R., Jablonski, S.: Nosql evaluation: a use case oriented survey. In: 2011 International Conference on Cloud and Service Computing, pp. 336–341 (2011)
23. IEEE: Task Force on Data Science and Advanced Analytics. http://www.dsaa.co/
24. IEEE: IEEE standard for extensible event stream (XES) for achieving interoperability in event logs and event streams. In: IEEE Std 1849–2016, pp. 1–50 (2016)
25. Kharbili, M.E., Ma, Q., Kelsen, P., Pulvermueller, E.: CoReL: policy-based and model-driven regulatory compliance management. In: IEEE 15th International Enterprise Distributed Object Computing Conference, IEEE, August 2011

26. Khasawneh, T.N., AL-Sahlee, M.H., Safia, A.A.: Sql, newsql, and nosql databases: a comparative survey. In: 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 013–021 (2020)
27. Knuplesch, D., Reichert, M.: A visual language for modeling multiple perspectives of business process compliance rules. Softw. Syst. Model. **16**(3), 715–736 (2016). https://doi.org/10.1007/s10270-016-0526-0
28. Knuplesch, D., Reichert, M., Ly, L.T., Kumar, A., Rinderle-Ma, S.: Visual modeling of business process compliance rules with the support of multiple perspectives. In: Ng, W., Storey, V.C., Trujillo, J.C. (eds.) ER 2013. LNCS, vol. 8217, pp. 106–120. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41924-9_10
29. Mariscal, G., Marbán, O., Fernández, C.: A survey of data mining and knowledge discovery process models and methodologies. Knowl. Eng. Rev. **25**(2), 137–166 (2010)
30. Papazoglou, M.P.: Making business processes compliant to standards and regulations. In: 15th International Enterprise Distributed Object Computing Conference, IEEE, August 2011
31. Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J. Data Warehouse. **5**(4), 13–22 (2000)
32. Sumathi, S., Sivanandam, S.N.: Introduction to Data Mining and its Applications, Studies in Computational Intelligence, vol. 29. Springer, Berlin (2006)
33. Tepandi, J., et al.: The Data Quality Framework for the Estonian Public Sector and Its Evaluation. In: Hameurlain, A., Küng, J., Wagner, R., Sakr, S., Razzak, I., Riyad, A. (eds.) Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXV. Lecture Notes in Computer Science(), vol. 10680, pp. 1–26. Springer, Berlin (2017). https://doi.org/10.1007/978-3-662-56121-8_1
34. Valverde, M.C., Vallespir, D., Marotta, A., Panach, J.I.: Applying a data quality model to experiments in software engineering. In: Indulska, M., Purao, S. (eds.) ER 2014. LNCS, vol. 8823, pp. 168–177. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12256-4_18
35. Verhulst, R.: Evaluating quality of event data within event logs:an extensible framework. Master's thesis, Eindhoven University of Technology (2016)