Towards a Data Science Framework Integrating Process and Data Mining for Organizational Improvement

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

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

Abstract: Organizations face many challenges in obtaining information and value from data for the improvement of their operations. For example, business processes are rarely modeled explicitly, and their data is coupled with business data and implicitly managed by the information systems, hindering a process perspective. This paper presents a proposal of a framework that integrates process and data mining techniques and algorithms, process compliance, data quality, and adequate tools to support evidence-based process improvement in organizations. It aims to help reduce the effort of identification and application of techniques, methodologies, and tools in isolation for each case, providing an integrated approach to guide each operative phase, which will expand the capabilities of analysis, evaluation, and improvement of business processes and organizational data.

1 INTRODUCTION

Over the last years the ”data explosion” phenomenon characterized by the amount of data available in internet and organizations, from several sources such as personal/enterprise computers, social media, digital cameras, servers, sensors, and others, has been impacting the world and the way data is perceived, stored, collected and analyzed (van der Aalst, 2016). Organizations face many challenges in managing these large volumes of data, being one of the most important ones to obtain information and value from the data in their information systems.

Although business processes (BPs) are the basis for the operation of organizations no matter which is their domain (i.e. banking, health, e-government) they are rarely modeled explicitly to guide the activities to perform, and they are implicitly stored and managed within the organizational information systems and associated with the business data. Both organizations and their processes, as well as the software systems that support such processes and data, are increasingly complex, defining ecosystems in which it is necessary to integrate different visions, techniques, and tools for the management of information, processes, and associated systems.

Data science (van der Aalst, 2013; IEEE, 2020) has emerged in recent years as a discipline in itself, interdisciplinary, to respond to the problem of management, analysis and discovery of information in large volumes of data that are generated at high speed (velocity) and with great variety (the three V) (Furht and Villanustre, 2016), also considering the veracity of the data (Ong et al., 2016), which is stored in structured or unstructured form. Organizations are increasingly incorporating tools and techniques for managing and analyzing the large volumes of data they have, but due to the variety of approaches, tools, and objectives, they often lack conceptual and objective guides that allow them to identify the solutions that best suit their needs and capabilities.

In this context, the compartmentalized vision of processes on the one hand and organizational data on the other are not adequate to provide the organization with the evidence-based business intelligence necessary to improve their daily operation. What is more, in inter-organizational collaborative environments business processes include several participants with their own internal processes (orchestrations) with their own internal data, which makes the scenario of data integration and analysis more complex. Also, a key element in data manipulation, both of the event logs from processes execution and of the organizational data that these processes manipulate, refers to their quality analysis, data cleaning, and assuring that the data analyzed complies with a minimum quality, in different dimensions. In light of the above, one of the main lines of research that remains

Delgado, A., Marotta, A., González, L., Tansini, L. and Calegari, D.
Towards a Data Science Framework Integrating Process and Data Mining for Organizational Improvement.
DOI: 10.5220/0009875004920500
In *Proceedings of the 15th International Conference on Software Technologies (ICSOFT 2020)*, pages 492-500
ISBN: 978-989-758-443-5

open in this area refers precisely to the integrated support for the analysis of processes and data in organizations.

This paper presents a proposal of an integrated framework for organizational Data Science, that includes process and data mining techniques and algorithms, the integration of process and organizational data, data quality assessment, process compliance assessment, methodologies and guides to support all of the above, and adequate tool support, for the improvement of organizations based on evidence. The main objective of this framework is to help reduce the effort of identification and application of techniques, methodologies, and tools in isolation for each case, providing an integrated package to guide each phase of the data analytic operation, which will expand the possibilities of analysis, evaluation, and improvement of the organizational business processes and corresponding data.

The main contributions of our work are as follows: i) an integrated view and complete support (elements mentioned above) for the manipulation and analysis of process and organizational data, that will serve as a basis to guide analytic efforts in organizations, ii) models and tools for process and organizational data integration, from different sources and scenarios, iii) models and tools for process and organizational data quality assessment and improvement, iv) adapted and new techniques and algorithms for integrated process and data mining analysis over the integrated data within different scenarios, and corresponding tool support, v) models, techniques, algorithms, and tools to support compliance analysis on business processes, over different scenarios.

As a research methodology, we follow Design Science guidelines (Hevner et al., 2004; Wieringa, 2014), where knowledge and understanding of a problem and its solution are based on two main processes: building and assessment (of the application) of an artifact. Artifacts that are useful to solve problems not yet solved are built, and they are assessed with respect to their usefulness in the solution of the defined problem (Hevner et al., 2004). For the evaluation of artifacts we will carry out experimentation on algorithms and their results as we build them, Action-Research (Iivari and Venable, 2009) and case study research (Yin, 2014) within the organization, to validate artifacts and the proposal with our counterpart. We are working with a team from the e-Government in our countrywhich has real processes and organizational data for our research work. We also carried out a systematic literature review (Kitchenham, 2004; Kitchenham and Charters, 2007) at the beginning of our research, to review existing work on the integrated view we are proposing and the main sub-topics. To the best of our knowledge, there are no other initiatives that integrate all of the dimensions of process and data analysis as we are in our framework.

The rest of the article is organized as follows: In Section 2 we introduce key concepts related to the main elements included in our proposal. In Section 3 we discuss related work. In Section 4 we describe our proposal including the definition of the framework and the main elements it comprises, as well as preliminary results. Finally in Section 5 we present some conclusions and future work.

2 BACKGROUND

Business Process Management (BPM) (van der Aalst et al., 2003; Weske, 2019; Dumas et al., 2018) refers to the activities that organizations perform for the explicit management and improvement of their business processes according to their organizational needs. In these terms, a business process (BP) is a set of activities carried out in coordination in an organizational and technical environment, to achieve a business objective (Weske, 2019). Its life cycle (e.g.: analysis & design, configuration, execution, and evaluation phases (Weske, 2019)) is usually supported by a Business Process Management System (BPMS) (Chang, 2016).

Process discovery is a complex task, especially when trying to describe not only the activities but also the participants and resources involved in a BP. In this context, organizations not only use strategies based on interviews with process participants, but also automatic methods based on learning from the information systems that support BPs. Process mining (van der Aalst, 2016) exploits the data registered by such information systems when supporting the real executions of BPs to discover process models. Complimentary, process archaeology (Pérez-Castillo et al., 2011) can be used to extract information from the source code of such information systems, when available.

Using runtime information from information systems it is possible not only to describe a BP, but also to verify the compliance of the enacted BP concerning the expected one (the one that can be modeled from interviews). Moreover, it is possible to obtain key execution measures, e.g., about bottlenecks, used resources, time duration, etc. In this context, in previous works (Delgado et al., 2014; Delgado et al., 2012) we have presented a framework and methodology for BPs continuous improvement to define and analyze execution measures with the Business Process Exe-

cution Measurement Model (BPEMM), including a plug-in for the ProM tool[1].

In turn, compliance management aims to ensure that organizations act following multiple established regulations (e.g. laws, standards) (Tran et al., 2012). It comprises several activities including the modeling, implementation, maintenance, verification, and reporting of compliance requirements (Ramezani et al., 2012)(El Kharbili, 2012). In particular, compliance control involves assessing the fulfillment of such requirements and acting accordingly. In general, most current approaches control compliance at design time, execution (i.e. runtime), or after execution (Hashmi et al., 2018). Also, compliance controls may be preventive, detective, or corrective (Elgammal et al., 2016).

To monitor processes execution including process compliance, BPMS platforms may be integrated with middleware infrastructures such as the enterprise service bus (ESB) (González and Ruggia, 2011), and complex event processing (CEP) engines (Flouris et al., 2017), for example, to signal an alarm when a violation of policies occur during process execution. Furthermore, the traceability of collaborative BPs between participants is another important element for the discovery as well as monitoring and analysis of processes execution (Delgado et al., 2017).

Another perspective on the operation of the organization can be obtained by analyzing the data involved in the execution of these BPs, adding the extra information on when, how and by whom these data were created, modified, deleted, etc.

Data mining techniques allow exploring large databases to find repetitive patterns, trends, or rules that explain the behavior of the data in a given context (Sumathi and Sivanandam, 2006). Given the large amount of data that organizations generate in their daily activity and the need to take advantage of it, data mining techniques have become fundamental tools to assist in business decision making involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. A wide range of algorithms or methods is used to carry out data mining functions based on data mining techniques. For example, the Apriori algorithm, Naıve Bayesian, k-Nearest Neighbour, k-Means, CLIQUE, STING, etc. (Gupta and Chandra, 2020). Data mining has been used in a variety of domains, such as time-series data mining, web mining, temporal data mining, spatial data mining, tempo-spatial data mining, educational data mining, business, medical, science, and engineering, etc. Each domain can have one or more applications of data mining (Han et al., 2011).

Finally, a key element in data manipulation, both of the event logs from processes execution and of the data that these processes manipulate, refers to their quality. As remarked in (van der Aalst, 2016), data quality is of great importance in process mining, since its results are less valuable if the data is not complete enough or trustful. The author focuses on event logs data quality, providing some basic guidelines for addressing this problem. Data quality evaluation, data cleaning, and enforcement of a minimum quality of the managed data, according to several quality dimensions, are the kind of tasks that should be present in this context. Quality management in a data set involves the complex tasks of evaluating, improving, and monitoring its data quality (Batini and Scannapieco, 2016). To carry out these tasks it is necessary to define a quality model that works as a base and conducts all the processes involved. A quality model is a set of quality dimensions and metrics, where the former represent general aspects of data quality and the latter define how these dimensions are measured to evaluate the quality in a particular data set.

3 RELATED WORK

Although process mining (van der Aalst, 2016) and data mining (Sumathi and Sivanandam, 2006) are extensive research areas in which many techniques, algorithms, and tools are being currently developed, the exploitation of both process data and organizational data altogether has not been analyzed much yet. When dealing with process execution, the problem is mostly observed from the perspective of process mining. In (van der Aalst and Damiani, 2015) the relation between data science and process science through process mining is explored, and in (van der Aalst, 2013) a process cube is defined to analyze and explore processes interactively based on a multidimensional view on event data.

In line with our interests, in (de Murillas et al., 2019) the authors propose a comprehensive integration of process and organizational data in a consistent and unified format through the definition of a metamodel. However, they focus on the extraction of read/write event logs from a database, thus business-level activities are hidden, and the analysis is focused on the lower level of database operations. In (Tsoury et al., 2018) the authors discuss the aforementioned problem and define a conceptual framework for a deep exploration of process behavior, combining information from three sources: the event log (business-level), the database (low level), and the transaction

[1]ProM: http://www.promtools.org/

(redo) log, as we do, but they do not provide a uniform way of expressing all the information. Finally, in (Radeschütz et al., 2008; Radeschütz et al., 2015) the authors describe concrete matching techniques between process and organizational data, that are later integrated into a business impact analysis framework based on a data warehouse. To support our vision, we are extending the unified model for process execution in (Delgado et al., 2016) to include mining concepts, and link it with other metamodels such as for business data (de Murillas et al., 2019), inter-organizational collaborative processes (Delgado et al., 2020), and process compliance (González and Ruggia, 2018).

In turn, during the last two decades, a large body of knowledge has been developed in the field of business process compliance, mostly focusing on controlling compliance within intra-organizational processes (Fdhila et al., 2015), at design time and runtime (Fellmann and Zasada, 2014)(Hashmi et al., 2018). The COMPAS project defined a model-driven approach for runtime compliance governance in the context of a process-driven SOA (Tran et al., 2012). The approach proposed languages and tools for modeling compliance requirements, linking them to business processes, monitoring process execution using CEP, displaying the current state of compliance, and analyzing cases of non-compliance (Birukou et al., 2010). The C^{3}Pro Project focused on providing a theoretical framework for enabling change and compliance of collaborative business processes, at design time, runtime, and a-posteriori (i.e. after execution) by processing execution logs (Knuplesch et al., 2017). Finally, in our previous work, we proposed a policy-based approach to compliance management within inter-organizational integration platforms (González and Ruggia, 2018). This approach enables compliance control at runtime in collaborative business processes, by leveraging an integration platform and a compliance policy language (PL4C)(González and Ruggia, 2018). This language enables specifying how the platform has to control compliance requirements for each one of the processes.

Regarding data quality, in recent years a wide set of data quality dimensions has been defined, currently, there is a sub-set used by most of the authors (Scannapieco and Catarci, 2002; Shankaranarayanan and Blake, 2017), but without reaching total agreement about the set of dimensions that characterize data quality. In (Batini and Scannapieco, 2016) the existing quality dimensions are organized and 6 clusters have proposed that try to cover the main dimensions: accuracy, completeness, redundancy, readability, accessibility, consistency, usefulness and trust. The definition of quality dimensions for process mining, focusing only on event logs data, was also studied. In (Verhulst, 2016) the author proposes a set of dimensions for a generic model, discarding the dimensions that depend on specific domains or users. These dimensions are selected taking into account previous proposals and following the guidelines proposed in (van der Aalst, 2016). In (Andrews et al., 2019) an approach for process mining and practical experience is presented, where data quality is an essential step, and certain dimensions and metrics are selected.

4 FRAMEWORK PROPOSAL

The framework integrates process and data mining techniques and algorithms for the analysis of process execution and organizational data, and tool support, to help improve an organization's operation based on evidence. We have named it PRICED for Process and Data sCience for oRganIzational improvEment, and although we integrate elements for intra-organizational business processes (orchestrations) we focus on inter-organizational collaborative business processes.

4.1 Framework Definition

The framework defines a general strategy including methodologies, techniques, and tools, both existing and new, to provide organizations with key elements to analyze their processes and data in an integrated manner. The framework will help organizations reducing the effort of identifying and applying suitable techniques, methodologies, and tools to analyze operational data (event logs and organizational data) in order to evaluate and improve their daily operation. It provides an integrated and accessible package of proposals for each operative phase, which will translate in better possibilities for analysis, evaluation, and improvement of processes and related data in organizations. In the following, we describe the dynamic and static views of the framework.

4.1.1 Dynamic View

Figure 1 presents the dynamic view of the framework including the three phases we have defined.

In the Enactment Phase, several different systems are operating in the organization, which can be categorized in two main types: i) Systems that are Process-Aware (PAIS) where business processes are explicit and generally enforced within a process engine, and ii) traditional Systems where processes are

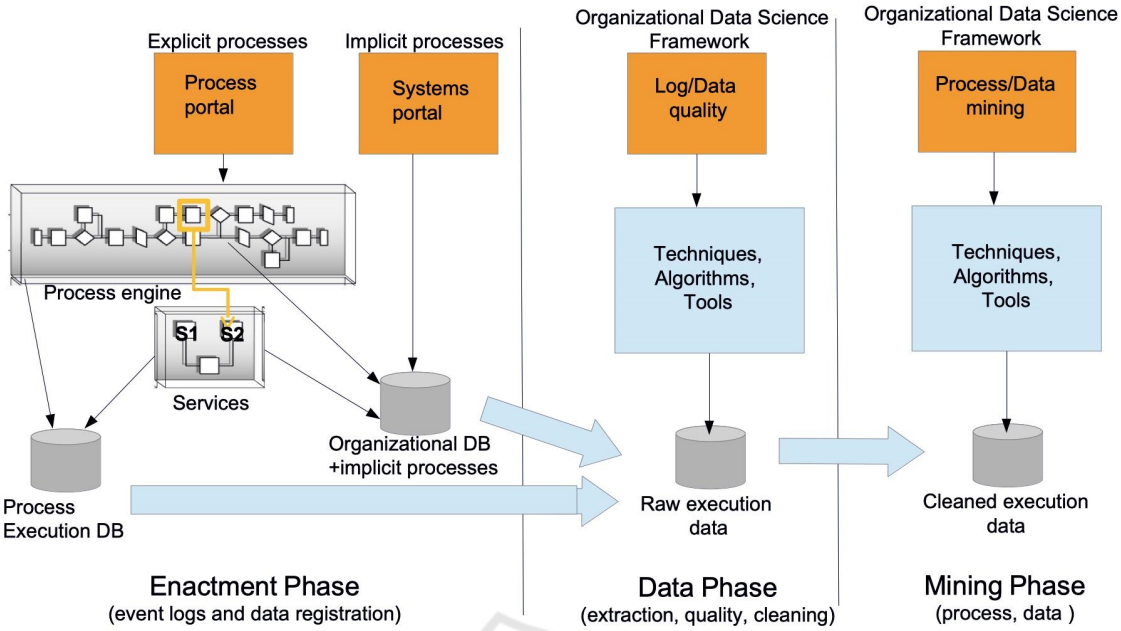

Figure 1: Framework proposal Phases.

implicitly defined and embedded in it. Traces of process execution (user tasks, services, business rules) are registered in the process engine database (i.e. in a BPMS), whereas organizational data are registered in the organizational database, along with data from the implicit processes.

Although some organizational data is registered in the process engine database, the complete data is often implemented within activities and registered directly into the organizational database, without knowledge of the process engine. Then, at least two (internal) data sources should be taken into account as input for analysis and evaluation of organizational processes and data. These sources are not automatically connected (i.e. records in a process - event log- and the business data that flows with it - organizational data-) for which the first challenge to tackle refers to linking enhanced event logs with the corresponding data in the organizational database.

The Data Phase deals with all aspects of data preparation, in order to be used as input in the next phase for process and data mining. The first step refers to extract data from the sources and put it together in event logs and database query results. After the data is in place and in the correct format, data quality aspects are reviewed, in order to remove undesirable elements before the mining phase, cleaning the data. Regarding event log data quality aspects we consider an existing work (Verhulst, 2016), and for data quality aspects we integrate a data quality framework already defined within the participating research groups. Finally, in the Mining Phase an integrated view on process and data mining is used, to provide organizations with the complete information regarding the operation of their processes and the associated data. For this, we are working on mining both processes and data based on existing algorithms and techniques, enhanced with the correlation of data and their visualization in an integrated manner.

4.1.2 Static View

The framework comprises seven dimensions in which elements are defined. These elements are used within the phases to go from input data to output information and business value regarding the real operation of the organization. In Figure 2 these dimensions are presented.

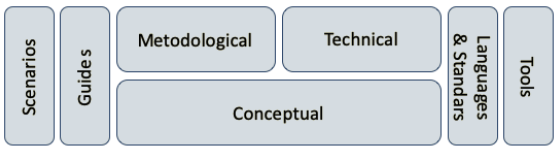

Figure 2: Framework proposal dimensions.

Conceptual Dimension: includes the definition of key concepts for process and data mining, data quality, and process compliance, that are used within the framework. Elements such as process traces, event logs, quality dimensions, policies specification, techniques, and algorithms for process and data mining such as process discovering and conformance, data clustering analysis, decision trees, regression, among

others, are included with detailed definitions and examples.
Technical Dimension: builds upon the conceptual dimension, and includes an exhaustive list of approaches, techniques, and algorithms used for process and data mining and their categorization, including description and operation of each one, data quality, and process compliance approaches and techniques.
Methodological Dimension: it also builds upon the conceptual dimension and provides methodological guides to carry out process and data mining activities, data quality activities, and process compliance activities. It defines support processes, roles, artifacts, and guides for the selection and use of techniques and algorithms, among others. All processes will be specified in the Eclipse Process Framework (EPF) [2] and published on the web site of the framework.
Languages & Standards: includes a description of languages and standards that are used for process and data mining, such as MXML and XES, BPMN 2.0, Petri nets, and others for data quality and process compliance.
Tools: includes a list and description of existing tools (open source and proprietary) that provide support for process and data mining, such as the ProM framework or Disco, and tools for data quality and process compliance.
Guides: includes guides, templates, FAQs, best practices, and general knowledge management to support carrying out the activities defined within the framework, as well as related artifacts and documents.
Scenarios: in this dimension, scenarios, and examples of different identified use cases are provided, in order to illustrate the adoption of the framework in organizations.

4.2 Preliminary Results

The main preliminary result is the definition and conceptualization of the framework itself, its phases, and dimensions, as presented above. We have identified several scenarios for the integration of process and organizational data, which includes intra-organizational processes (orchestrations) and inter-organizational collaborative processes, and defined and initial metamodel to support this integration. In Figure 3 b) we present the initial definition of the integrated metamodel which extends existing metamodels from the previous works (Delgado et al., 2016; Delgado et al., 2020).

The integrated metamodel is composed of a process view and a data view, both of them with two levels of information: definition of elements and their instances. The left-hand side focuses on the definition of elements such as process which are composed of process elements (tasks, messages, etc.), variables, roles, the variables that can refer to data entities composed by attributes. The right-hand side focuses on the instances of elements defined in the left-hand side, such as cases (process instances) which are composed of element instances, variable instances connected with entity and attribute instances, and users, which are related within each other in the same way as their definitions. These elements specify values that evolve over time. The ElementDefinition concept is used to connect this metamodel with the specification of processes (e.g. BPMN 2.0) and compliance requirements (described below). The definition of this metamodel is part of ongoing work.

The aim of this metamodel is to build an extended event log which contains not only the traditional process data regarding process execution and related variables, such as case id, tasks (business tasks), events (start, complete), timestamps, the originator (resource), variables data, but also elements from services execution (internal, external), organizational data objects (in external BDs such as client, order, etc.), messages exchanged between process participants and the associated data, among others. For doing this, we added concepts for data elements definition (Entity, Attribute) related to the corresponding instances (EntityInstance, AttributeInstance) in the organizational database. We are automating the integration of data from all the sources mentioned to populate the integration metamodel, as well as algorithms to generate and analyze the extended event log from these integrated data, including inter-organizational collaborative process records with several participants.

Regarding business process compliance, we are working on processing event logs in a post mortem fashion in order to analyze each case execution, and checking whether it presents a violation of the compliance requirements that were specified for the process. We are exploring the definition of clusters of traces that presents the same behavior with respect to the compliance requirements, in order to further analyze the causes of the violations (i.e. by the organization, employee, among others) as well as to be able, for example, to perform preventive actions. For compliance elements definition we use PL4C specifications (González and Ruggia, 2018).

On the modeling side, we are extending the Business Process Model and Notation (BPMN 2.0) to specify compliance requirements directly over BPMN 2.0 business processes and choreographies, with a focus on inter-organizational collaborative business

[2] https://www.eclipse.org/epf/

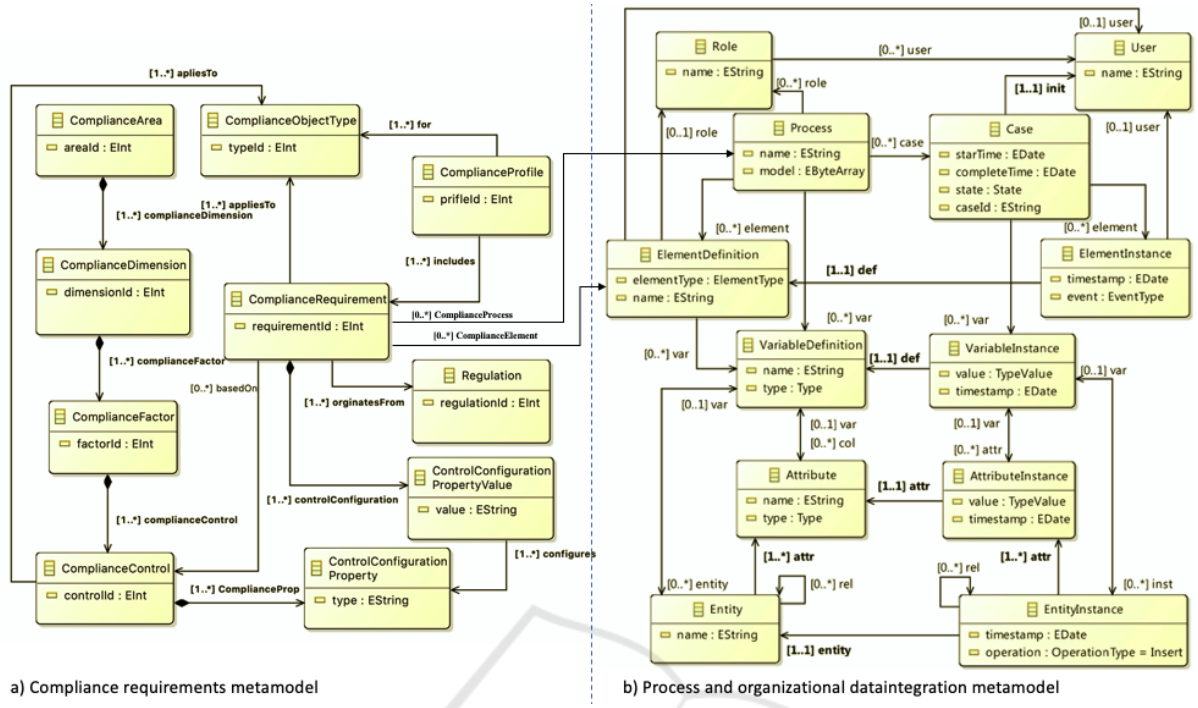

Figure 3: Defined metamodels for: a) compliance requirements, b) process and organizational data integration.

processes. Figure 3 a) presents the compliance requirements metamodel defined in the context of (González and Ruggia, 2018), where the ComplianceRequirements element references the ElementDefinition concept which abstracts the main modeling elements from BPMN 2.0. Based on this specification, we will automatically generate the PL4C specifications which would enable compliance control not only at runtime but also after execution, by processing event logs with our process mining approach.

Compliance models define the focal compliance areas (e.g. Quality of Service, Data quality) and relevant characteristics within these areas (e.g. availability, completeness). Compliance requirements specify general requirements (e.g. response time greater than 1ms) applicable to specific objects types (e.g. operation, service). Compliance profiles define a set of requirements for the same object type enabling a more agile specification of a set of requirements for different objects of the same type. Applicable regulations are also managed, which allows relating requirements with the regulations from which they come from.

Data quality dimensions and factors are mostly universal, but depending on the specific data under analysis some aspects will be more important than others. As we are working with an extended event log, we consider several elements that are not usually present in a traditional event log, such as organizational data, services, messages, etc. We are working on defining the specific dimensions and factors to integrate into the framework for the extended event log. These elements will be added to the previous metamodel to specify quality requirements over the log.

5 CONCLUSION

We have presented a proposal towards an integrated framework that helps to analyze execution data in an integrated manner, both from processes and organizational data that are handled by those processes, with a focus on inter-organizational collaborative processes. The framework aims to support and guide organizations in the complete process of analyzing their data, from data extraction, data quality assessment, data format and selection, data integration, application of process and data mining techniques and algorithms, and tool support.

Although initial definitions and conceptualizations have been made for the framework proposal which we presented here, many challenges remain. We are working on obtaining an integrated vision of execution data from any source within the organization and from other participant organizations, and how to apply process and data mining techniques to the extended execution log we are building. For doing so, we are extending our previously defined metamodels to provide support for that integrated view, includ-

ing adding specific data quality elements and process compliance elements to analyze processes behavior.

We believe the framework will help organizations in getting the most of their data, in an integrated manner, and to use the best tools to support the activities within each phase, which will be accessible within the framework.

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

Andrews, R., Wynn, M., Vallmuur, K., ter Hofstede, A., Bosley, E., Elcock, M., and Rashford, S. (2019). Leveraging data quality to better prepare for process mining: An approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in queensland. *Int. Journal Environment Research and Public Health*, 16(7):1138.

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 *SOC*, pages 122–136. Springer.

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

de Murillas, E. G. L., Reijers, H. A., and van der Aalst, W. M. P. (2019). Connecting databases with process mining: a meta model and toolset. *Software and Systems Modeling*, 18(2):1209–1247.

Delgado, A., Calegari, D., and Arrigoni, A. (2016). Towards a Generic BPMS User Portal Definition for the Execution of Business Processes. *Electronic Notes in Theoretical Computer Science*, 329:39 – 59. CLEI 2016 - The Latin American Computing Conference.

Delgado, A., Calegari, D., González, L., Montarnal, A., and Benaben, F. (2020). Towards a metamodel supporting e-government collaborative business processes management within a service-based interoperability platform. In *The 53rd Hawaii International Conference on System Sciences (HICSS-53)*.

Delgado, A., González, L., and Calegari, D. (2017). Towards setting up a collaborative environment to support collaborative business processes and services with social interactions. In *Service-Oriented Computing - ICSOC Work. and Sat. Events, Revised Selected Papers*, volume 10797, pages 308–320. Springer.

Delgado, A., Ruiz, F., de Guzmán, I. G.-R., and Piattini, M. (2008-2012). MINERVA: Model drIveN and sErvice oRiented framework for the continuous business process improVement & relAted tools. http://alarcos.esi.uclm.es/MINERVA/.

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., Rosa, M. L., Mendling, J., and Reijers, H. A. (2018). *Fundamentals of BPM, 2nd Edition*. Springer.

El Kharbili, M. (2012). Business process regulatory compliance management solution frameworks: A comparative evaluation. In *Procs. Eighth Asia-Pacific Conf. on Conceptual Modelling*, APCCM '12, pages 23–32. Australian Comp. Soc., Inc.

Elgammal, A., Turetken, O., van den Heuvel, W.-J., and Papazoglou, M. (2016). Formalizing and appling compliance patterns for business process compliance. *Software & Systems Modeling*, 15(1):119–146.

Fdhila, W., Rinderle-Ma, S., Knuplesch, D., and Reichert, M. (2015). Change and compliance in collaborative processes. In *2015 IEEE International Conference on Services Computing*. IEEE.

Fellmann, M. and Zasada, A. (2014). State of the art of business process compliance approaches. a survey. In *22nd European Conference on Information Systems*.

Flouris, I., Giatrakos, N., Deligiannakis, A., Garofalakis, M. N., Kamp, M., and Mock, M. (2017). Issues in complex event processing: Status and prospects in the big data era. *Journal of Systems and Software*, 127:217–236.

Furht, B. and Villanustre, F. (2016). Introduction to big data. In Furht, B. and Villanustre, F., editors, *Big Data Technologies and Applications*, pages 3–11. Springer.

González, L. and Ruggia, R. (2011). Addressing qos issues in service based systems through an adaptive ESB infrastructure. In *Proc. 6th Workshop on Middleware for Service Oriented Computing, MW4SOC 2011*, page 4. ACM.

González, L. and Ruggia, R. (2018). A comprehensive approach to compliance management in inter-organizational service integration platforms. In *Procs. 13th International Conference on Software Technologies, ICSOFT 2018*, pages 722–730. SciTePress.

González, L. and Ruggia, R. (2018). Policy-based compliance control within inter-organizational service integration platforms. In *Procs. 11th Conference on Service Oriented Computing and Applications (SOCA)*. IEEE.

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

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

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.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004).

Design science in information systems research. *MIS Quarterly*, 28(1):75–105.

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

Iivari, J. and Venable, J. (2009). Action research and design science research - seemingly similar but decisively dissimilar. In *17th European Conference on Information Systems, ECIS, Verona, Italy*, pages 1642–1653.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report TR/SE-0401, Keele University.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, EBSE.

Knuplesch, D., Reichert, M., and Kumar, A. (2017). A framework for visually monitoring business process compliance. *Information Systems*, 64:381–409.

Ong, K.-L., De Silva, D., Boo, Y. L., Lim, E. H., Bodi, F., Alahakoon, D., and Leao, S. (2016). *Big Data Applications in Eng. and Science*, pages 315–351. Springer.

Pérez-Castillo, R., de Guzmán, I. G. R., and Piattini, M. (2011). Business process archeology using MARBLE. *Journal of Information and Software Technology*, 53(10):1023–1044.

Radeschütz, S., Mitschang, B., and Leymann, F. (2008). Matching of process data and operational data for a deep business analysis. In *Procs. 4th Int. Conference on Interoperability for Enterprise SW and Applications, IESA*, pages 171–182. Springer.

Radeschütz, S., Schwarz, H., and Niedermann, F. (2015). Business impact analysis - a framework for a comprehensive analysis and optimization of business processes. *Computer Science and Research Development*, 30(1):69–86.

Ramezani, E., Fahland, D., van der Werf, J. M., and Mattheis, P. (2012). Separating compliance management and bpm. In *BPM Workshops*, pages 459–464. Springer.

Scannapieco, M. and Catarci, T. (2002). Data quality under a computer science perspective. *Archivi & Computer*, page 2:1–15.

Shankaranarayanan, G. and Blake, R. (2017). From content to context: The evolution and growth of data quality research. *J. Data and Inf. Quality*, 8(2):9:1–9:28.

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

Tran, H., Zdun, U., Holmes, T., Oberortner, E., Mulo, E., and Dustdar, S. (2012). Compliance in service-oriented architectures: A model-driven and view-based approach. *Information and Software Technology*, 54(6):531–552.

Tsoury, A., Soffer, P., and Reinhartz-Berger, I. (2018). A conceptual framework for supporting deep exploration of business process behavior. In *Conceptual Modeling - 37th Int. Conference, ER 2018, Procs.*, volume 11157 of *LNCS*, pages 58–71. Springer.

van der Aalst, W. M. P. (2013). Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining. In *Asia Pacific BPM Conf. AP-BPM, Selected Papers*, volume 159 of *LNBIP*, pages 1–22. Springer.

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

van der Aalst, W. M. P. and Damiani, E. (2015). Processes meet big data: Connecting data science with process science. *IEEE Transactions on Services Computing*, 8(6):810–819.

van der Aalst, W. M. P., ter Hofstede, A. H. M., and Weske, M. (2003). Business process management: A survey. In *Business Process Management, International Conference, BPM 2003, Proceedings*, volume 2678 of *LNCS*, pages 1–12. Springer.

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

Weske, M. (2019). *BPM - Concepts, Languages, Architectures, 3rd Edition*. Springer.

Wieringa, R. J. (2014). *Design Science Methodology for Inf. Systems and Software Engineering*. Springer.

Yin, R. K. (2014). *Case Study Research: Design and Methods, 5th edition*. SAGE Publications, Inc.