Process Mining in Software Process Improvement, a Systematic Literature Review

P. E. Velazquez-Solis, B. L. Flores-Rios, F. F. Gonzalez-Navarro
Institute of Engineering
Autonomous University of Baja California
Mexicali, Mexico
{paola.velazquez, brenda.flores, fernando.gonzalez}@uabc.edu.mx

M. A. Astorga-Vargas, J. E. Ibarra-Esquer and C. Hernández-Castro
Faculty of Engineering Mexicali Campus
Autonomous University of Baja California
Mexicali, Mexico
{angelicaastorga, jorge.ibarra, a172106}@uabc.edu.mx

Abstract— **In this review, we observed there is a growing interest on the strategy of applying the discipline of Process Mining in Software Process Improvement projects. Process Mining aims to discover, monitor and improve processes through analysis of the various event logs generated by organizational processes. In order to identify the predominant contributions, the results were classified in three categories: Theoretical Foundations and Algorithms, Proposals and Tools. We found emerging methodologies and tools with characteristics that enable the application of this discipline into Software Improvement projects.**

Keywords—Process Mining, Software Process Improvement Project, Systematic Literature Review.

I. Introduction

Software Process Improvement (SPI) has become the most recommended discipline to drive the increase of business competitiveness [1]. The successful implementation of SPI has provided important benefits to software organizations; increase in customer satisfaction, efficiency in development process, definition and accomplishment of quality objectives among others [2].

In an SPI project, the assessment process establishes that information must be compiled, registered, stored, processed, analyzed and presented via tools based on paper or software. Mainly, software tools are a valuable support for evaluators allowing them to compile, register and classify evidence especially when volume and information complexity are significant [3]. However, the amount of information increases for each process or improvement cycle, causing scalability or evidence control problems. In general, size of knowledge databases or repositories of processes execution logs are based on criteria of capacity and efficiency of storage and not on the future usage, exploration, exploitation or analysis of those logs. Hence the limited access to such information prevents the enrichment of criteria for decision making needed in the discovery of new business rules, continuous improvement, monitoring and control of the already established processes [4].

This document presents by means of a Systematic Literature Review (SLR), how Process Mining (PM) can provide techniques, algorithms or tools when an SPI project has already generated a considerable volume of information. Thus, deployed processes and information would be perfectly aligned to the related requirements with norm compliance, efficiency and customer support [5].

This document is structured as follows: In section two, we propose a conceptual framework with information regarding Data Mining, PM and SPI; Section three describes the research methodology; to later address, in section four, the results on the interrelationship of PM with the improvement of software processes based on CMMI-DEV. Finally, conclusions and future work are presented.

II. State f Art

Data Mining and PM are common disciplines that provide the techniques and tools for such analysis, making use of algorithms that search for patterns and relationships. For this reason, it is important to present the formal definition and properties of both disciplines.

The Statistical Analysis Systems Institute defines the term Data Mining as the process of selecting, exploring, modifying, modeling and valuating huge amounts of data with the goal of discovering unknown patterns which can be used as a competitive advantage for an organization [6]. Nowadays, it is used to refer to a set of global processes for knowledge discovery from data and data analysis algorithms.

Table I shows definitions of PM found in literature. In general, it is specified as a discipline which aims to discover, monitor, and improve processes through the extraction of knowledge from event logs in information systems [6]. Event logs relate to the execution of organizational processes history, where instances or process cases are found. Case studies were found about the deployment of PM in hospital processes [4], industrial systems [7], government dependencies, biology, medicine, economy, among others. These are available in Process Aware Information Systems (PAIS), as are in workflows, BPMS, ERP or CRM [6].

According to Aguirre Mayorga [8], PM provides tools (ProM, CPN, DisCo) and techniques that help organizations to: 1) Discover the real process execution model; 2) Determine if the process complies with the regulation and documented procedures; 3) Analyze the interaction of staff running the process; 4) Discover bottlenecks; 5) Predict the cycle time of a

TABLE I
LIST OF PROPOSALS FOR THE DEFINITION OF PROCESS MINING

Year	Author	Definition
2011	van der Aalst	Discipline that allows to discover, monitor and improve processes identified in information systems held by organizations from their event logs [5].
2014	Gupta	Discipline that builds the bridge between data mining as an approach to Business Intelligence and Business Process Management and aims the discovery of models based on the event log available [9].
2014	Rendon-Mattogno	Relatively young research discipline that has emerged from data modeling in the business environment, in order to facilitate the work of extracting models [10].
2015	Aguirre Mayorga	Discipline which is composed by a number of tools and techniques based on data mining to analyze business processes which event log actual implementation are available in information systems [8].

case; and 6) Determine through the application of classification techniques if it is possible to assess the relationship between the different use cases.

Based on the six previously mentioned points, we can distinguish the conceptual difference between Data Mining and PM; while the first is used to find patterns in large amounts of data, or in a relationship of patterns and data, PM finds the relationship of the process inside the data, targeting information about the organization processes.

While reviewing the different reference process models (ISO/IEC 15504, ISO 12207, COMPETISOFT and ISO 29110), we found that they don't propose techniques, tools or algorithms to extract information and knowledge existing in their event logs or that analyze the evidence of the execution of activities in their database or repository of the organization knowledge. In this manner, we exposed the importance of implementing techniques or algorithms from PM which allow the adoption or application by the software developer organization to find out the actual status of their processes.

III. METHODOLOGY FOR SLR

We performed a Systematic Literature Review (SLR) which allows to identify, interpret, and synthetize all the existent investigation relevant in a certain topic [12]. The SLR methodology proposes three phases: 1) Planning the Review, 2) Conducting the Review and 3) Reporting the Review.

The stages associated with first phase are: Identification of the need for a review and Development of a review protocol. The stages listed above may appear to be sequential, but it is important to recognize that many of the stages involve iteration. In particular, many activities are initiated during the protocol development stage, and refined when the review proper takes place [12].

PM is relatively a new discipline, so the theoretical fundamentals are of great importance for researchers and practitioners to summarize all existing information about this discipline and SPI projects in a thorough and unbiased manner. The Development or review protocol step is documented below.

A. Research questions

This phase is focused on identifying problems in adoption of PM using SPI models. We formulated the following research questions: a) Are there any studies on the use of Process Mining using SPI projects? and b) What are the main issues of adopting Process Mining using SPI projects?

First question deals with identifying other studies about this subject, while the second question emphasizes identifying problems for adopting PM using a model for conducting the improvement process, a reference process model to follow and processes assessment methods. It consists in seeking for reference list about PM, its techniques and algorithms used in reports, study cases or areas which have benefited from their usage during 2005-2015 (search period). This search was done with the support from main information sources with a technical-scientific focus, under the same terms or search strings for Process Mining and software processes. Once the protocol has been agreed, the stages associated with second phase were performed.

B. Identification of Research and Data collection

Use of some important scientific research libraries was necessary, such as: IEEExplore, ACM DL, Scopus, ResearchGate and Springer Link; they help as primary sources of information. The goal was to review the most recent publications from the search period, aiming for specialized Journals, Conferences proceedings and the Internet [12]. We used a search structure, based on keyword and logic connectors, dealing with the following search string: `"Process Mining" and "Software Process Improvement"`.

The search string was adapted to each library and the search was restricted to title, abstract and keywords. In order to identify the most relevant studies for this SLR methodology, we used the following procedure: 1) Choosing the research keyword; 2) Looking in the digital library, trying with research keywords based on inclusion and exclusion criteria; 3) Analyzing each paper through title and abstract; 4) Downloading papers covering search criteria; 5) Reviewing introduction and conclusions conscientiously, and taking a fast view of the rest of the sections in a paper. This process was applied for each electronic library.

C. Study selection criteria

The largest number of assessments in CMMI-DEV v1.3 during 2015 corresponds to maturity level 3 (Defined), which might be the level that would require most effort taking into account the higher number of process areas associated being evaluated, regardless of the lower levels.

We considered those papers that included keywords such as: “reference process model”, “CMMI-DEV”, “SCAMPI”. Selected papers had to be published between 2005 and 2015 in Journals and Conferences, written in English language.

D. *Study Quality Assessment*

We formulated three questions that helped with relevance and accuracy in the papers contribution to our research goals.

QA1: Does the paper focus on solutions with software process reference models or SPI and Process Mining?

QA2: Does the paper focus on solutions with CMMI-DEV and Process Mining?

QA3: Does the paper expose application frameworks, methodology and technology (software resources) in Process Mining?

Each paper was evaluated using the same quality questions.

IV. Results From The Review

The search string was redefined for documents associated to PM with process improvement as follows: `("Process Mining" and ("Software Process Improvement" or "CMMI-DEV" or "SCAMPI"))`, obtaining as result 302 documents. Based on criteria of abstract and title, the number decreased to 68 documents, discarding those who did not consider the terms SPI, CMMI and SCAMPI; such as those documents that contained repeated information.

Table III shows the reference list of seven documents selected in the SLR [5, 8, 9, 17, 19, 20, 22], focusing on last contributions to the discipline of "Process Mining" and the solutions with "*Reference Process Models*" or "*Software Process Improvement*". The results were organized on four categories, focusing on the type of contribution made to the discipline: Theoretical Foundations and Algorithms, Proposals and Tools (Software Resources).

A. *Theoretical Foundations and Algorithms*

The document selection shows the importance of PM integrated to information system, database, evaluation processes, control and process monitoring, on those, diverse PM techniques (Discovery, Conformance Checking and Enhancement) and algorithms are proposed so it is necessary to know which algorithms are better for each process in particular. The most complete contributions were made by van der Aalst [5, 7, 13, 14, 15, 16], which were used as base of this work.

PM implements three basic algorithms: heuristic [17], fuzzy and genetic algorithms [10]. The extracted model by any of those three algorithms, could provide a view of the process(es) at a given point of the execution; this heads us to search diverse ways to analyze the process for the manipulations of the changes on it [17].

B. *Proposals*

According to Rubin, it is proposed a centered process in the environment of software engineering or software processes, where PM is the information needed as input in the software processes model [18]. It is pointed out that PM would ease facilitate the appraisal process type SCAMPI, giving the information needed about the process areas defined for a reference process model, such as CMMI Development v1.3.

The author Gruañas researched the potential of PM to support an improvement project and software process evaluation based on CMMI [19]. He proposed a specific methodology to find associations between PM and the evaluation of two process areas for CMMI maturity level 3.

On the other hand, the authors Valle *et al.* developed a framework (structure and content) to apply PM techniques in SCAMPI appraisals with the goal of carrying out a search with an improved deepness and coverage, while keeping the same effort level [20].

C. *Tools (Software Resource)*

In the last decade event data has become readily available and PM techniques in various academic and commercial systems [21]. Examples of commercial systems that support Process Mining are: ARIS Process Performance Manager by Software AG, Disco by Fyxicon, Enterprise Visualization Suite by Businesscape, Interstage BPME by Fujitsu, Process Discovery Focus by Iontas, Reflect one by Pallas Athena, and Reflect BY Futura Process Intelligence [26]. On the contrary, the open-source Process Mining tool ProM is widely used all over the globe and provides an easy starting point for practitioners, students and academics. ProM has been developed by the PM group at the Eindhoven University of

TABLE III
List of Selected Documents

Year	Database	Author(s)	Title	Category	Doc. Type
2007	ResearchGate	V. Rubin [18]	A Workflow Mining Approach for Deriving Software Process Models	Proposal	Thesis
2009	Springer	W. M. P. Van der Alst and B. F. Van Dongen [22]	ProM: The Process Mining Toolkit	Tools	Paper
2011	Springer	W. M.P. van der Aalst, *et al*. [5]	Process Mining Manifesto	Theoretical Foundations	Paper
2011	ResearchGate	J. Riera Gruañas [19]	Process Mining Opportunities for CMMI Assessments	Proposal	Thesis
2013	IEEExplore	J. Wang, R. K. Wong, J. Ding, Q. Guo, and Lijie Wen [17]	Efficient Selection of Process Mining Algorithms	Algorithms	Paper
2014	ResearchGate	E. P. Gupta [9]	Process Mining a comparative study	Algorithms	Paper
2015	ResearchGate	H. S. Aguirre Mayorga and Nicolás Rincón García [8]	Process mining: Development, applications and critical factors.	Theoretical Foundations and proposal	Paper
2015	Springer	A. Valle, E. Rocha Loures and E. Portela [20]	Process Mining Extension to SCAMPI	Proposal	Paper

Technology [22]. It integrates the functionality of several existing Process Mining tools and provides different PM plugins [15]. ProM contains more than 280 pluggable algorithms, developed to provide a wide range of functionalities and techniques. E.g., Petri nets, EPCs or Social Networks [16].

D. Implications for practice

PM techniques assume that it is possible to log events in a sequential manner, and by such way every single event refers to an activity. It is claimed that if every event (activity) could log within itself additional information of the process execution defined in a process reference model, then the event logs could be used to perform any of the three different types of PM techniques and/or algorithms through the application of data analysis tools, depending of the improvement cycle and/or the level of capacity which the organization aims to reach. E.g., in the event logs generated by process areas of CMMI, could obtain a process model to compare it with the required evidence in the SCAMPI assessment process, review the fulfillment of the established process areas for a specific process capability level and/or improve the process model using the following improvement cycle.

V. Conclusions

This paper presents a systematic literature review on interdisciplinary PM in SPI projects. A software developer organization faces problems associated with complexity in the management of a big amount of information or to the lack of experience in the technique or algorithms improvement team. This situation makes evident that deployment of formal techniques allows organizations to have more visibility in the execution of project or process, focusing in those which help to satisfy the business objectives making them more competitive.

The contributions of the review are 1) to identify the predominant contributions, the results were classified in three categories: Theoretical Foundations and Algorithms, Proposals and Tools; and 2) to visualize the incidence of processes mining in the event logs generated by organizational processes and support the process evaluation cycle or iteration in SPI project.

We considered that when making use and taking advantage of event logs in an improvement project based on CMMI-DEV, the appraisal method for process SCAMPI could be more efficient. Currently, the project is applying elements of the conceptual component of PM in a certified CMMI-DEV v1.3 with a maturity level 3 company, which wants to analyze event logs generated for one year in process areas, and thus obtain a model of the process that is compared with the evidence required in the SCAMPI assessment process.

References

[1] SPIRE Cases Studies. Software Process Improvement in Regions of Europe,. ESSI project, [Online]. Available: http://www.cse.dcu.ie/spire/main.html.

[2] B. L. Durón del Villar and M. A. Muñoz-Mata, "Selección de estrategias para la implementación de mejoras de procesos de Software", *ReCIBE,* vol. 2, nº 3, p. 15, 2013.

[3] A. Astorga-Vargas, J. Morales-Bustamante, B. Flores-Rios and J. Ibarra-Esquer. Determinación de la capacidad de procesos de software siguiendo el modelo de evaluación de procesos NMX-I-15504 aplicado en el modelo de referencia NMX-I-059 bajo la herramienta AURAP. Actas de la 9ª CISTI. España. 2014, pp. 283-288.

[4] A. Orellana-Garcia and D. Perez, "Generador de Registros de Eventos para el Análisis de procesos en el Sistema de Información Hospitalaria xavia HIS", 2015.

[5] W. M. P. Van der Aalst, A. Adriansyah, A. K. Alves de Medeiros et al, "Process Mining Manifesto", vol. 99, 2012.

[6] C. Pérez-López and D. Santín-González., "Data Mining", de *Data Mining, Soluciones con Enterprise Miner*, Mexico, Alfaomega, 2007, pp. 1-21.

[7] W. M. P. Van der Aalst, Netherlands, K. Hee, J. Werf and M. Verdonk, "Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor", *Computer,* vol. 43, nº 3, pp. 90-93, 2010.

[8] Aguirre, Mayorga, H. "Metodología para la aplicación de minería de procesos.", Pontificia universidad javeriana, *Tesis Doctoral,* Colombia, 2015.

[10] E. P. Gupta, "Process Mining a Comparative Study", *International Journal of Advanced Research in Computer and Communications Engineering,* vol. 3, nº 11, p. 5, 2014.

[11] M. Rendon-Mottgno, "Inteligencia Empresarial combinando técnicas de Minería de Procesos y Minería de Datos", *Thesis,* p. 48, 2014.

[12] B. Kitchenham. "Guidelines for performing Systematic Literature Reviews in Software Engineering", *Technical Repor*t, Software Engineering Group, Department of Computer Science, University of Durham, 2007.

[13] A.J.M.M. Weijters and W. M. P. Van der Aalst, "Process Mining: Discovering Workflow Models from Event-Based Data", p. 8, 2001.

[14] W. M. P. Van der Aalst, "Process Mining: Discovery, conformance and enhancement of business process", *Springer,* 2011.

[15] B. F. van Dongen, , A. de Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters and W. M. P. van der Aalst. "The ProM Framework: A New Era in Process Mining Tool Support." *APN*, 2005.

[16] I. Ailenei, "Process Mining Tools: A Comparative Analysis", *Eindhoven University of Technology. Masther Thesis*, *Eindhoven,* Netherlands, 2011.

[17] J. Wang, R. K. Wong, J. Ding, Q. Guo and L. Wen, *"Efficient Selection of Process Mining Algorithms",* IEEE Transactions on Services Computing, vol. 6, nº 4, p. 13, 2013.

[18] V. Rubin, "Process Mining Framework form Software Procesess". *Software Process Dynamics and Agility*, 2011, pp. 169-181.

[19] J. Gruañas, "Process Mining Opportunities for CMMI Assessments", *Tesis*, 2011, pp. 10-18.

[20] A. Valle, E. Rocha and E. Portela, *"Process Mining Extension to SCAMPI",* 4th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA), *Vol. 1293, 2011, pp. 179-183.*

[21] W. M. P. van der Aalst and B. F. van Dongen, "Discovering Petri Nets from Event Logs", *Springer,* ToPNoC VII, LNCS 7480, pp. 372–422, 2013.