Getting the Data in Shape for Your Process Mining Analysis: An In-Depth Analysis of the Pre-Analysis Stage

DOI: 10.1145/3712587
Handle: http://hdl.handle.net/1942/45193

Getting the Data in Shape for Your Process Mining Analysis: An In-Depth Analysis of the Pre-Analysis Stage

SHAMEER K. PRADHAN*, Faculty of Business Economics, Hasselt University, Diepenbeek, Belgium and Digital Future Lab, Hasselt University, Diepenbeek, Belgium
MIEKE JANS, Faculty of Business Economics, Hasselt University, Diepenbeek, Belgium, Digital Future Lab, Hasselt University, Diepenbeek, Belgium, and School of Business and Economics, Maastricht University, Maastricht, Netherlands
NIELS MARTIN, Faculty of Business Economics, Hasselt University, Diepenbeek, Belgium and Digital Future Lab, Hasselt University, Diepenbeek, Belgium

Process mining enables organizations to analyze the data stored in their information systems and derive insights regarding their business processes. However, raw data needs to be converted into a format that can be fed into process mining algorithms. Various pre-analysis activities can be performed on the raw data, such as imperfection removal or granularity level change. Although pre-analysis activities play a crucial role in process mining, there is currently a limited overview available regarding their scope and the extent of their examination. This study presents a systematic literature review of the pre-analysis activities in process mining projects. To better understand this stage and its current state of research, we explore which activities constitute the pre-analysis stage, their goals, the applied research methodologies, the proposed research outcomes, and the data used to evaluate the research outcomes. We identify 15 pre-analysis activities and concepts, e.g., data extraction, generation, and cleaning. We also discover that design science research is the methodology and methods that are the primary research outcome in previous studies. We also realize that the proposed outcomes have been evaluated using only real-life data most of the time. This study reveals that research on pre-analysis is a growing field of interest in process mining.

CCS Concepts: • **Applied computing → Business process monitoring**.

Additional Key Words and Phrases: process mining, process mining pre-analysis, data preprocessing, event log

1 Introduction

Information Technology (IT) resources are extensively utilized in organizations of various types and sizes. One significant application of the applications of information systems is to support business process management and enactment. Consequently, the increasing use of information systems has generated a large volume of business process data. Process mining is a set of techniques designed to enhance operational processes through the

*Corresponding author

Authors' Contact Information: Shameer K. Pradhan, Faculty of Business Economics, Hasselt University, Diepenbeek, Flanders, Belgium and Digital Future Lab, Hasselt University, Diepenbeek, Flanders, Belgium; e-mail: shameer.pradhan@uhasselt.be; Mieke Jans, Faculty of Business Economics, Hasselt University, Diepenbeek, Belgium and Digital Future Lab, Hasselt University, Diepenbeek, Flanders, Belgium and School of Business and Economics, Maastricht University, Maastricht, LI, Netherlands; e-mail: mieke.jans@uhasselt.be; Niels Martin, Faculty of Business Economics, Hasselt University, Diepenbeek, Belgium and Digital Future Lab, Hasselt University, Diepenbeek, Flanders, Belgium; e-mail: niels.martin@uhasselt.be.

ACM 1557-7341/2025/1-ART
https://doi.org/10.1145/3712587

systematic use of such process data [209]. It is a valuable tool to understand better and improve business processes by leveraging the process data.

The availability of high-quality data is a foremost requirement for a successful process mining project [204]. However, in most situations, the data stored in information systems is not in a form directly usable for process mining algorithms. The data must be extracted from the information systems and converted into an *event log*, which is a collection of events executed and stored in information systems that serve as the input for process mining algorithms [207].

Transforming data stored in information systems into event logs is a non-trivial task. An extensive set of tasks must be performed on process data before process mining analysis can commence. These tasks, that can be referred to as the "pre-analysis stage," are time-consuming, labor- and knowledge-intensive, as evidenced by their inclusion as a distinct stage in various process mining methodologies [24, 213]. This stage is regarded as one of the most critical in process mining, often accounting for more than 80% of the time and effort involved [4, 46]. Furthermore, a novel object-centric event data format is currently undergoing standardization [210], which may introduce additional challenges during the pre-analysis stage. It is evident that data preprocessing, a sub-stage within pre-analysis, can enhance the performance of discovered process models in terms of their fitness, precision, and F1 measure [67]. Despite the apparent benefits of pre-analysis activities on event data, a comprehensive overview of the pre-analysis stage of process mining, including its components and current research efforts, is lacking.

In this study, we conducted a systematic literature review (SLR) of the process mining pre-analysis stage to highlight the topics and extent of research efforts in this area. An in-depth exploration of the pre-analysis phase advances the field by providing researchers with a comprehensive overview of the state-of-the-art. Such a holistic perspective assists in identifying trends and gaps. Additionally, synthesizing knowledge across various aspects of the pre-analysis stage can foster innovation and new research directions that integrate findings from diverse areas. This overview can also serve as a resource for early-career researchers or those new to the field, providing a broad understanding of the current research landscape.

The primary contribution of this study is the categorization of topics investigated in process mining pre-analysis, which we have divided into four broad categories and 15 primary topics. We also identified four goals of process mining pre-analysis from the reviewed papers. The study also examined the research methodologies, types of research outcomes, and the data used to evaluate proposed solutions in these papers. Additionally, we analyzed the industries from which evaluation data was extracted. Understanding applied research methodologies is crucial as they provide specific direction to research and guide it throughout its life cycle [39, 167]. Finally, our study showcases the evolutionary trend of process mining pre-analysis research and specific topics within the field.

The remainder of this paper is structured as follows. Section 2 provides background information on the pre-analysis stage of process mining and positions this study in comparison with other existing systematic literature reviews. Section 3 describes the methodology used in this study. The results are presented in Section 4 and discussed in Section 5, which also offers directions for future research. Section 6 concludes the paper.

2 Background

The pre-analysis stage of process mining is a crucial phase in the success of any process mining project. Research on various aspects of the pre-analysis stage of process mining has gained more attraction in recent years. For example, research on data extraction [19, 187], assessment and fixing of quality issues [74, 133, 225], and privacy of event data [9, 56, 156] have been conducted. Furthermore, many papers related to the pre-analysis stage of process mining have been presented in the annual *International Conference in Process Mining*[1].

[1]https://icpmconference.org/

Previous studies have explored the subject of the pre-analysis stage of process mining and associated concepts such as data preprocessing to some extent. We have identified ten SLRs related to our study. One of the SLRs explores the preprocessing of event logs in general and classifies the reviewed literature based on various research topics [131]; however, the dimensions and scope of our review differ from their review. The rest of the SLRs can be categorized into three different settings.

Firstly, some reviews employ an SLR to explore state-of-the-art research and immediately integrate the insights gained into the development of an artifact. For example, Zandkarimi et al. [237] identified *common concepts and building blocks* of trace clustering techniques used to develop a trace clustering framework for process mining. Similarly, Koschmider et al. [111] conducted an SLR to develop a framework for incorporating context information while mapping log events and process model activities. Another study employed an SLR to identify various approaches to querying process data in order to devise an improved process data query language [81]. A taxonomy of the event abstraction methods was devised based on the review of the preprocessing techniques that enable the abstraction of event data into the right granularity level [218]. Finally, Stein Dani et al. [187] reviewed the role of humans in data extraction and developed a taxonomy of manual tasks performed during such extraction.

Secondly, some SLRs include event logs as part of a broader review scope. However, unlike our study, those SLRs focus on data preprocessing as a subpart of their primary objective, such as process prediction. One study explores, among other things, the types of preprocessing steps and data encoding in the domain of deep learning-based process prediction [141]. Another SLR focuses on process mining in the manufacturing industry [186]. They review the literature on data extraction, generation, and preprocessing.

Thirdly, some SLRs relate to preprocessing process data as their primary objective but zoom in on one specific aspect. One study is dedicated to reviewing the literature on extracting data from information systems [41]. Similarly, another SLR reviews the literature regarding automated event log abstraction techniques, focusing on the quality aspects and accuracy of the proposed techniques [53].

Table 1 presents an overview of the earlier studies and their characteristics. The table acts as a comparison between those studies and our study. While the previous SLRs have focused on specific aspects of the pre-analysis stage of process mining, we have taken a broader and more detailed approach by exploring the entire pre-analysis phase. Additionally, we have analyzed literature using dimensions that have not been considered before, i.e., the goals, the applied research methodologies, research outcomes, and evaluation data. Such information can enrich the study of the pre-analysis stage and guide it throughout its life cycle [167].

Table 1. Overview of existing SLRs and their comparison with this review.
(Legend: ● - Included as a review dimension. ◐ - Mentioned but not included as a review dimension. ○ - Not mentioned in the review.)

	González López de Murillas et al. (2017) [81]	Koschmider et al. (2019) [111]	van Zelst et al. (2020) [218]	Dakic et al. (2020) [41]	El-Mastri et al. (2020) [53]	Stefanovic et al. (2020) [186]	Zandkarimi et al. (2020) [237]	Neu et al. (2021) [141]	Marin-Castro and Tello-Leal (2021) [131]	Stein Dani et al. (2022) [187]	**This paper**
Total number of papers reviewed	25	16	28	15	89	14	103	32	70	46	260
Scope	Querying	Event-activity mapping	Abstraction	Extraction	Automated log abstraction	Preprocessing for an application	Trace clustering	Preprocessing for process prediction	Preprocessing	Extraction	**Pre-Analysis Stage of Process Mining**
Reporting period	1994-2016	Not reported	2009-2020	>2008	2000-2018	Not reported	2004-2020	Not reported	2006-2020	2000-2020	>2022
Review dimensions											
Research topic	●	●	●	◐	●	●	◐	●	●	◐	●
Goals	○	○	○	○	○	○	○	○	○	○	●
Research methodology	○	○	○	○	○	○	○	○	○	○	●
Research outcome	◐	○	○	○	○	○	○	○	○	○	●
Type of data used for evaluation of the proposed solution	○	○	●	○	○	○	○	○	○	○	●
Type of industry of the evaluation data	○	○	○	○	○	○	○	○	○	○	●
Number of data set used in evaluation	○	○	○	○	○	○	○	○	○	○	●

3 Methodology

We conducted an SLR using the method proposed by Xiao and Watson [230] to explore the topics investigated in the pre-analysis stage of process mining. The method begins by understanding the problem and developing a review protocol, which we supported by formulating the following research questions.

3.1 Research Questions

To guide our objective of exploring existing research dealing with the pre-analysis stage of process mining and developing an overview of investigation topics in the field, we formulated the following research questions:

- **RQ 1:** Which research topics currently relate to the pre-analysis stage of process mining?
- **RQ 2:** Which research methodologies are employed to study the pre-analysis stage of process mining?
- **RQ 3:** What kind of research outcomes have been proposed by studies conducted in the pre-analysis stage of process mining?
- **RQ 4:** Which type of data is used to evaluate the artifacts proposed by the studies conducted in the pre-analysis stage of process mining?
- **RQ 5:** How has the research field of the pre-analysis stage of process mining evolved in the last 20 years?

3.2 Initial Literature Search

We initiated our investigation by sourcing papers from Web of Science, IEEE Xplore, and Scopus databases. The review process encompassed papers published until April 2022 without imposing any specific start date for the literature search.

To identify studies pertinent to the pre-analysis stage of process mining, we employed the search query: *("process mining" OR "process discovery") AND "event*" AND "log*").* We used the logical *OR* between "*process mining*" and "*process discovery*" since some studies used the latter terminology instead of the former. Additionally, we utilized the asterisk (*) symbol as a wildcard to represent different variations of the terms, such as *event logs* and *events are logged*, ensuring comprehensive coverage.

In IEEE Xplore, we conducted searches across *All Metadata* without restricting by publication year. For Scopus, we searched in *Article title, Abstract, Keywords* across all available years. In Web of Science, searches were conducted across *All fields* within the *Web of Science Core Collection*.

Figure 1 provides an overview of the search strategy and the number of papers included and excluded at each stage. The initial search across the three databases yielded 3,923 results, comprising 1,286 papers from Web of Science, 513 from IEEE Xplore, and 2,124 from Scopus. Upon removing duplicates, we screened 2,255 papers for eligibility.

3.3 Screening

A paper was excluded when it met one of the following exclusion criteria:

(1) The paper's primary focus is not on the pre-analysis stage of process mining.
(2) The paper is written in a language other than English.
(3) The paper is not published in a peer-reviewed scientific journal or peer-reviewed conference proceedings.
(4) The paper is not electronically available.
(5) The paper is a conference paper with an extension published as a journal paper.
(6) The paper is a poster, one-pager, executive summary, abstract, position paper, literature review, or commentary.

The screening process was conducted in three phases. Initially, papers were screened based on their title alone, followed by screening based on both title and abstract, and finally, full-text assessment. If there was insufficient information to decide the inclusion or exclusion of a paper, the paper advanced to the subsequent phase. During

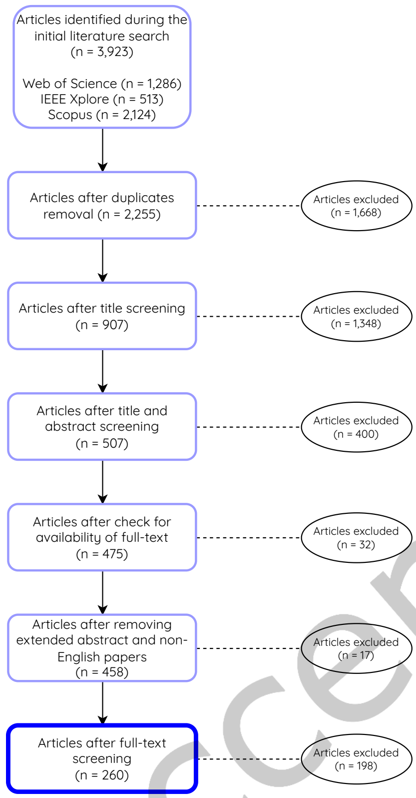

Fig. 1. Paper selection strategy

the first phase, 1,348 papers were excluded outright, with 907 papers proceeding to the next phase. Screening based on title and abstract further excluded 400 papers, resulting in 507 papers being considered for full-text review. Among these, 32 papers were inaccessible in full text, 12 contained only extended abstracts, and five were written in languages other than English. The remaining 458 papers underwent full-text screening, resulting in the exclusion of 198 papers. Ultimately, 260 papers were included in this literature review.

3.4 Quality Assurance

The first author screened and coded all papers. Additionally, a random subset of papers was individually screened by the two other co-authors. Fifty papers were individually screened on title. Each researcher independently assessed whether to include or exclude each of these papers. In instances where there were conflicting viewpoints regarding the inclusion or exclusion of a paper, the issue was thoroughly discussed. When consensus among the researchers was not reached, decisions were subsequently made based on a majority vote. Furthermore, inter-rater reliability among the researchers was evaluated using Fleiss' kappa [177] and Krippendorff's alpha [113]. Fleiss' kappa for the title screening phase was 0.232, indicating "fair agreement" [114] among the researchers. Krippendorff's alpha was recorded at 0.238, with positive values suggesting agreement and a value of 1 denoting perfect agreement [113].

In the subsequent phase of the screening process, which involved evaluating both the title and abstract, thirty papers were individually reviewed to test for reliability. During this phase, Fleiss' kappa improved to 0.453, signifying "moderate agreement," while Krippendorff's alpha increased to 0.460, indicating a higher level of agreement compared to the first phase.

Notably, the primary researcher's cautious screening approach may have impacted Fleiss' kappa and Krippendorff's alpha values, potentially reducing inter-rater reliability. For instance, in the first phase, the primary researcher included 10 out of 50 randomly selected papers that the other researchers had excluded. Had the primary researcher agreed with the others and excluded these papers, Fleiss' kappa and Krippendorff's alpha would have increased to 0.488 and 0.492, respectively. Similarly, in a hypothetical scenario for the second phase, alignment with other researchers would have raised the values to 0.665 and 0.669 for Fleiss' kappa (indicating a "substantial agreement") and Krippendorff's alpha, respectively. However, the primary researcher's cautious approach aimed to retain more papers for further screening, prioritizing inclusivity over the potential exclusion of relevant papers early on, which is generally considered advantageous.

3.5 Coding Scheme

We employed a coding scheme that utilized both inductive and deductive coding techniques [38, 71] to extract relevant information from the literature. The inductive coding technique assigns codes based on the terms and phrases mentioned in the paper, allowing us to extract information regarding research topics and goals [134]. We synthesized and refined these codes over several iterations and multiple coding cycles, as Giola et al. [77] and Skjott Linneberg and Korsgaard [182] suggested. The deductive coding technique uses predefined codes to extract information concerning the research methodology, research outcome, and evaluation data. The predefined set of codes and their descriptions can be found in Section 1 of a separate supplementary document[2].

We utilized a spreadsheet to track the codes and synthesize them into high-level categories systematically. Each researcher actively participated in developing the coding scheme and contributed to its refinement. When there was uncertainty regarding the appropriate codes for a particular dimension of a specific paper, collaborative brainstorming sessions were conducted among the researchers to reach a consensus on the most suitable codes. Additionally, it was determined that a single paper could be assigned multiple codes within any given dimension. This is because a paper might address multiple topics, employ various research methodologies, produce different types of research outcomes, or use diverse types of evaluation data from a variety of sources. The following dimensions were included as part of the coding scheme developed for this study.

Research topic. We employed inductive coding to identify various topics related to the pre-analysis phase of process mining, addressing research question 1. Following the Gioia method [37], we first extracted the general subject of investigation from each paper as a first-order conceptualization. We then combined similar concepts into topical themes, which were further aggregated into a final collection of study topics. The identification of these research topics was conducted using a bottom-up approach, allowing the topics to naturally emerge from the content of the reviewed papers.

Goal: To determine the objectives of the papers, we employed inductive coding to identify the goals of the reviewed papers. This dimension also supports research question 1.

Research methodology. To code the applied research methodology of the reviewed papers, we used a predefined set of codes suggested by Recker and Mendling [168]. The codes include *action research*, *case study*, *controlled experiment*, *design science / engineering*, *field experiment*, *formal proof*, *illustration*, *interviews*, *simulation*, and *survey*. This dimension supports research question 2.

Research outcome. We examined the research outcome of the reviewed papers, using the types of research outcomes listed by Recker [167] as a starting point. The set of codes for answering research question 3 includes

[2]https://doi.org/10.5281/zenodo.14623497

constructs, *design theories*, *instantiations*, *methods*, and *models*. March and Smith [130] delineates four primary research outcomes: constructs, models, methods, and instantiations. A construct is conceptualized as a vocabulary of a domain used to describe problems within that domain and to delineate their solutions. Constructs serve as the foundation for defining key terms within that domain. In contrast, a model is defined as a set of propositions that articulate the relationships among constructs. Models can represent various aspects, such as problems and their solutions, entities and their relationships, or even data structures. A method is a series of steps designed to perform a particular task, grounded in the underlying constructs and models. Finally, an instantiation refers to the implementation of an artifact within an environment, thereby operationalizing constructs, models, and methods [130]. A design theory, on the other hand, refers to a systematic approach to investigating, understanding, and improving the processes, principles, and outcomes of design [167].

Evaluation – data type. We coded the type of data used to evaluate the proposed methods as either *artificial data* or *real-life data*, supporting research question 4. If the information regarding the type of evaluation data was missing, the paper was coded as *not reported*. Additionally, a paper did not need to evaluate the proposed method using data to be included in our study. Consequently, the distinction between *real-life evaluation data*, *artificial evaluation data*, and *not reported* was only made if the paper included a data-based evaluated. Therefore, the number of papers coded in this dimension can be lower than the total number of included papers.

Evaluation – data industry. To further investigate which type of data is used to evaluate research contributions to process mining pre-analysis, we recorded the industry from which the data originates in case of real-life data evaluations, using a predetermined set of codes derived from the *Statistical Classification of Economic Activities in the European Community* [62]. If the industry information was missing, the paper was coded as *not reported*. A paper could evaluate the proposed solution using data from multiple industries, resulting in the paper being coded for multiple industries. Hence, the number of papers coded for this dimension may differ from the total number included in the review.

Evaluation – number of data sets. For the papers utilizing real-life data, we tracked the number of data sets used during the evaluation phase, categorizing them as either *single data set* or *multiple data sets*. Papers coded as "single data set" used one data set for evaluation, while those coded as "multiple data sets" used more than one.

The coding scheme, topics, and subtopics were developed through a series of iterative brainstorming sessions among the researchers.

4 Results

This section outlines the results of our SLR. We have categorized our findings using the Giola method [37], which organizes data into lower-level first-order concepts, groups them into higher-level second-order themes, and combines them into aggregate dimensions. Using this method, we identified four broad categories: *Getting Data*, *Subsetting Data*, *Data Quality*, and *Privacy*, encompassing the various topics discussed in the papers. Moreover, the research topics have been further classified into subtopics. We have described our findings in more detail in the following subsections.

While synthesizing the research topics, we identified common objectives underlying the studies, leading us to conclude that outlining these objectives is also pertinent to the study of research topics. Including the goals provides a more comprehensive understanding of the research landscape, highlighting the broader intentions and motivations driving the studies. Therefore, in Section 4.1, we first provide our findings regarding the goals of the reviewed studies. The subsequent subsections are organized according to the research questions outlined in Section 3.1. Specifically, our analysis covers the research topics (Section 4.2), research methodologies (Section 4.3), research outcomes (Section 4.4), and types of evaluation data (Section 4.5), corresponding to research questions 1, 2, 3, and 4, respectively. Furthermore, Section 4.6 details our observations on the evolution of the research field pertaining to the pre-analysis stage of process mining over the past two decades.

We also developed an interactive dashboard, containing a list of all the papers in this study. This dashboard allows viewers to interactively search, sort, and filter the data according to their requirements. The dashboard can be accessed from our online repository[3]. Similarly, a list of the reviewed papers are provided in Section 2 of a separate supplementary document[4].

4.1 Goals

During our examination of the literature, we identified four principal objectives for the pre-analysis phase in process mining: *enabling process mining* (157 papers), *enabling privacy-aware process mining* (19 papers), *enabling quality-aware process mining* (78 papers), and *enabling semantics-aware process mining* (6 papers). *Enabling process mining* is a broad goal applicable across all subtopics of pre-analysis, whereas the remaining three objectives are more specialized.

Privacy, defined as the right of individuals to protect personal information from unauthorized access, dissemination, and use by others [171], is critical when event data contains sensitive personal details. When event data contains sensitive personal information, privacy-aware process mining becomes necessary to safeguard the privacy of individuals whose data is present within the event logs [129]. Therefore, privacy-aware process mining has emerged as a research goal aimed at enhancing the privacy of individuals whose personal data is present in event data.

Ensuring the quality of event data is essential for obtaining accurate process mining outcomes [22]. Techniques to detect and enhance data quality have been extensively explored in literature, emphasizing the importance of *quality-aware process mining*.

Researchers have also concentrated on understanding the semantic meaning of data recorded in event logs. Semantic process mining endeavors to interpret the conceptual underpinnings of textual labels within the logs [44], underscoring *semantics-aware process mining* as a significant pre-analysis objective.

Figure 2 illustrates the distribution of papers across these distinct goals, highlighting the field's focal points.

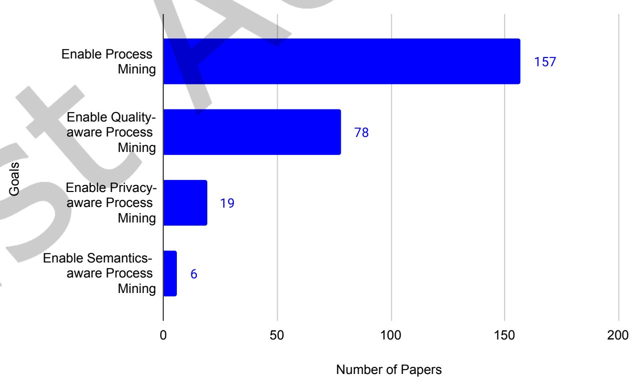

Fig. 2. Goals of the pre-analysis stage of process mining

[3]https://shapradhan.shinyapps.io/pre-analysis_dashboard/
[4]https://doi.org/10.5281/zenodo.13790927

4.2 Research Topics

This section presents the research topics identified during the literature review, which are grouped into four main categories - *Getting Data*, *Subsetting Data*, *Data Quality*, and *Privacy*. This section addresses the first research question, which pertains to identifying the research topics of interest during the pre-analysis stage of process mining. The following subsections describe the topics classified under these categories. The topics and subtopics are graphically visualized in Section 3 of a separate supplementary document [5].

As observed in Figure 3 topics related to getting data, such as *event log-building* and *data generation*, have been extensively researched. Moreover, there has been significant research on data quality-related topics such as *log cleaning*, *log assessment*, and *log enrichment*, which are located in the upper half of the chart in Figure 3.

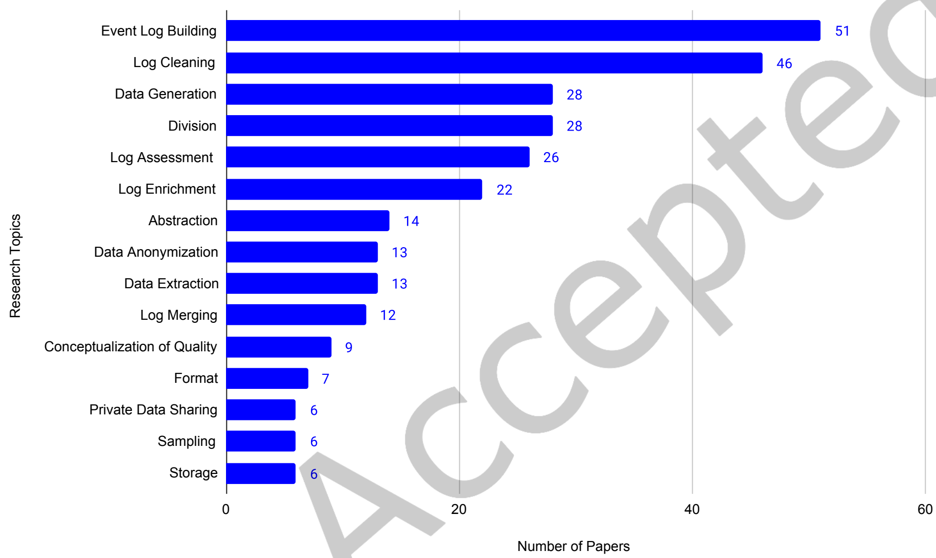

Fig. 3. Research topics in the field of pre-analysis stage of process mining

4.2.1 Data Extraction (Getting Data). Of the reviewed papers, 13 (5%) focus on *data extraction*, which involves obtaining data from information systems. The data extraction process may include manual and (semi-)automated steps [187]. In this study, we identified only one paper that studies the manual human intervention required during data extraction, while the rest propose (semi-) automatic approaches to extract data. Stein Dani et al. [187] developed a taxonomy of the manual tasks necessary to understand the human role in data extraction. The authors argue that comprehending the human role in data extraction is critical for developing automated techniques. They identified five primary categories of tasks requiring human intervention: defining the context and scope, assessing the data source, selecting attributes, extracting the data source, and assessing the event log.

Numerous (semi-)automatic techniques for data extraction have also been proposed. The contributions of these papers primarily focus on extracting data from information systems that store data in various media, such as relational databases [145, 180, 223], graph databases [61], redo logs [80], and unstructured process-related documents [119].

[5]https://doi.org/10.5281/zenodo.13790927

In addition to extracting raw data from various sources, there is also a need to query the event data stored in databases. Two papers focus on querying event data by developing query languages to extract data from databases while considering different aspects of data [81, 176].

4.2.2 Data Generation (Getting Data). Twenty-eight (10.8%) papers are dedicated to *data generation*, which involves producing raw data for process mining. This topic can be categorized into two types: generating artificial event logs and collecting real-life data.

Fourteen (5.4%) papers focus on using process simulation to generate artificial event logs. For instance, event logs have been generated by simulating Petri net models [60, 139], BPMN models [135], DECLARE models [49, 183], process tree models [99, 100], and business specifications [2]. Additionally, the generation of *negative events* – events that did not occur in real life but could have – has been investigated [78, 124, 219]. These *negative events* can be added to existing event logs to make them more comprehensive and reflect all potential events that could have happened [79]. Furthermore, the artificial generation of anomalous event data has also been explored [108].

Collecting real-life event data involves capturing and gathering data from two primary sources: information systems and non-information systems, such as blockchain. Fourteen papers (5.4%) delve into this subtopic. Techniques for data collection by analyzing the legacy code of information systems have been developed [94, 140, 151, 153–155], where legacy code can be analyzed statically, dynamically, or in a hybrid manner and annotated with business process information, which can then be executed to generate an event log files. Additionally, data can be captured and collected from other sources such as blockchain [14, 136], Internet of Things (IoT) [178], and unplugged processes [93].

4.2.3 Event Log-Building (Getting Data). Event log-building involves creating event logs from raw data. Among the papers reviewed, 50 (19.2%) address this topic. Research on event log-building primarily focuses on the general conceptualization and techniques of building event log from raw data. These techniques can be subdivided into those that extract logs from data stored in databases of information systems and those that use data stored in non-database systems, such as blockchain.

Concerning the general conceptualization of building an event log, a series of steps have been devised to support process analysts in effectively building event logs [98]. These steps aim to enhance the understanding of decisions made by analysts regarding process instances, activities, and attributes as well as the consequences of those decisions. Another study presents a set of options for building event logs in the auditing domain, focusing on process instances, activities, and attributes [97].

A majority of papers discuss event log-building from data stored in databases. For example, an ontology-based technique, in which a high-level representation of the domain of interest is connected to the data sources, has been used to extract and build event logs from relational databases [27–30]. Similarly, techniques have been proposed to build event logs using data from domain-specific information systems such as health information systems [75, 84, 85, 169, 220] and enterprise resource planning systems [19, 92, 137, 146, 147].

Research on building event logs from non-database systems includes sources such as blockchain [136, 138], CSV files [52, 101], electronic data interchange (EDI) messages [58, 59], Internet of Things (IoT) devices [15, 16, 25, 106, 178, 179, 199, 214], and unstructured text data such as emails and natural language conversations [7, 57, 104, 150]. For instance, a natural language processing technique proposed by Kecht et al. [104] constructs an event log from customer service conversations on social media sites such as Twitter.

4.2.4 Storage (Getting Data). Six (2.3%) of the reviewed papers concentrate on *storing* event logs, exploring various storage media such as relational databases [176, 194, 212], graph-based databases [61, 96], and a 3D process-aware datacube [82]. Graph-based and relational databases facilitate the preprocessing of intermediate structures, like the *Directly Follows Graph*, within the database engine. This approach eliminates the need to transfer data to memory and reload and recompute previous data [96, 194]. Such computation reduces the processing time as they are executed during data insertion rather than analysis [194]. Furthermore, storing event logs in a relational database allows for including events in multiple traces and supports random access [212]. Another paper presents a solution aimed at optimizing event data performance and storage by recommending an appropriate database configuration [176].

4.2.5 Format (Getting Data). Seven (2.7%) of the reviewed papers focus on the *format* of event logs, identifying four main formats: *Mining eXtensible Markup Language (MXML)* [51], *eXtensible Event Stream (XES)* [223], *eXtensible Object-Centric (XOC)* [117], and *Object-Centric Event Log (OCEL)* [76].

MXML, one of the earliest event log formats, was developed to address the lack of a standard event log format for process mining [51]. An extension known as *Semantically Annotated MXML* supports semantic process mining [48]. However, MXML format has notable limitations, such as its lack of support for extensive semantic information of the additional attributes and a rigid naming schema for various concepts. In order to alleviate these weaknesses, the XES format was developed [223]. The XES format provides flexibility in the event data structure through attribute extensions covering concept, life cycle, organizational, time, and semantic perspectives. Furthermore, XES allows the identification and comparison of individual events via event classifiers [223].

XOC format was designed to mitigate the shortcomings of XES, such as the mandatory need to identify case IDs for processes and the problem of data convergence (where a single event is associated with multiple cases [208]) and divergence (where a single case may have multiple independent, repeated executions of a group of activities [208]) [117]. XOC correlates events based on objects rather than cases. Similarly, the OCEL format addresses convergence and divergence issues by focusing on object-centric event correlations.

4.2.6 Log Merging (Getting Data). Twelve (4.6%) papers explore *merging of event logs* from different sources. Techniques have been developed to create event logs using data from multiple systems and organizations, in both inter-organizational [35, 86, 232] and intra-organizational [48, 163, 232] contexts.

During data merging, it is essential to correlate the events belonging to the same process instance. Various correlation techniques have been developed utilizing different types of event data, such as object-centric event data [118], unlabeled process event data [83], data from non-process-aware information systems [152], and data from supplementary sources [127].

4.2.7 Sampling (Subsetting Data). Six (2.3%) of the reviewed papers focus on the topic of *sampling* data. Sampling an event log generates a subset of data that can be analyzed while preserving the characteristics of the entire dataset [3]. In process mining, sampling can extract a subset of events from the entire event log, thereby enhancing the computational performance of analysis algorithms [67].

A method has been proposed to sample event data using a graph-based ranking model to evaluate the quality of the sample log [121, 122]. This approach ranks the data based on similarity and selects the most representative sample. Similarly, Sun et al. [189] presented a method to derive high-quality samples by considering various quality factors, such as risk level and risk diversity. Additionally, an incremental sampling technique for event data in process discovery has been developed, allowing data preprocessing and subsequent steps to be performed with partial data [11]. Moreover, specific sampling techniques have been proposed to facilitate applications such as determining trace relevance for conformance checking [103] and predictive monitoring [70].

4.2.8 Division (Subsetting Data). Twenty-eight (10.8%) of the reviewed papers address the *division* of data as a strategy to mitigate the complexity of event logs. Analyzing process data with finer granularity can result in complex and incomprehensible "spaghetti" process models [205] particularly for highly variable processes. Two primary methods of division are identified: trace clustering and decomposition.

Trace clustering involves grouping traces in a heterogeneous event log into homogeneous clusters that a process model can adequately represent [23, 63]. Process models are subsequently discovered for each cluster. In contrast, decomposition entails breaking down an event log into smaller sub-logs with process models being discovered for each sub-logs, which are then merged into a comprehensive process model [221].

Several methods have been proposed for clustering traces, including alignment [20, 63, 228] and similarity-based clustering [110, 144, 195]. Similarly, clustering techniques have been developed based on case attribute values [87, 216] and distance measures [5, 21, 32]. Moreover, various trace clustering approaches employ different clustering strategies and focus on specific evaluation schemas. For instance, the technique proposed by Sun and Bauer [190] emphasizes the accuracy and complexity of the sub-process models, while the one proposed by De Koninck and De Weerdt [42] focuses on model fitness. Furthermore, a technique to detect changes in event logs using clustering to divide the log has also been proposed [112].

Several papers propose techniques that utilize a divide and conquer approach to decompose event logs [31, 206, 221, 222]. Other decomposition methods, such as the one proposed by Nguyen et al. [142], decompose the data into stages that individual discovery of process models. Similar techniques have also been proposed for the decomposition of unlabeled event logs [227].

4.2.9 Conceptualization of Quality (Data Quality). Among the reviewed papers, nine (3.5%) focus on exploring the *concept of quality* in the context of event logs. Three of these papers delve into the definitions and estimations of completeness of event logs, offering novel definitions or characterizations of incompleteness [34, 116, 148]. Several papers investigate conceptual aspects, such as different types, characteristics, and measures of quality [4, 22, 192, 200]. Additionally, the tendencies of data, including typical characteristics of missing values in event logs, are examined, as demonstrated by the study conducted by Horita et al. [89]. Furthermore, Martin [132] presents a domain-specific study on the quality of event data within the healthcare sector.

4.2.10 Log Assessment (Data Quality). Twenty-six (10%) of the reviewed papers concentrate on *assessing* event logs, a process essential for identifying and addressing quality-related issues. High-quality event logs are imperative for accurate process mining results, making the detection of quality issues is a crucial task [203]. The papers address four primary quality concerns: completeness, complexity, noise, and outliers, along with the overall quality of the logs.

To accurately reflect actual business processes, event logs must be complete [204]. Many organizations manually verify event logs to ensure they meet the required standards. However, various techniques have been proposed to evaluate and ensure completeness [6, 215, 233, 235]. For instance, methods have been introduced to assess the local completeness of an event log [233, 235] and probabilistic approaches have been employed to evaluate overall completeness [6, 215].

A subset of papers focuses on the complexity of event logs. Metrics for evaluating complexity include methods to compute trace diversity within event logs [17]. Ramos-Gutiérrez et al. [164], for example, have presented a method and a set of metrics for assessing and ranking possible traces from relational databases, which can help to reduce computational costs.

Several techniques have been developed to assess the noise in event data, including methods to identify and measure labeling issues, such as synonymous and polluted labels [165, 173, 174]. Additionally, timestamp-related imperfections, such as inaccurate or incomplete timestamps, have been identified and addressed [73, 74].

Techniques for dealing with outliers – data points that deviate significantly from other data points [8] – have also been proposed. These techniques detect anomalies related to event data [102, 202, 226, 234] and infrequent behaviors [217].

Several tools and frameworks have been developed for log quality assessment. For example, the *DaQAPO* tool [133] enables a flexible and fine-grained assessment of event log quality. The technique proposed by Knols and van der Werf [107] supports measuring the behavioral quality of a sample of an event log, allowing for the detection of optimally sampled or oversampled behaviors.

4.2.11 Abstraction (Data Quality). Fourteen (5.4%) of the reviewed papers concentrate on *abstraction*, which involves transforming event data from a low-level granularity to a higher level. In real-life processes, events are often recorded in a highly detailed manner, leading to the generation of the so-called spaghetti models [95, 218]. The papers within this subtopic can be classified into two categories: those that devise abstraction techniques and those that propose abstraction frameworks.

Clustering-based techniques have been developed to abstract lower-level events to higher-level activities [43, 191]. By using these techniques, fine-grained events are simplified into more manageable and interpretable higher-level activities. Furthermore, sector-specific abstraction techniques, employing clustering or other methods like trace comparison, are also prevalent. For instance, Leonardi et al. [115] have developed abstraction techniques to medical processes.

Furthermore, comprehensive abstraction frameworks have also been proposed [40, 111, 198], extending beyond mere abstraction techniques. Papers addressing these frameworks often discuss related concepts such as log segmentation, event classification, and activity labeling, providing a more holistic approach to event data abstraction.

4.2.12 Log Cleaning (Data Quality). Forty-six (17.7%) of the reviewed papers focus on *log cleaning*, which involves rectifying imperfections in event logs. The papers on log cleaning can be categorized into several themes: event order imperfection repair, filtering, noise repair, and outlier handling.

Various methods have been developed to address imperfection in event order. For instance, van der Aa et al. [201] proposed using partial events to estimate the probability of the overall event order, while Dixit et al. [50] developed timestamp-based indicators to correct these imperfections. Additionally, Xiangsheng [229] proposed a framework for managing out-of-order RFID event streams.

Filtering out infrequent and erratic behavior enhances the quality of the discovered process model [36]. To this end, several filtration techniques have been developed [18, 36, 69, 88, 90, 91, 126, 188, 197, 217, 224, 238]. For example, Tax et al. [197] introduced a method that filters activities based on their entropy rather than their frequency, while Vidgog et al. [224] proposed an approach that uses a multi-range filter to extract insights from infrequent data.

Several techniques have also been proposed to address noise-related imperfections, which includes issues like incorrect data (e.g., mislabeled data) and incomplete data [175, 211]. The literature presents methods for repairing missing values [22, 47, 123, 181, 184, 231, 236], resolving labeling issues [33, 143, 173], and correcting timestamp discrepancies [22, 143, 170]. Furthermore, some papers focus on handling outliers, such as anomalous or nonconforming behavior, including trace deviation [68, 109, 185].

4.2.13 Log Enrichment (Data Quality). Twenty-two (8.5%) of the reviewed papers address *log enrichment*, which involves integrating additional information into event logs to facilitate more detailed analysis. Various techniques for refining activity labels in event logs have been proposed [1, 7, 26, 45, 125, 126, 165, 197]. For instance, Tax et al. [196] proposed a method that leverages event timestamps to enhance event labels, while De Oliveira et al. [45] developed a technique for analyzing complex events and generating meaningful labels for them. Additionally, the literature includes methods for deducing case IDs from unlabeled event logs to enable the event correlation [12, 13, 72, 83, 120]. Moreover, approaches to enriching event logs with semantic information have been developed, allowing for adding semantic meaning to the data contained within the logs [44, 105, 166].

4.2.14 Data Anonymization (Privacy). Fourteen (5.4%) of the reviewed papers focus on *data anonymization*, presenting various techniques developed for both general and sector-specific applications. For instance, a privacy-preserving process mining model has been created to anonymize data in the healthcare sector, where protecting patients' medical data is paramount [149, 172]. Similarly, Rafiei et al. [161, 162] proposed methods for anonymizing personally-identifiable information to maintain confidentiality across broader applications. These techniques aim to prevent unauthorized access to confidential information while ensuring that the anonymized event logs produce results comparable to the original data.

Several papers have proposed *differential privacy*, which involves adding noise to the data to reduce the likelihood of re-identifying individual in an event log [56, 64, 128]. Moreover, privacy-preserving techniques that protect against specific types of attacks have been proposed. For instance, Batista and Solanas [9] developed methods to guard against location-oriented targeted attacks, while Rafiei and van der Aalst [157] focused on defenses against frequency-based attacks. Furthermore, event log sanitization methods have also been introduced to enhance data privacy [10, 65].

4.2.15 Private Data Sharing (Privacy). Six (2.3%) of the reviewed papers address the topic of *private data sharing* in the context of process mining. One key privacy-related issue in this field is the sharing event data between organizations for analytical purposes. Although integrating event data from multiple organizations can enhance analysis, it raises significant privacy concern. In particular, the study by Rafiei and van der Aalst [159] delved into a specific type of privacy threat known as *correspondence attacks*, where an attacker attempts to identify personal information by cross-referencing published event logs. Various techniques for sharing private data without exposing sensitive information have been proposed to mitigate such risks [66, 156, 158]. Rafiei and van der Aalst [160] contributed to this effort by proposing methods to quantify data disclosure risk and enhance data privacy. Additionally, Elkoumy et al. [54, 55] introduced an architecture called *secure multi-party computation* that enables sharing of computations among multiple parties while preserving the confidentiality of the original data. It employs secret sharing algorithms to partition the data into several segments, which are then uploaded by the respective parties to compute nodes. The event logs are encrypted, ensuring that private information remains accessible only to the log owner.

4.3 Research Methodologies

In this section, we address the second research question, which focuses on the types of research methodologies applied in the study of the pre-analysis stage of process mining. The majority (94.2%) of the studies employed the *design science* approach, with 245 papers utilizing this methodology. Notably, some papers incorporated multiple research methodologies, leading to a higher total number of papers in Figure 4 compared to the number included in the full-text review. The *case study* and *illustration* methodologies were used by 62 (23.8%) and 50 (19.2%) papers, respectively. All methodologies except for *formal proof* were utilized in at least one paper.

For example, De Oliveira et al. [45] developed a technique for automatically labeling medical event logs, a method classified under design science. The authors conducted a case study on a laparotomy process to validate their technique, demonstrating its application to real healthcare data.

4.4 Research Outcome

This section addresses the third research question of this study, which pertains to the types of research outcomes proposed by studies conducted in the pre-analysis stage of process mining. Our analysis identified that the primary type of outcome is a *method*, with 231 papers (88.8%) contributing to this category. However, it should be noted that papers may have multiple types of contributions. Therefore, the total number of papers in Figure 5 may be exceed number of papers included in the full-text review. Other types of research outcomes identified were *instantiations*, *models*, and *constructs*, with 138 (53.1%), 17 (6.5%), and 17 (6.5%) papers contributing to these categories, respectively. Notably, no paper the review contributed a *design theory* to the field.

The study by Li et al. [116] introduces a novel event log completeness definition termed *#TAR Completeness*, which addresses implicit dependencies in the model. This definition is categorized as a *construct* type of research outcome, as it establishes fundamental terminology essential for theoretical clarity and consistency. Subsequently, the authors propose an algorithm to generate a #TAR complete event log, classified under the *method* category

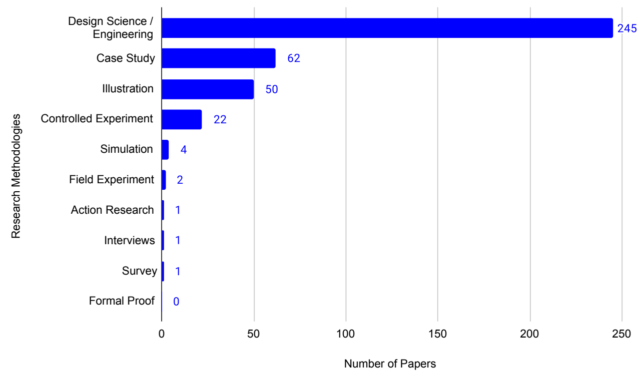

Fig. 4. Research methodologies employed by the reviewed papers

of research outcomes, as algorithms in this context denote systematic procedures designed to resolve specific challenges or achieve defined objectives within a specified domain. Additionally, the authors develop an executable version of their algorithm, exemplifying the *instantiation* category of research outcomes, wherein theoretical concepts are applied and operationalized into practical tools or processes.

Similarly, another study includes a *model* research outcome, specifically a meta-model designed to specify the types of process knowledge extractable from emails, thereby enabling process mining from email data [57]. This model categorization illustrates the conceptual frameworks that guide and structure research endeavors in applied settings.

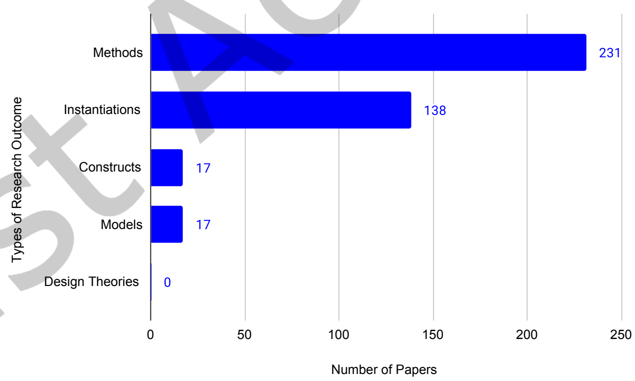

Fig. 5. Types of research outcome proposed by the reviewed papers

4.5 Evaluation Data

This section addresses the fourth research question of this study, which investigates the types of data utilized to evaluate the outcomes proposed in research focused on the pre-analysis stage of process mining. Out of 260 papers reviewed, only 200 (76.9%) mentioned the use of evaluation data. This data was categorized into two main types: real-life and artificial. According to the findings illustrated in Figure 6, most papers utilizing evaluation data (167 papers, or 83.5% of those using evaluation data) employed real-life data for their assessments. Among these, 132 papers (66% of those using evaluation data) exclusively utilized real-life data. In contrast, 59 papers (29.5% of those using evaluation data) used artificial data, with 24 papers (12% of those using evaluation data) exclusively utilizing artificial data. Additionally, 35 papers (17.5% of those using evaluation data) employed both real-life and artificial data in their evaluations. In comparison, nine papers (4.5% of those using evaluation data) did not explicitly specify the data type utilized.

For instance, Raichelson et al. [163] employed real-life data from the 2014 BPI challenge, alongside noise-free and noise-laden artificial data, to evaluate their proposed method for merging event logs.

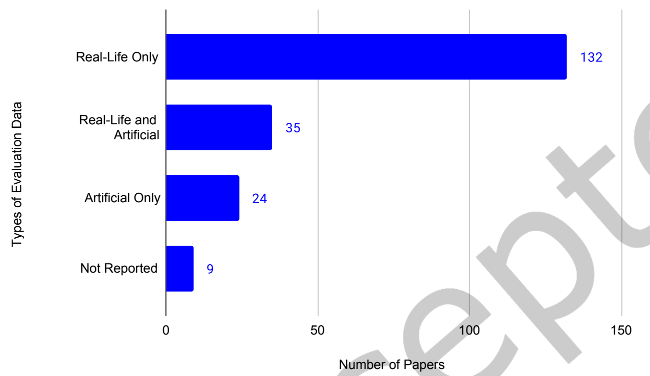

Fig. 6. Evaluation data types used to test the proposed outcome of the reviewed papers

Among the papers using real-life data, the top three industries providing evaluation data were *human health and social work activities* (56 papers, 33.5% of those papers that used real-life data), *financial and insurance activities* (54 papers, 32.3% of those papers that used real-life data), and *public administration and defense; compulsory social security* (46 papers, 27.5% of those papers that used real-life data).

Additionally, regarding the number of datasets used for evaluation, 110 papers (55% of those papers that used data for evaluation) used multiple datasets, while 90 papers (45% of those papers that used data for evaluation) used a single dataset.

4.6 Evolution of the Research Field

This section addresses the fifth research question regarding the evolution of the research field. As illustrated in Figure 7, there has been a consistent increase in the number of papers published in the field of the pre-analysis stage of process mining, highlighting the growing interest in this sub-field within process mining research. This growth has been particularly pronounced since the second half of the 2010s, with the number of papers published in 2019 alone surpassing the total published between 2005 and 2013. This trend supports the premise of this study that research on the pre-analysis stage of process mining is expanding substantially.

Figure 8 provides additional insight into trends in process mining research by categorizing the annual publication counts based on research goals. Initially, there was a significant focus on developing techniques for general process mining. However, starting from the latter half of the 2010s, there has been a notable proliferation of research targeting specific objectives such as privacy-aware and quality-aware process mining.

Overall, the upward trend in research on pre-analysis topics since the latter half of the 2010s, as depicted in Figures 8 and 9, suggests that researchers are placing more emphasis on the process mining pre-analysis stage, including data collection, cleaning, and preparation. This trend is further supported by the increase in research on privacy-aware and quality-aware process mining, as shown in Figure 8, indicating a growing awareness of the importance of these issues in the pre-analysis stage of process mining.

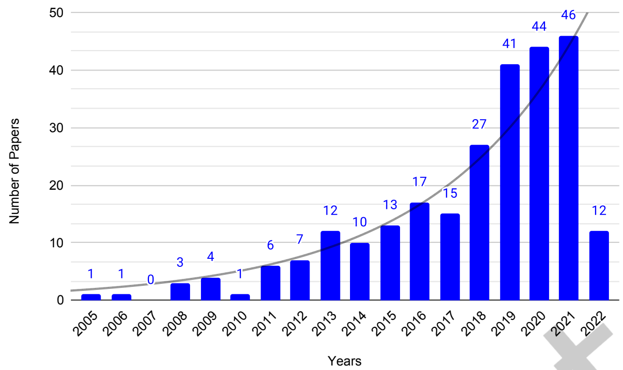

Fig. 7. Papers published in the research field of process mining pre-analysis (Note: the number for 2022 does not reflect the entire year)

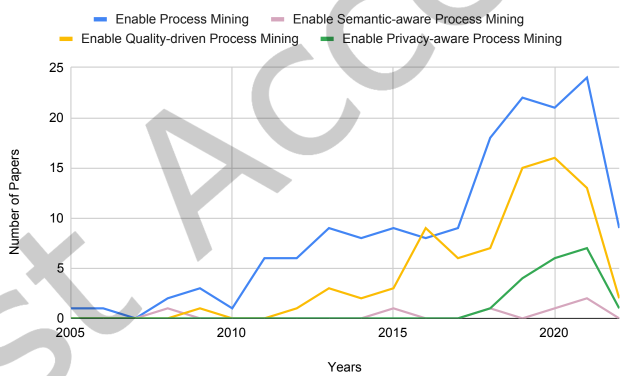

Fig. 8. Goals of the papers on the process mining pre-analysis stage

5 Discussion

This section elaborates on the results of our study. As depicted in Figures 8 and 9, there is a discernible increase in interest in the pre-analysis stage of process mining. Due to the critical dependence of process mining on high-quality process data, researchers and practitioners have made significant efforts to devise tools and techniques to aid various tasks during the pre-analysis stage. These efforts are crucial in ensuring the generation of suitable event logs for effective process mining.

Based on our findings (Section 4.1), the predominant objective of a significant portion of pre-analysis research in process mining has been to facilitate general process mining. However, with the emergence of novel paradigms

Fig. 9. Papers published on pre-analysis process mining sorted by research topics. *X-axis*: years. *Y-axis*: number of papers.

in recent years—such as predictive process mining, quality-aware process mining, and privacy-aware process mining—the significance of the pre-analysis stage has become increasingly apparent in these domains as well. An interesting trend in the research on the pre-analysis stage of process mining is the temporal growth in publications, as shown in Figure 8. Initially, from 2005 to the mid-2010s, there was an increase in research aimed at enabling general process mining. This was followed by a surge in studies focused on quality-driven process mining from the early to late 2010s. The late 2010s witnessed a rise in privacy-aware process mining research. In the early 2020s, there was a modest increase in semantics-aware process mining. This progression suggests that research objectives in process mining evolve in distinct phases. Moreover, the objective of privacy-aware process mining can be viewed as a subset of broader societal concerns, such as fairness, diversity, equity, and inclusion[6]. Over the past few years, the focus on privacy has been predominant within process mining pre-analysis research. However, this emphasis may shift in the future as other societal concerns gain prominence.

Concerning RQ1 of this study, which explores the research topics in process mining pre-analysis, we identified 15 distinct topics that we categorized into four broad themes: *getting the data*, *subsetting the data*, *data quality*, and *privacy*. Our analysis revealed that the majority of research has focused on *data quality* (117 papers, accounting for 45% of the total) and *getting the data* (116 papers, representing 44.5% of the total). Together, these two themes, which encompass 11 of the identified topics, account for over 80% of the papers reviewed in this study. It is important to note that this figure is not closer to 90% because some papers contribute to multiple themes. The prominence of these themes is unsurprising, given that the success of any process mining project is fundamentally dependent on (i) the availability of data and (ii) the quality of that data. The rapid global adoption of digitalization by both individuals and organizations has generated vast volumes of data[7] [8], yet the utility of this data is contingent upon its quality. This trend highlights the universal challenge of acquiring data that meets the necessary quality standards, not only within process mining but across various domains.

At the research topic level, the prominence of *event log-building* (51 papers, representing 19.6% of the total) as the leading research topic underscores the necessity of converting process data stored in information systems into event logs for subsequent process mining analyses. This transformation frequently exposes a range of data quality issues such missing data, timestamp issues, and formatting errors among others. Addressing these issues is crucial for effective analysis, which is reflected in the prominence of *log cleaning* as the second most researched topic in the pre-analysis stage (46 papers, representing 17.7% of the total). Lastly, *data generation*, identified as the third most researched topic (28 papers, accounting for 10.8% of the total), highlights the necessity of generating process data for testing the tools and techniques developed in process mining. It is noteworthy to highlight the distinction between the target end-users of the first two research topics and the third. While *event log-building* and *log cleaning* are pertinent to process mining applications within industry settings, *data generation* holds greater relevance for academic researchers as *data generation* primarily aims to advance process mining algorithms and contribute to the research field's overall development. As illustrated in Figure 9, there has been a notable increase in publications on topics such as *event log-building*, *log assessment*, and *log cleaning* since 2017. This parallel rise in publications underscores the necessity of assessing the quality of event logs and performing log cleaning when building event logs. Moreover, the increasing trend towards privacy in data management is reflected in the growth of publications on *data anonymization* and *private data sharing* during the latter half of the 2010s. Additionally, topics like *log division* and *log enrichment* have also seen an increase in research. The higher publication numbers in these areas may be attributed to the growing size of event logs and datasets, which necessitates efficient management and division into manageable, useful segments, as well as the enrichment of logs with contextual information to derive deeper insights.

Concerning RQ2 of this study, which addresses the applied research methodologies, it is evident from Figure 4 that a large majority of studies (245, accounting for 94.2% of the total) utilizes *design science research* as their primary methodology. This prevalence can be attributed to the focus of many processes mining pre-analysis research studies on developing novel methods, techniques, and approaches. Notably, only a quarter of the

[6]https://dictionary.cambridge.org/dictionary/english/dei
[7]https://www.oecd.org/en/about/news/press-releases/2024/05/growth-of-digital-economy-outperforms-overall-growth-across-oecd.html
[8]https://www.worldbank.org/en/news/immersive-story/2024/03/05/global-digitalization-in-10-charts

papers utilizing design science research also incorporate *case study* and *illustration* methodologies. This disparity underscores the challenge researchers face in conducting real-world studies. Furthermore, the number of papers developing instantiations of their contributions is significantly lower than those employing design science research. There are also some methodologies that occur less frequently such as *simulation*, *field experiment*, *action research*, *interviews*, and *surveys*. This scarcity can be attributed to two possible explanations. First, these methodologies may not be particularly suitable for process mining pre-analysis research. Alternatively, the research topic may not yet be sufficiently mature to warrant extensive investigation using these broader methodologies.

Our analysis of the data for RQ3 concerning the type of research outcomes reveals that most papers contribute either *methods* (231, representing 88.8% of the total) or *instantiations* (138, accounting for 53.1% of the total) because these research outcomes directly address the practical implementation and application of theories and concepts in process mining pre-analysis. The emphasis on *methods* and *instantiations* reflects the inherently applied nature of process mining, which relies on the development of practical solutions and tools that can be effectively implemented in real-world contexts. The relatively smaller number of papers producing *constructs* (17, representing 6.5% of the total) and *models* (17 representing 6.5% of the total) suggests two hypotheses: (i) few researchers are interested in conducting research at this level of process mining pre-analysis, or (ii) once constructs and models are devised and proposed, subsequent researchers use them as foundational elements for their work, thus not pursuing further research on the constructs and models themselves.

An observation from our review regarding RQ4 concerning the type of data that is used to evaluate solutions is the prevalent use of *real-life* data for evaluating research contributions, in contrast to the use of *artificial* data. While sourcing real-life data for evaluation is often perceived as challenging, particularly among researchers, our findings indicate that many researchers have successfully obtained such data. Nonetheless, this achievement should not be interpreted as an indication that the process is straightforward. Figure 6 illustrates that over five times as many (132, representing 50.8% of the total) studies utilize real-life data exclusively compared to those using exclusively artificial data (24, representing 9.2% of the total). However, it is plausible that multiple studies rely on the same dataset, as evidenced by the extensive use of the BPI Challenge[9] dataset in process mining research.

Regarding RQ5, which addresses the evolution of the research field of process mining pre-analysis, our study reveals that early research in this field had a sporadic character, with only ten papers published between 2005 and 2010. However, 2011 marked a significant increase, with the number of published papers rising from one in 2010 to six in 2011. Since then, the field has experienced steady growth, particularly in the late 2010s and early 2020s, when there was a rapid surge in the number of studies conducted on process mining pre-analysis. This growth is aligned with the increasing adoption of process mining as a crucial analytics tool by businesses[10], coupled with a significant rise in investments in and acquisitions of companies providing process mining tools and services[11] [12] [13]. Our data collection concluded in April 2022, at which point only 12 papers had been published that year. This figure is notably lower compared to the previous five years, which experienced at least a threefold increase in annual publications. If this downward trend continues, it may suggest that research in process mining pre-analysis reached its peak during the 2019-2021 period.

Despite the comprehensive nature of our systematic review, certain limitations must be acknowledged. First, while we selected prestigious and frequently utilized databases for data collection, this study did not include other databases that might be relevant. As a result, pertinent papers in those omitted databases may not be covered in our review. However, we carefully chose multiple well-regarded databases to mitigate this limitation to ensure

[9]https://www.tf-pm.org/competitions-awards/bpi-challenge
[10]https://www.gartner.com/en/documents/3991229
[11]https://earlybird.com/wp-content/uploads/2020/01/191016_Minit_Press-Release_English.pdf
[12]https://www.celonis.com/press/crunchbase-celonis-raises-1billion/
[13]https://ir.uipath.com/news/detail/69/uipath-acquires-processgold-to-deliver-unparalleled-end-to-end-process-understanding-solution

comprehensive coverage and reduce the risk of missing significant contributions. Second, the effectiveness of the review is inherently tied to the search terms used. To mitigate this limitation, we deliberately chose broad search terms to ensure comprehensive coverage across various aspects of the field, thereby reducing the likelihood of missing significant studies. Third, the selection criteria significantly influenced the inclusion and exclusion of studies. Although we adopted a conservative approach during the screening phases to ensure rigor, this does not entirely eliminate the possibility that some relevant papers might have been inadvertently overlooked. However, by applying carefully defined criteria and conducting multiple screening rounds, we aimed to minimize this risk and enhance the comprehensiveness of our review.

We foresee several potential avenues for future research in process mining pre-analysis. There has been a notable increase in research on topics such as *event log-building*, *log assessment*, and *log cleaning*, as illustrated by Figure 9. However, other topics, such as *privacy* and *log enrichment*, will likely be explored further. These two topics are particularly compelling as they resonate with contemporary societal concerns. Privacy has been a significant issue globally, especially following the unauthorized use of user data in the Facebook-Cambridge Analytica scandal in 2018[14]. Similarly, data integration has also gained prominence. In this context, enriching event logs with external data, such as relevant contextual information [193] or semantics [44], could be interesting. Further research could explore enriching event logs with additional contextual information, such as the state of mind of the person executing an activity or the temperature when the activity occurred.

There are several promising directions of future research for this particular literature review as well. Currently, we have identified four broad goals in pre-analysis process mining. Future investigations could benefit from examining these objectives at a finer level of granularity to uncover more specific goals. For instance, a detailed analysis of papers focused on *enabling process mining* might reveal more nuanced objectives. Another potential research avenue involves studying the types of information systems organizations employ from which data can be extracted to build event logs. This understanding could help process mining practitioners target specific types of information systems used in businesses for data extraction. Additionally, a further extension of this work could involve analyzing how the investigated dimensions interact with each other. For example, examining the relationship between research topics and methodologies could provide valuable insights, enabling researchers to identify the methodologies that have been, or should be, employed to explore specific topics within process mining pre-analysis.

6 Conclusion

This paper presents a systematic literature review of studies focusing on the pre-analysis stage of process mining. Our analysis has covered diverse dimensions, including research topics, methodologies employed, types of research outcomes, and the evaluation data utilized in assessing contributions. The primary contribution of this work lies in its comprehensive categorization of research topics pertinent to the pre-analysis stage of process mining, coupled with the identification of associated goals. Additionally, we have explored various research-related dimensions such as methodologies, outcome types, and evaluation data, uncovering emerging topics such as privacy and semantics in this domain.

This study bears significant implications for both researchers and practitioners in process mining. For researchers, it establishes a foundational baseline of research topics by categorizing pre-analysis topics comprehensively, thereby guiding future investigations and identifying avenues for further research. By synthesizing existing literature, our study also highlights areas where additional exploration is needed, facilitating the development of new research directions in process mining pre-analysis.

Practitioners can benefit from our study as a valuable reference tool to deepen their understanding of critical activities during event data pre-analysis. The detailed examination of topics and subtopics serves as a practical

[14]https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

guide for practitioners involved in data extraction, preprocessing, privacy assurance, and data quality verification. With enhanced knowledge of these pivotal pre-analysis tasks and concepts, practitioners can streamline their process analysis workflows, ultimately improving the accuracy and relevance of their findings in real-world applications.

Acknowledgments

This study was supported by the Special Research Fund (BOF) of Hasselt University under Grant Nos. BOF21OWB22 and BOF24TT02.

References

[1] Amine Abbad Andaloussi, Andrea Burattin, and Barbara Weber. 2018. Toward an Automated Labeling of Event Log Attributes. In *Enterprise, Business-Process and Information Systems Modeling*, Jens Gulden, Iris Reinhartz-Berger, Rainer Schmidt, Sérgio Guerreiro, Wided Guédria, and Palash Bera (Eds.). Springer International Publishing, Cham, 82–96. https://doi.org/10.1007/978-3-319-91704-7_6

[2] Rafael Accorsi and Thomas Stocker. 2013. SecSy: Synthesizing Process Event Logs. In *Enterprise Modelling and Information Systems Architectures (EMISA 2013)*. Gesellschaft für Informatik e.V., Bonn, 71–84. https://dl.gi.de/handle/20.500.12116/17247

[3] Anita S Acharya, Anupam Prakash, Pikee Saxena, and Aruna Nigam. 2013. Sampling: Why and how of it. *Indian Journal of Medical Specialities* 4, 2 (Dec. 2013), 330–333.

[4] Robert Andrews, Moe T. Wynn, Kirsten Vallmuur, Arthur H. M. ter Hofstede, Emma Bosley, Mark Elcock, and Stephen Rashford. 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. *International Journal of Environmental Research and Public Health* 16, 7 (Jan. 2019), 1138. https://doi.org/10.3390/ijerph16071138 Publisher: Multidisciplinary Digital Publishing Institute.

[5] Hanane Ariouat, Kamel Barkaoui, and Jacky Akoka. 2015. Improving process models discovery using AXOR clustering algorithm. In *Information Science and Applications*, Kuinam J. Kim (Ed.). Springer, Berlin, Heidelberg, 623–629. https://doi.org/10.1007/978-3-662-46578-3_73

[6] Femi Emmanuel Ayo, Olusegun Folorunso, and Friday Thomas Ibharalu. 2017. A probabilistic approach to event log completeness. *Expert Systems with Applications* 80 (Sept. 2017), 263–272. https://doi.org/10.1016/j.eswa.2017.03.039

[7] Rolf B. Banziger, Artie Basukoski, and Thierry Chaussalet. 2018. Discovering Business Processes in CRM Systems by Leveraging Unstructured Text Data. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, Exeter, UK, 1571–1577. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00257

[8] Vic Barnett and Toby Lewis. 1994. *Outliers in Statistical Data* (3 ed.). John Wiley & Sons, New York, NY, USA.

[9] Edgar Batista and Agusti Solanas. 2021. A uniformization-based approach to preserve individuals' privacy during process mining analyses. *Peer-to-Peer Networking and Applications* 14, 3 (May 2021), 1500–1519. https://doi.org/10.1007/s12083-020-01059-1

[10] Martin Bauer, Stephan A. Fahrenkrog-Petersen, Agnes Koschmider, Felix Mannhardt, Han van der Aa, and Matthias Weidlich. 2019. ELPaaS: Event Log Privacy as a Service. *159-163* (2019). https://sintef.brage.unit.no/sintef-xmlui/handle/11250/2725043

[11] Martin Bauer, Arik Senderovich, Avigdor Gal, Lars Grunske, and Matthias Weidlich. 2018. How Much Event Data Is Enough? A Statistical Framework for Process Discovery. In *Advanced Information Systems Engineering*, John Krogstie and Hajo A. Reijers (Eds.). Springer International Publishing, Cham, 239–256. https://doi.org/10.1007/978-3-319-91563-0_15

[12] Dina Bayomie, Ahmed Awad, and Ehab Ezat. 2016. Correlating Unlabeled Events from Cyclic Business Processes Execution. In *Advanced Information Systems Engineering*, Selmin Nurcan, Pnina Soffer, Marko Bajec, and Johann Eder (Eds.). Springer International Publishing, Cham, 274–289. https://doi.org/10.1007/978-3-319-39696-5_17

[13] Dina Bayomie, Iman M. A. Helal, Ahmed Awad, Ehab Ezat, and Ali ElBastawissi. 2016. Deducing Case IDs for Unlabeled Event Logs. In *Business Process Management Workshops*, Manfred Reichert and Hajo A. Reijers (Eds.). Springer International Publishing, Cham, 242–254. https://doi.org/10.1007/978-3-319-42887-1_20

[14] Paul Beck, Hendrik Bockrath, Tom Knoche, Mykola Digtiar, Tobias Petrich, Daniil Romanchenko, Richard Hobeck, Luise Pufahl, Christopher Klinkmüller, and Ingo Weber. 2021. BLF: A Blockchain Logging Framework for Mining Blockchain Data. In *Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2021*. Rome, Italy, 111–115.

[15] Yana A. Bekeneva. 2020. Algorithm for Generating Event Logs Based on Data from Heterogeneous Sources. In *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, St. Petersburg and Moscow, Russia, 233–236. https://doi.org/10.1109/EIConRus49466.2020.9039350 ISSN: 2376-6565.

[16] Yana A. Bekeneva. 2020. An Approach to the Distributed Generation of Event Logs Based on Data from Heterogeneous Monitoring Devices. In *2020 9th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, Budva, Montenegro, 1–4. https://doi.org/10.1109/MECO49872.2020.9134276 ISSN: 2637-9511.

[17] Marian Benner-Wickner, Matthias Book, Tobias Brückmann, and Volker Gruhn. 2014. Examining Case Management Demand Using Event Log Complexity Metrics. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference Workshops and Demonstrations*. IEEE, Ulm, Germany, 108–115. https://doi.org/10.1109/EDOCW.2014.25 ISSN: 2325-6605.

[18] Gaël Bernard and Periklis Andritsos. 2020. Truncated Trace Classifier. Removal of Incomplete Traces from Event Logs. In *Enterprise, Business-Process and Information Systems Modeling*, Selmin Nurcan, Iris Reinhartz-Berger, Pnina Soffer, and Jelena Zdravkovic (Eds.). Springer International Publishing, Cham, 150–165. https://doi.org/10.1007/978-3-030-49418-6_10

[19] Alessandro Berti, Gyunam Park, Majid Rafiei, and Wil M. P. Van Der Aalst. 2021. An event data extraction approach from SAP ERP for process mining. Springer.

[20] Mathilde Boltenhagen, Thomas Chatain, and Josep Carmona. 2019. Generalized Alignment-Based Trace Clustering of Process Behavior. In *Application and Theory of Petri Nets and Concurrency*, Susanna Donatelli and Stefan Haar (Eds.). Springer International Publishing, Cham, 237–257. https://doi.org/10.1007/978-3-030-21571-2_14

[21] R.P. Jagadeesh Chandra Bose and van der Aalst, W.M.P. 2009. Context aware trace clustering : towards improving process mining results. In *Proceedings of the Ninth SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics (SIAM), Sparks, NV, USA, 401–412.

[22] R.P. Jagadeesh Chandra Bose, Ronny S. Mans, and Wil M.P. van der Aalst. 2013. Wanna improve process mining results?. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, Singapore, 127–134. https://doi.org/10.1109/CIDM.2013.6597227

[23] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. 2010. Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models. In *Business Process Management Workshops (Lecture Notes in Business Information Processing)*, Stefanie Rinderle-Ma, Shazia Sadiq, and Frank Leymann (Eds.). Springer, Berlin, Heidelberg, 170–181. https://doi.org/10.1007/978-3-642-12186-9_16

[24] Melike Bozkaya, Joost Gabriels, and Jan Martijn van der Werf. 2009. Process Diagnostics: A Method Based on Process Mining. In *2009 International Conference on Information, Process, and Knowledge Management*. IEEE, Cancun, Mexico, 22–27. https://doi.org/10.1109/eKNOW.2009.29

[25] Edyta Brzychczy and Agnieszka Trzcionkowska. 2018. Creation of an Event Log from a Low-Level Machinery Monitoring System for Process Mining Purposes. In *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, Hujun Yin, David Camacho, Paulo Novais, and Antonio J. Tallón-Ballesteros (Eds.). Springer International Publishing, Cham, 54–63. https://doi.org/10.1007/978-3-030-03496-2_7

[26] Scott Buffett and Liqiang Geng. 2010. Using classification methods to label tasks in process mining. *Journal of Software Maintenance and Evolution: Research and Practice* 22, 6-7 (2010), 497–517. https://doi.org/10.1002/smr.463

[27] Diego Calvanese, Tahir Emre Kalayci, Marco Montali, and Ario Santoso. 2017. OBDA for Log Extraction in Process Mining. In *Reasoning Web. Semantic Interoperability on the Web: 13th International Summer School 2017, London, UK, July 7-11, 2017, Tutorial Lectures*, Giovambattista Ianni, Domenico Lembo, Leopoldo Bertossi, Wolfgang Faber, Birte Glimm, Georg Gottlob, and Steffen Staab (Eds.). Springer International Publishing, Cham, 292–345. https://doi.org/10.1007/978-3-319-61033-7_9

[28] Diego Calvanese, Tahir Emre Kalayci, Marco Montali, and Ario Santoso. 2017. The onprom Toolchain for Extracting Business Process Logs using Ontology-based Data Access. In *Proceedings of the BPM Demo Track and BPM Dissertation Award (CEUR Workshop Proceedings, Vol. 1920)*. CEUR Workshop Proceedings, Barcelona, Spain.

[29] Diego Calvanese, Tahir Emre Kalayci, Marco Montali, and Stefano Tinella. 2017. Ontology-Based Data Access for Extracting Event Logs from Legacy Data: The onprom Tool and Methodology. In *Business Information Systems*, Witold Abramowicz (Ed.). Springer International Publishing, Cham, 220–236. https://doi.org/10.1007/978-3-319-59336-4_16

[30] Diego Calvanese, Marco Montali, Alifah Syamsiyah, and Wil M. P. van der Aalst. 2016. Ontology-Driven Extraction of Event Logs from Relational Databases. In *Business Process Management Workshops*, Manfred Reichert and Hajo A. Reijers (Eds.). Springer International Publishing, Cham, 140–153. https://doi.org/10.1007/978-3-319-42887-1_12

[31] Josep Carmona, Jordi Cortadella, and Michael Kishinevsky. 2009. Divide-and-Conquer Strategies for Process Mining. In *Business Process Management*, Umeshwar Dayal, Johann Eder, Jana Koehler, and Hajo A. Reijers (Eds.). Springer, Berlin, Heidelberg, 327–343. https://doi.org/10.1007/978-3-642-03848-8_22

[32] Paolo Ceravolo, Ernesto Damiani, Mohammadsadegh Torabi, and Sylvio Barbon. 2017. Toward a New Generation of Log Pre-processing Methods for Process Mining. In *Business Process Management Forum*, Josep Carmona, Gregor Engels, and Akhil Kumar (Eds.). Springer International Publishing, Cham, 55–70. https://doi.org/10.1007/978-3-319-65015-9_4

[33] L. Chen, S. Kang, S. Karimidorabati, and C. T. Haas. 2019. Improving the Quality of Event Logs in the Construction Industry for Process Mining. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, Vol. 36. IAARC Publications, Waterloo, Canada, 804–811. https://doi.org/10.22260/ISARC2019/0108

[34] Federico Chesani, Riccardo De Masellis, Chiara Di Francescomarino, Chiara Ghidini, Paola Mello, Marco Montali, and Sergio Tessaris. 2016. Abducing Compliance of Incomplete Event Logs. In *AI*IA 2016 Advances in Artificial Intelligence*, Giovanni Adorni, Stefano Cagnoni, Marco Gori, and Marco Maratea (Eds.). Springer International Publishing, Cham, 208–222. https://doi.org/10.1007/978-3-319-49130-1_16

[35] Jan Claes and Geert Poels. 2014. Merging event logs for process mining: A rule based merging method and rule suggestion algorithm. *Expert Systems with Applications* 41, 16 (Nov. 2014), 7291–7306. https://doi.org/10.1016/j.eswa.2014.06.012

[36] Raffaele Conforti, Marcello La Rosa, and Arthur H.M. ter Hofstede. 2017. Filtering Out Infrequent Behavior from Business Process Event Logs. *IEEE Transactions on Knowledge and Data Engineering* 29, 2 (Feb. 2017), 300–314. https://doi.org/10.1109/TKDE.2016.2614680 Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[37] Kevin G. Corley and Dennis A. Gioia. 2004. Identity Ambiguity and Change in the Wake of a Corporate Spin-Off. *Administrative Science Quarterly* 49, 2 (2004), 173–208. https://www.jstor.org/stable/4131471 Publisher: Sage Publications, Inc., Johnson Graduate School of Management, Cornell University.

[38] Benjamin F. Crabtree and William F. Miller. 1992. A template approach to text analysis: Developing and using codebooks. In *Doing qualitative research*. Sage Publications, Inc, Thousand Oaks, CA, US, 93–109.

[39] John W. Creswell. 2003. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Thousand Oaks, CA, USA.

[40] A. Cuzzocrea, E. Damiani, H. Al-Ali, R. Mizouni, G. Tello, and E. Fadda. 2021. From Low-Level-Event-Logs to High-Level-Business-Process-Model-Activities: An Advanced Framework based on Machine Learning and Flexible BPMN Model Translation. CEUR Workshop Proceedings. https://air.unimi.it/handle/2434/881403

[41] Dusanka Dakic, Darko Stefanovic, Teodora Lolic, Dajana Narandzic, and Nenad Simeunovic. 2020. Event Log Extraction for the Purpose of Process Mining: A Systematic Literature Review. In *Innovation in Sustainable Management and Entrepreneurship*, Gabriela Prostean, Juan José Lavios Villahoz, Laura Brancu, and Gyula Bakacsi (Eds.). Springer International Publishing, Cham, 299–312. https://doi.org/10.1007/978-3-030-44711-3_22

[42] Pieter De Koninck and Jochen De Weerdt. 2017. Multi-objective Trace Clustering: Finding More Balanced Solutions. In *Business Process Management Workshops*, Marlon Dumas and Marcelo Fantinato (Eds.). Springer International Publishing, Cham, 49–60. https://doi.org/10.1007/978-3-319-58457-7_4

[43] Massimiliano de Leoni and Safa Dündar. 2020. Event-log abstraction using batch session identification and clustering. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC '20)*. Association for Computing Machinery, New York, NY, USA, 36–44. https://doi.org/10.1145/3341105.3373861

[44] Ana Karla Alves de Medeiros, Wil van der Aalst, and Carlos Pedrinaci. 2008. Semantic process mining tools: core building blocks. Galway, Ireland. http://is2.lse.ac.uk/asp/aspecis/20080169.pdf

[45] Hugo De Oliveira, Vincent Augusto, Baptiste Jouaneton, Ludovic Lamarsalle, Martin Prodel, and Xiaolan Xie. 2020. Automatic and Explainable Labeling of Medical Event Logs With Autoencoding. *IEEE Journal of Biomedical and Health Informatics* 24, 11 (Nov. 2020), 3076–3084. https://doi.org/10.1109/JBHI.2020.3021790 Conference Name: IEEE Journal of Biomedical and Health Informatics.

[46] Jochen De Weerdt and Moe Thandar Wynn. 2022. Foundations of Process Event Data. In *Process Mining Handbook*. Springer, Aachen, Germany.

[47] Vadim Denisov, Dirk Fahland, and Wil M. P. van der Aalst. 2020. Repairing Event Logs with Missing Events to Support Performance Analysis of Systems with Shared Resources. In *Application and Theory of Petri Nets and Concurrency*, Ryszard Janicki, Natalia Sidorova, and Thomas Chatain (Eds.). Springer International Publishing, Cham, 239–259. https://doi.org/10.1007/978-3-030-51831-8_12

[48] Amit V. Deokar and Jie Tao. 2015. Semantics-based event log aggregation for process mining and analytics. *Information Systems Frontiers* 17, 6 (Dec. 2015), 1209–1226. https://doi.org/10.1007/s10796-015-9563-4

[49] Claudio Di Ciccio, Mario Luca Bernardi, Marta Cimitile, and Fabrizio Maria Maggi. 2015. Generating Event Logs Through the Simulation of Declare Models. In *Enterprise and Organizational Modeling and Simulation*, Joseph Barjis, Robert Pergl, and Eduard Babkin (Eds.). Springer International Publishing, Cham, 20–36. https://doi.org/10.1007/978-3-319-24626-0_2

[50] Prabhakar M. Dixit, Suriadi Suriadi, Robert Andrews, Moe T. Wynn, Arthur H. M. ter Hofstede, Joos C. A. M. Buijs, and Wil M. P. van der Aalst. 2018. Detection and Interactive Repair of Event Ordering Imperfection in Process Logs. In *Advanced Information Systems Engineering*, John Krogstie and Hajo A. Reijers (Eds.). Springer International Publishing, Cham, 274–290. https://doi.org/10.1007/978-3-319-91563-0_17

[51] Boudewijn Dongen and Wil Aalst. 2005. A Meta Model for Process Mining Data. In *Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability*, Vol. 2. Porto, Portugal.

[52] Yutika Amelia Effendi and Fitri Retrialisca. 2021. Transforming Timestamp of a CSV Database to an Event Log. In *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, Kuala Lumpur, Malaysia, 367–372. https://doi.org/10.1109/ICCCE50029.2021.9467226

[53] Diana El-Masri, Fabio Petrillo, Yann-Gaël Guéhéneuc, Abdelwahab Hamou-Lhadj, and Anas Bouziane. 2020. A systematic literature review on automated log abstraction techniques. *Information and Software Technology* 122 (June 2020), 106276. https://doi.org/10.1016/j.infsof.2020.106276

[54] Gamal Elkoumy, Stephan A. Fahrenkrog-Petersen, Marlon Dumas, Peeter Laud, Alisa Pankova, and Matthias Weidlich. 2020. Secure Multi-party Computation for Inter-organizational Process Mining. In *Enterprise, Business-Process and Information Systems Modeling*, Selmin Nurcan, Iris Reinhartz-Berger, Pnina Soffer, and Jelena Zdravkovic (Eds.). Springer International Publishing, Cham, 166–181.

https://doi.org/10.1007/978-3-030-49418-6_11

[55] Gamal Elkoumy, Stephan A. Fahrenkrog-Petersen, Marlon Dumas, Peeter Laud, Alisa Pankova, and Matthias Weidlich. 2020. Shareprom: A Tool for Privacy-Preserving Inter-Organizational Process Mining. In *Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2020*, Vol. 2673. Sevilla, Spain, 72–76.

[56] Gamal Elkoumy, Alisa Pankova, and Marlon Dumas. 2021. Mine Me but Don't Single Me Out: Differentially Private Event Logs for Process Mining. In *2021 3rd International Conference on Process Mining (ICPM)*. IEEE, Eindhoven, Netherlands, 80–87. https://doi.org/10.1109/ICPM53251.2021.9576852

[57] Marwa Elleuch, Nour Assy, Nassim Laga, Walid Gaaloul, Oumaima Alaoui Ismaili, and Boualem Benatallah. 2020. A Meta Model for Mining Processes from Email Data. In *2020 IEEE International Conference on Services Computing (SCC)*. IEEE, Beijing, China, 152–161. https://doi.org/10.1109/SCC49832.2020.00028 ISSN: 2474-2473.

[58] R. Engel, R.P. Jagadeesh Chandra Bose, C. Pichler, M. Zapletal, and H. Werthner. 2013. EDIminer : a toolset for process mining from EDI messages. *Proceedings of the CAiSE'13 Forum at the 25th International Conference on Advanced Information Systems Engineering (CAiSE, Valencia, Spain, June 20, 2013))* (2013), 146–153. CEUR Workshop Proceedings.

[59] Robert Engel, Worarat Krathu, Marco Zapletal, Christian Pichler, R. P. Jagadeesh Chandra Bose, Wil van der Aalst, Hannes Werthner, and Christian Huemer. 2016. Analyzing inter-organizational business processes. *Information Systems and e-Business Management* 14, 3 (Aug. 2016), 577–612. https://doi.org/10.1007/s10257-015-0295-2

[60] Eren Esgin and Pinar Karagoz. 2019. Process Profiling based Synthetic Event Log Generation.. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*. Vienna, Austria, 516–524. https://doi.org/10.5220/0008363805160524

[61] Stefan Esser and Dirk Fahland. 2021. Multi-Dimensional Event Data in Graph Databases. *Journal on Data Semantics* 10, 1 (June 2021), 109–141. https://doi.org/10.1007/s13740-021-00122-1

[62] Eurostat. 2008. *Statistical Classification of Economic Activities in the European Community, Rev. 2.* Eurostat. https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC&IntCurrentPage=1

[63] Joerg Evermann, Tom Thaler, and Peter Fettke. 2016. Clustering Traces Using Sequence Alignment. In *Business Process Management Workshops*, Manfred Reichert and Hajo A. Reijers (Eds.). Springer International Publishing, Cham, 179–190. https://doi.org/10.1007/978-3-319-42887-1_15

[64] Stephan A. Fahrenkog-Petersen, Martin Kabierski, Fabian Rösel, Han van der Aa, and Matthias Weidlich. 2021. SaCoFa: Semantics-aware Control-flow Anonymization for Process Mining. In *2021 3rd International Conference on Process Mining (ICPM)*. IEEE, Eindhoven, Netherlands, 72–79. https://doi.org/10.1109/ICPM53251.2021.9576857

[65] Stephan A. Fahrenkrog-Petersen, Han van der Aa, and Matthias Weidlich. 2019. PRETSA: Event Log Sanitization for Privacy-aware Process Discovery. In *2019 International Conference on Process Mining (ICPM)*. IEEE, Aachen, Germany, 1–8. https://doi.org/10.1109/ICPM.2019.00012

[66] Stephan A. Fahrenkrog-Petersen, Han van der Aa, and Matthias Weidlich. 2020. PRIPEL: Privacy-Preserving Event Log Publishing Including Contextual Information. In *Business Process Management*, Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas (Eds.). Springer International Publishing, Cham, 111–128. https://doi.org/10.1007/978-3-030-58666-9_7

[67] Mohammadreza Fani Sani. 2020. Preprocessing Event Data in Process Mining. In *Doctoral Consortium Papers Presented at the 32nd International Conference on Advanced Information Systems Engineering (CAiSE 2020)*, Vol. 2613. CEUR Workshop Proceedings, Grenoble, France, 1–10. http://ceur-ws.org/Vol-2613/paper1.pdf

[68] Mohammadreza Fani Sani, Sebastiaan J. van Zelst, and Wil M. P. van der Aalst. 2018. Repairing Outlier Behaviour in Event Logs. In *Business Information Systems (Lecture Notes in Business Information Processing)*, Witold Abramowicz and Adrian Paschke (Eds.). Springer International Publishing, Cham, 115–131. https://doi.org/10.1007/978-3-319-93931-5_9

[69] Mohammadreza Fani Sani, Sebastiaan J. van Zelst, and Wil M. P. van der Aalst. 2018. Improving Process Discovery Results by Filtering Outliers Using Conditional Behavioural Probabilities. In *Business Process Management Workshops (Lecture Notes in Business Information Processing)*, Ernest Teniente and Matthias Weidlich (Eds.). Springer International Publishing, Cham, 216–229. https://doi.org/10.1007/978-3-319-74030-0_16

[70] Mohammadreza Fani Sani, Mozhgan Vazifehdoostirani, Gyunam Park, Marco Pegoraro, Sebastiaan J. van Zelst, and Wil M. P. van der Aalst. 2022. Event Log Sampling for Predictive Monitoring. In *Process Mining Workshops (Lecture Notes in Business Information Processing)*, Jorge Munoz-Gama and Xixi Lu (Eds.). Springer International Publishing, Cham, 154–166. https://doi.org/10.1007/978-3-030-98581-3_12

[71] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (March 2006), 80–92. https://doi.org/10.1177/160940690600500107 Publisher: SAGE Publications Inc.

[72] Diogo R. Ferreira and Daniel Gillblad. 2009. Discovering Process Models from Unlabelled Event Logs. In *Business Process Management*, Umeshwar Dayal, Johann Eder, Jana Koehler, and Hajo A. Reijers (Eds.). Springer, Berlin, Heidelberg, 143–158. https://doi.org/10.1007/978-3-642-03848-8_11

[73] Dominik Andreas Fischer, Kanika Goel, Robert Andrews, Christopher Gerhard Johannes van Dun, Moe Thandar Wynn, and Maximilian Röglinger. 2020. Enhancing Event Log Quality: Detecting and Quantifying Timestamp Imperfections. In *Business Process Management*, Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas (Eds.). Springer International Publishing, Cham, 309–326. https://doi.org/10.1007/978-3-030-58666-9_18

[74] D. A. Fischer, K. Goel, R. Andrews, C. G. J. van Dun, M. T. Wynn, and M. Röglinger. 2022. Towards interactive event log forensics: Detecting and quantifying timestamp imperfections. *Information Systems* 109 (Nov. 2022), 102039. https://doi.org/10.1016/j.is.2022.102039

[75] Roberto Gatta, Mauro Vallati, Jacopo Lenkowicz, Calogero Casà, Francesco Cellini, Andrea Damiani, and Vincenzo Valentini. 2018. A Framework for Event Log Generation and Knowledge Representation for Process Mining in Healthcare. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, Volos, Greece, 647–654. https://doi.org/10.1109/ICTAI.2018.00103 ISSN: 2375-0197.

[76] Anahita Farhang Ghahfarokhi, Gyunam Park, Alessandro Berti, and Wil M. P. van der Aalst. 2021. OCEL: A Standard for Object-Centric Event Logs. In *New Trends in Database and Information Systems*, Ladjel Bellatreche, Marlon Dumas, Panagiotis Karras, Raimundas Matulevičius, Ahmed Awad, Matthias Weidlich, Mirjana Ivanović, and Olaf Hartig (Eds.). Springer International Publishing, Cham, 169–175. https://doi.org/10.1007/978-3-030-85082-1_16

[77] Dennis A. Gioia, Kevin G. Corley, and Aimee L. Hamilton. 2013. Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods* 16, 1 (Jan. 2013), 15–31. https://doi.org/10.1177/1094428112452151 Publisher: SAGE Publications Inc.

[78] Stijn Goedertier, David Martens, Bart Baesens, Raf Haesen, and Jan Vanthienen. 2008. Process Mining as First-Order Classification Learning on Logs with Negative Events. In *Business Process Management Workshops*, Arthur ter Hofstede, Boualem Benatallah, and Hye-Young Paik (Eds.). Springer, Berlin, Heidelberg, 42–53. https://doi.org/10.1007/978-3-540-78238-4_6

[79] Stijn Goedertier, David Martens, Jan Vanthienen, and Bart Baesens. 2009. Robust Process Discovery with Artificial Negative Events. *Journal of Machine Learning Research* 10, 2009 (June 2009), 1305–1340.

[80] E. González López de Murillas, G. E. Hoogendoorn, and Hajo A. Reijers. 2018. Redo Log Process Mining in Real Life: Data Challenges & Opportunities. In *Business Process Management Workshops*, Ernest Teniente and Matthias Weidlich (Eds.). Springer International Publishing, Cham, 573–587. https://doi.org/10.1007/978-3-319-74030-0_45

[81] Eduardo González López de Murillas, Hajo A. Reijers, and Wil M. P. van der Aalst. 2017. Everything You Always Wanted to Know About Your Process, but Did Not Know How to Ask. In *Business Process Management Workshops*, Marlon Dumas and Marcelo Fantinato (Eds.). Springer International Publishing, Cham, 296–309. https://doi.org/10.1007/978-3-319-58457-7_22

[82] Seong-Hun Ham, Dihn-Lam Pham, Kyoung-Sook Kim, and Kwanghoon Pio Kim. 2020. Process-aware Event Log Datacubes for Workflow Process and Knowledge Mining, Predicting and Analyzing Frameworks. In *2020 22nd International Conference on Advanced Communication Technology (ICACT)*. IEEE, Phoenix Park, Korea (South), 651–657. https://doi.org/10.23919/ICACT48636.2020.9061449 ISSN: 1738-9445.

[83] Iman M. A. Helal and Ahmed Awad. 2022. Online correlation for unlabeled process events: A flexible CEP-based approach. *Information Systems* 108 (Sept. 2022), 102031. https://doi.org/10.1016/j.is.2022.102031

[84] Emmanuel Helm, Oliver Krauss, Anna Lin, Andreas Pointner, Andreas Schuler, and Josef Küng. 2021. Process Mining on FHIR - An Open Standards-Based Process Analytics Approach for Healthcare. In *Process Mining Workshops*, Sander Leemans and Henrik Leopold (Eds.). Springer International Publishing, Cham, 343–355. https://doi.org/10.1007/978-3-030-72693-5_26

[85] E. Helm and F. Paster. 2015. First Steps Towards Process Mining in Distributed Health Information Systems. *International Journal of Electronics and Telecommunications* Vol. 61, No. 2 (2015), 137–142. https://doi.org/10.1515/eletel-2015-0017

[86] Jaciel David Hernandez-Resendiz, Edgar Tello-Leal, Heidy Marisol Marin-Castro, Ulises Manuel Ramirez-Alcocer, and Jonathan Alfonso Mata-Torres. 2021. Merging Event Logs for Inter-organizational Process Mining. In *New Perspectives on Enterprise Decision-Making Applying Artificial Intelligence Techniques*, Julian Andres Zapata-Cortes, Giner Alor-Hernández, Cuauhtémoc Sánchez-Ramírez, and Jorge Luis García-Alcaraz (Eds.). Springer International Publishing, Cham, 3–26. https://doi.org/10.1007/978-3-030-71115-3_1

[87] Markku Hinkka, Teemu Lehto, and Keijo Heljanko. 2020. Exploiting Event Log Event Attributes in RNN Based Prediction. In *Data-Driven Process Discovery and Analysis*, Paolo Ceravolo, Maurice van Keulen, and María Teresa Gómez-López (Eds.). Springer International Publishing, Cham, 67–85. https://doi.org/10.1007/978-3-030-46633-6_4

[88] A. Hmami, H. Sbai, and M. Fredj. 2021. Enhancing change mining from a collection of event logs: Merging and Filtering approaches. *Journal of Physics: Conference Series* 1743, 1 (Jan. 2021), 012020. https://doi.org/10.1088/1742-6596/1743/1/012020 Publisher: IOP Publishing.

[89] Hiroki Horita, Yuta Kurihashi, and Nozomi Miyamori. 2020. Extraction of Missing Tendency Using Decision Tree Learning in Business Process Event Log. *Data* 5, 3 (Sept. 2020), 82. https://doi.org/10.3390/data5030082 Publisher: Multidisciplinary Digital Publishing Institute.

[90] Ying Huang, Yingxu Wang, and Yiwang Huang. 2018. Filtering Out Infrequent Events by Expectation from Business Process Event Logs. In *2018 14th International Conference on Computational Intelligence and Security (CIS)*. IEEE, Hangzhou, China, 374–377.

https://doi.org/10.1109/CIS2018.2018.00089
[91] Ying Huang, Liyun Zhong, and Yan Chen. 2020. Filtering Infrequent Behavior in Business Process Discovery by Using the Minimum Expectation. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 14, 2 (April 2020), 1–15. https://doi.org/10.4018/IJCINI.2020040101 Publisher: IGI Global.
[92] Jon Espen Ingvaldsen and Jon Atle Gulla. 2008. Preprocessing Support for Large Scale Process Mining of SAP Transactions. In *Business Process Management Workshops*, Arthur ter Hofstede, Boualem Benatallah, and Hye-Young Paik (Eds.). Springer, Berlin, Heidelberg, 30–41. https://doi.org/10.1007/978-3-540-78238-4_5
[93] Constantina Ioannou, Andrea Burattin, and Barbara Weber. 2018. Mining Developers' Workflows from IDE Usage. In *Advanced Information Systems Engineering Workshops*, Raimundas Matulevičius and Remco Dijkman (Eds.). Springer International Publishing, Cham, 167–179. https://doi.org/10.1007/978-3-319-92898-2_14
[94] Md Rofiqul Islam and Tomas Cerny. 2021. Business Process Extraction Using Static Analysis. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, Melbourne, Australia, 1202–1204. https://doi.org/10.1109/ASE51524.2021.9678588 ISSN: 2643-1572.
[95] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. 2009. Abstractions in Process Mining: A Taxonomy of Patterns. In *Business Process Management*, Umeshwar Dayal, Johann Eder, Jana Koehler, and Hajo A. Reijers (Eds.). Springer, Berlin, Heidelberg, 159–175. https://doi.org/10.1007/978-3-642-03848-8_12
[96] Amin Jalali. 2021. Graph-Based Process Mining. In *Process Mining Workshops*, Sander Leemans and Henrik Leopold (Eds.). Springer International Publishing, Cham, 273–285. https://doi.org/10.1007/978-3-030-72693-5_21
[97] Mieke Jans. 2019. Auditor Choices during Event Log Building for Process Mining. *Journal of Emerging Technologies in Accounting* 16, 2 (Sept. 2019), 59–67. https://doi.org/10.2308/jeta-52496
[98] Mieke Jans, Pnina Soffer, and Toon Jouck. 2019. Building a valuable event log for process mining: an experimental exploration of a guided process. *Enterprise Information Systems* 13, 5 (May 2019), 601–630. https://doi.org/10.1080/17517575.2019.1587788
[99] Toon Jouck and Benoît Depaire. 2016. PTandLogGenerator: A Generator for Artificial Event Data. In *Proceedings of the BPM Demo Track 2016*, Vol. 1789. Rio de Janeiro, Brazil, 23–27.
[100] Toon Jouck and Benoît Depaire. 2019. Generating Artificial Data for Empirical Analysis of Control-flow Discovery Algorithms. *Business & Information Systems Engineering* 61, 6 (Dec. 2019), 695–712. https://doi.org/10.1007/s12599-018-0541-5
[101] Norrapon Joyfong, Sompong Tumswadi, Parham Porouhan, Poohridate Arpasat, and Wichian Premchaiswadi. 2019. Preparation of Smart Card Data for Food Purchase Analysis of Students through Process Mining. In *2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE)*. IEEE, Bangkok, Thailand, 1–4. https://doi.org/10.1109/ICTKE47035.2019.8966932 ISSN: 2157-099X.
[102] Sylvio Barbon Junior, Paolo Ceravolo, Ernesto Damiani, Nicolas Jashchenko Omori, and Gabriel Marques Tavares. 2020. Anomaly Detection on Event Logs with a Scarcity of Labels. In *2020 2nd International Conference on Process Mining (ICPM)*. IEEE, Padua, Italy, 161–168. https://doi.org/10.1109/ICPM49681.2020.00032
[103] Martin Kabierski, Hoang Lam Nguyen, Lars Grunske, and Matthias Weidlich. 2021. Sampling What Matters: Relevance-guided Sampling of Event Logs. In *2021 3rd International Conference on Process Mining (ICPM)*. IEEE, Eindhoven, Netherlands, 64–71. https://doi.org/10.1109/ICPM53251.2021.9576875
[104] Christoph Kecht, Andreas Egger, Wolfgang Kratsch, and Maximilian Röglinger. 2021. Event Log Construction from Customer Service Conversations Using Natural Language Inference. In *2021 3rd International Conference on Process Mining (ICPM)*. IEEE, Eindhoven, Netherlands, 144–151. https://doi.org/10.1109/ICPM53251.2021.9576869
[105] Aicha Khannat, Hanae Sbai, and Laila Kjiri. 2020. Event Logs Pre-processing for Configurable Process Discovery: Ontology-Based Approach. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*. IEEE, Agadir - Essaouira, Morocco, 139–144. https://doi.org/10.1109/CiSt49399.2021.9357303 ISSN: 2327-1884.
[106] Sönke Knoch, Shreeraman Ponpathirkoottam, and Tim Schwartz. 2020. Video-to-Model: Unsupervised Trace Extraction from Videos for Process Discovery and Conformance Checking in Manual Assembly. In *Business Process Management*, Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas (Eds.). Springer International Publishing, Cham, 291–308. https://doi.org/10.1007/978-3-030-58666-9_17
[107] Bram Knols and Jan Martijn E. M. van der Werf. 2019. Measuring the Behavioral Quality of Log Sampling. In *2019 International Conference on Process Mining (ICPM)*. IEEE, Aachen, Germany, 97–104. https://doi.org/10.1109/ICPM.2019.00024
[108] Jonghyeon Ko, Jongyup Lee, and Marco Comuzzi. 2020. AIR-BAGEL: An Interactive Root cause-Based Anomaly Generator for Event Logs. In *Proceedings of the ICPM Doctoral Consortium and Tool Demonstration Track 2020*. Padua, Italy, 35–38.
[109] Li Kong, Chuanyi Li, Jidong Ge, Zhongjin Li, Feifei Zhang, and Bin Luo. 2019. An Efficient Heuristic Method for Repairing Event Logs Independent of Process Models.. In *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security - Volume 1: IoTBDS*, Vol. 1. SciTePress, 83–93. https://doi.org/10.5220/0007676400830093
[110] Agnes Koschmider. 2017. Clustering Event Traces by Behavioral Similarity. In *Advances in Conceptual Modeling*, Sergio de Cesare and Ulrich Frank (Eds.). Springer International Publishing, Cham, 36–42. https://doi.org/10.1007/978-3-319-70625-2_4
[111] Agnes Koschmider, Felix Mannhardt, and Tobias Heuser. 2019. On the Contextualization of Event-Activity Mappings. In *Business Process Management Workshops*, Florian Daniel, Quan Z. Sheng, and Hamid Motahari (Eds.). Springer International Publishing, Cham,

445–457. https://doi.org/10.1007/978-3-030-11641-5_35

[112] Agnes Koschmider and Daniel Siqueira Vidal Moreira. 2018. Change Detection in Event Logs by Clustering. In *On the Move to Meaningful Internet Systems. OTM 2018 Conferences (Lecture Notes in Computer Science)*, Hervé Panetto, Christophe Debruyne, Henderik A. Proper, Claudio Agostino Ardagna, Dumitru Roman, and Robert Meersman (Eds.). Springer International Publishing, Cham, 643–660. https://doi.org/10.1007/978-3-030-02610-3_36

[113] Klaus Krippendorff. 2011. *Computing Krippendorff's Alpha-Reliability.* Technical Report. University of Pennsylvania - Annenberg School of Communication. https://repository.upenn.edu/asc_papers/43

[114] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310 Publisher: Wiley, International Biometric Society.

[115] Giorgio Leonardi, Manuel Striani, Silvana Quaglini, Anna Cavallini, and Stefania Montani. 2018. Leveraging semantic labels for multi-level abstraction in medical process mining and trace comparison. *Journal of Biomedical Informatics* 83 (July 2018), 10–24. https://doi.org/10.1016/j.jbi.2018.05.012

[116] Chuanyi Li, Jidong Ge, Lijie Wen, Li Kong, Victor Chang, Liguo Huang, and Bin Luo. 2020. A novel completeness definition of event logs and corresponding generation algorithm. *Expert Systems* 37, 4 (2020), e12529. https://doi.org/10.1111/exsy.12529

[117] Guangming Li, Eduardo González López de Murillas, Renata Medeiros de Carvalho, and Wil M. P. van der Aalst. 2018. Extracting Object-Centric Event Logs to Support Process Mining on Databases. In *Information Systems in the Big Data Era*, Jan Mendling and Haralambos Mouratidis (Eds.). Springer International Publishing, Cham, 182–199. https://doi.org/10.1007/978-3-319-92901-9_16

[118] Guangming Li, Renata Medeiros de Carvalho, and Wil M.P. van der Aalst. 2018. Configurable Event Correlation for Process Discovery from Object-Centric Event Data. In *2018 IEEE International Conference on Web Services (ICWS)*. IEEE, San Francisco, CA, USA, 203–210. https://doi.org/10.1109/ICWS.2018.00033

[119] Jiexun Li, Harry Jiannan Wang, and Xue Bai. 2015. An intelligent approach to data extraction and task identification for process mining. *Information Systems Frontiers* 17, 6 (Dec. 2015), 1195–1208. https://doi.org/10.1007/s10796-015-9564-3

[120] Tom Lichtenstein, Dorina Bano, and Mathias Weske. 2022. Attribute-Driven Case Notion Discovery for Unlabeled Event Logs. In *Business Process Management Workshops*, Andrea Marrella and Barbara Weber (Eds.). Springer International Publishing, Cham, 111–122. https://doi.org/10.1007/978-3-030-94343-1_9

[121] Cong Liu, Yulong Pei, Long Cheng, Qingtian Zeng, and Hua Duan. 2021. Sampling business process event logs using graph-based ranking model. *Concurrency and Computation: Practice and Experience* 33, 5 (2021), e5974. https://doi.org/10.1002/cpe.5974

[122] Cong Liu, Yulong Pei, Qingtian Zeng, Hua Duan, and Feng Zhang. 2020. LogRank+: A Novel Approach to Support Business Process Event Log Sampling. In *Web Information Systems Engineering – WISE 2020 (Lecture Notes in Computer Science)*, Zhisheng Huang, Wouter Beek, Hua Wang, Rui Zhou, and Yanchun Zhang (Eds.). Springer International Publishing, Cham, 417–430. https://doi.org/10.1007/978-3-030-62008-0_29

[123] Jie Liu, Jiuyun Xu, Ruru Zhang, and Stephan Reiff-Marganiec. 2021. A repairing missing activities approach with succession relation for event logs. *Knowledge and Information Systems* 63, 2 (Feb. 2021), 477–495. https://doi.org/10.1007/s10115-020-01524-6

[124] Daniela Loreti, Federico Chesani, Anna Ciampolini, and Paola Mello. 2020. Generating synthetic positive and negative business process traces through abduction. *Knowledge and Information Systems* 62, 2 (Feb. 2020), 813–839. https://doi.org/10.1007/s10115-019-01372-z

[125] Xixi Lu, Dirk Fahland, Frank J. H. M. van den Biggelaar, and Wil M. P. van der Aalst. 2016. Handling Duplicated Tasks in Process Discovery by Refining Event Labels. In *Business Process Management*, Marcello La Rosa, Peter Loos, and Oscar Pastor (Eds.). Springer International Publishing, Cham, 90–107. https://doi.org/10.1007/978-3-319-45348-4_6

[126] Xixi Lu, Dirk Fahland, and Wil M.P. van der Aalst. 2016. Interactively Exploring Logs and Mining Models with Clustering, Filtering, and Relabeling. In *Proceedings of the BPM Demo Track 2016*, Vol. 1789. Rio de Janeiro, Brazil, 44–49.

[127] Felix Mannhardt, Massimiliano de Leoni, and Hajo A. Reijers. 2015. Extending Process Logs with Events from Supplementary Sources. In *Business Process Management Workshops*, Fabiana Fournier and Jan Mendling (Eds.). Springer International Publishing, Cham, 235–247. https://doi.org/10.1007/978-3-319-15895-2_21

[128] Felix Mannhardt, Agnes Koschmider, Nathalie Baracaldo, Matthias Weidlich, and Judith Michael. 2019. Privacy-Preserving Process Mining. *Business & Information Systems Engineering* 61, 5 (Oct. 2019), 595–614. https://doi.org/10.1007/s12599-019-00613-3

[129] Felix Mannhardt, Sobah Abbas Petersen, and Manuel Fradinho Oliveira. 2018. Privacy Challenges for Process Mining in Human-Centered Industrial Environments. In *2018 14th International Conference on Intelligent Environments (IE)*. IEEE, Rome, Italy, 64–71. https://doi.org/10.1109/IE.2018.00017 ISSN: 2472-7571.

[130] Salvatore T. March and Gerald F. Smith. 1995. Design and natural science research on information technology. *Decision Support Systems* 15, 4 (Dec. 1995), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

[131] Heidy M. Marin-Castro and Edgar Tello-Leal. 2021. Event Log Preprocessing for Process Mining: A Review. *Applied Sciences* 11, 22 (Jan. 2021), 10556. https://doi.org/10.3390/app112210556 Publisher: Multidisciplinary Digital Publishing Institute.

[132] Niels Martin. 2019. Using Indoor Location System Data to Enhance the Quality of Healthcare Event Logs: Opportunities and Challenges. In *Business Process Management Workshops*, Florian Daniel, Quan Z. Sheng, and Hamid Motahari (Eds.). Springer International Publishing, Cham, 226–238. https://doi.org/10.1007/978-3-030-11641-5_18

[133] Niels Martin, Greg Van Houdt, and Gert Janssenswillen. 2022. DaQAPO: Supporting flexible and fine-grained event log quality assessment. *Expert Systems with Applications* 191 (April 2022), 116274. https://doi.org/10.1016/j.eswa.2021.116274
[134] Matthew B. Miles, A. Michael Huberman, and Johnny Saldana. 2018. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, London, UK.
[135] Alexey A. Mitsyuk, Ivan S. Shugurov, Anna A. Kalenkova, and Wil M. P. van der Aalst. 2017. Generating event logs for high-level process models. *Simulation Modelling Practice and Theory* 74 (May 2017), 1–16. https://doi.org/10.1016/j.simpat.2017.01.003
[136] Miguel Morales-Sandoval, José A. Molina, Heidy M. Marin-Castro, and Jose Luis Gonzalez-Compean. 2021. Blockchain support for execution, monitoring and discovery of inter-organizational business processes. *PeerJ Computer Science* 7 (Sept. 2021), e731. https://doi.org/10.7717/peerj-cs.731 Publisher: PeerJ Inc..
[137] Niels Mueller-Wickop and Martin Schultz. 2013. ERP Event Log Preprocessing: Timestamps vs. Accounting Logic. In *Design Science at the Intersection of Physical and Virtual Design*, Jan vom Brocke, Riitta Hekkala, Sudha Ram, and Matti Rossi (Eds.). Springer, Berlin, Heidelberg, 105–119. https://doi.org/10.1007/978-3-642-38827-9_8
[138] Roman Mühlberger, Stefan Bachhofner, Claudio Di Ciccio, Luciano García-Bañuelos, and Orlenys López-Pintado. 2019. Extracting Event Logs for Process Mining from Data Stored on the Blockchain. In *Business Process Management Workshops*, Chiara Di Francescomarino, Remco Dijkman, and Uwe Zdun (Eds.). Springer International Publishing, Cham, 690–703. https://doi.org/10.1007/978-3-030-37453-2_55
[139] Joyce Nakatumba, Michael Westergaard, and Wil M. P. van der Aalst. 2012. Generating Event Logs with Workload-Dependent Speeds from Simulation Models. In *Advanced Information Systems Engineering Workshops*, Marko Bajec and Johann Eder (Eds.). Springer, Berlin, Heidelberg, 383–397. https://doi.org/10.1007/978-3-642-31069-0_31
[140] Gleison S. Do Nascimento, Cirano Iochpe, Lucinéia Thom, André C. Kalsing, and Gleison S. do Nascimento. 2013. Discovering Business Processes in Legacy Systems Using Business Rules and Log Mining. In *2013 IEEE 10th International Conference on e-Business Engineering*. IEEE, Coventry, UK, 207–212. https://doi.org/10.1109/ICEBE.2013.31
[141] Dominic A. Neu, Johannes Lahann, and Peter Fettke. 2022. A systematic literature review on state-of-the-art deep learning methods for process prediction. *Artificial Intelligence Review* 55, 2 (Feb. 2022), 801–827. https://doi.org/10.1007/s10462-021-09960-8
[142] Hoang Nguyen, Marlon Dumas, Arthur H. M. ter Hofstede, Marcello La Rosa, and Fabrizio Maria Maggi. 2017. Mining Business Process Stages from Event Logs. In *Advanced Information Systems Engineering*, Eric Dubois and Klaus Pohl (Eds.). Springer International Publishing, Cham, 577–594. https://doi.org/10.1007/978-3-319-59536-8_36
[143] Hoang Thi Cam Nguyen, Suhwan Lee, Jongchan Kim, Jonghyeon Ko, and Marco Comuzzi. 2019. Autoencoders for improving quality of process event logs. *Expert Systems with Applications* 131 (Oct. 2019), 132–147. https://doi.org/10.1016/j.eswa.2019.04.052
[144] Phuong Nguyen, Aleksander Slominski, Vinod Muthusamy, Vatche Ishakian, and Klara Nahrstedt. 2016. Process Trace Clustering: A Heterogeneous Information Network Approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, 279–287. https://doi.org/10.1137/1.9781611974348.32
[145] Erik H. J. Nooijen, Boudewijn F. van Dongen, and Dirk Fahland. 2013. Automatic Discovery of Data-Centric and Artifact-Centric Processes. In *Business Process Management Workshops*, Marcello La Rosa and Pnina Soffer (Eds.). Springer, Berlin, Heidelberg, 316–327. https://doi.org/10.1007/978-3-642-36285-9_36
[146] Ana Pajić and Dragana Bečejski-Vujaklija. 2016. Metamodel of the Artifact-Centric Approach to Event Log Extraction from ERP Systems. *International Journal of Decision Support System Technology (IJDSST)* 8, 2 (April 2016), 18–28. https://doi.org/10.4018/IJDSST.2016040102 Publisher: IGI Global.
[147] Ana Pajić Simović, Slađan Babarogić, Ognjen Pantelić, and Stefan Krstović. 2021. Towards a Domain-Specific Modeling Language for Extracting Event Logs from ERP Systems. *Applied Sciences* 11, 12 (Jan. 2021), 5476. https://doi.org/10.3390/app11125476 Publisher: Multidisciplinary Digital Publishing Institute.
[148] Jisheng Pei, Lijie Wen, Hedong Yang, Jianmin Wang, and Xiaojun Ye. 2021. Estimating Global Completeness of Event Logs: A Comparative Study. *IEEE Transactions on Services Computing* 14, 2 (March 2021), 441–457. https://doi.org/10.1109/TSC.2018.2805912 Conference Name: IEEE Transactions on Services Computing.
[149] Anastasiia Pika, Moe T. Wynn, Stephanus Budiono, Arthur H. M. ter Hofstede, Wil M. P. van der Aalst, and Hajo A. Reijers. 2020. Privacy-Preserving Process Mining in Healthcare. *International Journal of Environmental Research and Public Health* 17, 5 (Jan. 2020), 1612. https://doi.org/10.3390/ijerph17051612 Publisher: Multidisciplinary Digital Publishing Institute.
[150] Sebastian Pospiech, Sven Mielke, Robert Mertens, Martin Pelke, Kavitha Jagannath, and Michael Stadler. 2014. Exploration and Analysis of Undocumented Processes Using Heterogeneous and Unstructured Business Data. In *2014 IEEE International Conference on Semantic Computing*. IEEE, Newport Beach, CA, USA, 191–198. https://doi.org/10.1109/ICSC.2014.24
[151] Ricardo Pérez-Castillo, Barbara Weber, Ignacio García-Rodríguez de Guzmán, Mario Piattini, and Jakob Pinggera. 2014. Assessing event correlation in non-process-aware information systems. *Software & Systems Modeling* 13, 3 (July 2014), 1117–1139. https://doi.org/10.1007/s10270-012-0285-5
[152] Ricardo Pérez-Castillo, Barbara Weber, Ignacio García-Rodríguez, and Mario Piattini. 2012. Improving Event Correlation for Non-process Aware Information Systems. In *Proceedings of the 7th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: ENASE*, Vol. 2. SCITEPRESS, Wrocław, Poland, 33–42. https://doi.org/10.5220/0003983100330042

[153] Ricardo Pérez-Castillo, Barbara Weber, Ignacio García-Rodríguez de Guzmán, and Mario Piattini. 2011. Toward Obtaining Event Logs from Legacy Code. In *Business Process Management Workshops*, Michael zur Muehlen and Jianwen Su (Eds.). Springer, Berlin, Heidelberg, 201–207. https://doi.org/10.1007/978-3-642-20511-8_18

[154] R. Pérez-Castillo, B. Weber, I. G.-R. de Guzmán, and M. Piattini. 2011. Process mining through dynamic analysis for modernising legacy systems. *IET Software* 5, 3 (June 2011), 304–319. https://doi.org/10.1049/iet-sen.2010.0103 Publisher: IET Digital Library.

[155] Ricardo Pérez-Castillo, Barbara Weber, Jakob Pinggera, Stefan Zugal, Ignacio García-Rodríguez de Guzmán, and Mario Piattini. 2011. Generating event logs from non-process-aware systems enabling business process mining. *Enterprise Information Systems* 5, 3 (Aug. 2011), 301–335. https://doi.org/10.1080/17517575.2011.587545

[156] Majid Rafiei, Alexander Schnitzler, and Wil M. P. van der Aalst. 2021. PC4PM: A Tool for Privacy/Confidentiality Preservation in Process Mining. https://doi.org/10.48550/arXiv.2107.14499

[157] Majid Rafiei and Wil M. P. van der Aalst. 2019. Mining Roles from Event Logs While Preserving Privacy. In *Business Process Management Workshops*, Chiara Di Francescomarino, Remco Dijkman, and Uwe Zdun (Eds.). Springer International Publishing, Cham, 676–689. https://doi.org/10.1007/978-3-030-37453-2_54

[158] Majid Rafiei and Wil M. P. van der Aalst. 2020. Privacy-Preserving Data Publishing in Process Mining. In *Business Process Management Forum*, Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas (Eds.). Springer International Publishing, Cham, 122–138. https://doi.org/10.1007/978-3-030-58638-6_8

[159] Majid Rafiei and Wil M. P. van der Aalst. 2021. Privacy-Preserving Continuous Event Data Publishing. In *Business Process Management Forum*, Artem Polyvyanyy, Moe Thandar Wynn, Amy Van Looy, and Manfred Reichert (Eds.). Springer International Publishing, Cham, 178–194. https://doi.org/10.1007/978-3-030-85440-9_11

[160] Majid Rafiei and Wil M. P. van der Aalst. 2021. Towards Quantifying Privacy in Process Mining. In *Process Mining Workshops (Lecture Notes in Business Information Processing)*, Sander Leemans and Henrik Leopold (Eds.). Springer International Publishing, Cham, 385–397. https://doi.org/10.1007/978-3-030-72693-5_29

[161] Majid Rafiei, Leopold von Waldthausen, and Wil M.P. van der Aalst. 2018. Ensuring Confidentiality in Process Mining. In *Proceedings of the 8th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2018)*, Vol. 2270. Seville, Spain, 3–17.

[162] Majid Rafiei, Leopold von Waldthausen, and Wil M. P. van der Aalst. 2020. Supporting Confidentiality in Process Mining Using Abstraction and Encryption. In *Data-Driven Process Discovery and Analysis*, Paolo Ceravolo, Maurice van Keulen, and María Teresa Gómez-López (Eds.). Springer International Publishing, Cham, 101–123. https://doi.org/10.1007/978-3-030-46633-6_6

[163] Lihi Raichelson, Pnina Soffer, and Eric Verbeek. 2017. Merging event logs: Combining granularity levels for process flow analysis. *Information Systems* 71 (Nov. 2017), 211–227. https://doi.org/10.1016/j.is.2017.08.010

[164] Belén Ramos-Gutiérrez, Luisa Parody, and María Teresa Gómez-López. 2020. Towards the Detection of Promising Processes by Analysing the Relational Data. In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, Ladjel Bellatreche, Mária Bieliková, Omar Boussaïd, Barbara Catania, Jérôme Darmont, Elena Demidova, Fabien Duchateau, Mark Hall, Tanja Merčun, Boris Novikov, Christos Papatheodorou, Thomas Risse, Oscar Romero, Lucile Sautot, Guilaine Talens, Robert Wrembel, and Maja Žumer (Eds.). Springer International Publishing, Cham, 283–295. https://doi.org/10.1007/978-3-030-55814-7_24

[165] Belén Ramos-Gutiérrez, Ángel Jesús Varela-Vaca, F. Javier Ortega, María Teresa Gómez-López, and Moe Thandar Wynn. 2021. A NLP-Oriented Methodology to Enhance Event Log Quality. In *Enterprise, Business-Process and Information Systems Modeling*, Adriano Augusto, Asif Gill, Selmin Nurcan, Iris Reinhartz-Berger, Rainer Schmidt, and Jelena Zdravkovic (Eds.). Springer International Publishing, Cham, 19–35. https://doi.org/10.1007/978-3-030-79186-5_2

[166] Adrian Rebmann and Han van der Aa. 2021. Extracting Semantic Process Information from the Natural Language in Event Logs. In *Advanced Information Systems Engineering*, Marcello La Rosa, Shazia Sadiq, and Ernest Teniente (Eds.). Springer International Publishing, Cham, 57–74. https://doi.org/10.1007/978-3-030-79382-1_4

[167] Jan Recker. 2021. Research Methods. In *Scientific Research in Information Systems: A Beginner's Guide*, Jan Recker (Ed.). Springer International Publishing, Cham, 87–160. https://doi.org/10.1007/978-3-030-85436-2_5

[168] Jan Recker and Jan Mendling. 2016. The State of the Art of Business Process Management Research as Published in the BPM Conference. *Business & Information Systems Engineering* 58, 1 (Feb. 2016), 55–72. https://doi.org/10.1007/s12599-015-0411-3

[169] Simon Remy, Luise Pufahl, Jan Philipp Sachs, Erwin Böttinger, and Mathias Weske. 2020. Event Log Generation in a Health System: A Case Study. In *Business Process Management*, Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas (Eds.). Springer International Publishing, Cham, 505–522. https://doi.org/10.1007/978-3-030-58666-9_29

[170] Andreas Rogge-Solti, Ronny S. Mans, Wil M. P. van der Aalst, and Mathias Weske. 2013. Repairing Event Logs Using Timed Process Models. In *On the Move to Meaningful Internet Systems: OTM 2013 Workshops*, Yan Tang Demey and Hervé Panetto (Eds.). Springer, Berlin, Heidelberg, 705–708. https://doi.org/10.1007/978-3-642-41033-8_89

[171] Jed Rubenfeld. 1989. The Right of Privacy. *Harvard Law Review* 102, 4 (1989), 737–807. https://doi.org/10.2307/1341305 Publisher: The Harvard Law Review Association.

[172] Sebastian Saavedra, José Llatas, and Jimmy Armas-Aguirre. 2021. Process Mining Model to Guarantee the Privacy of Personal Data in the Healthcare Sector. In *Proceedings of the International Congress on Educational and Technology in Sciences*, Vol. 3037. Chiclayo, Peru,

34–43.
[173] Sareh Sadeghianasl, Arthur H.M ter Hofstede, Suriadi Suriadi, and Selen Turkay. 2020. Collaborative and Interactive Detection and Repair of Activity Labels in Process Event Logs. In *2020 2nd International Conference on Process Mining (ICPM)*. IEEE, Padua, Italy, 41–48. https://doi.org/10.1109/ICPM49681.2020.00017
[174] Sareh Sadeghianasl, Arthur H. M. ter Hofstede, Moe T. Wynn, and Suriadi Suriadi. 2019. A Contextual Approach to Detecting Synonymous and Polluted Activity Labels in Process Event Logs. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, Hervé Panetto, Christophe Debruyne, Martin Hepp, Dave Lewis, Claudio Agostino Ardagna, and Robert Meersman (Eds.). Springer International Publishing, Cham, 76–94. https://doi.org/10.1007/978-3-030-33246-4_5
[175] Cátia M. Salgado, Carlos Azevedo, Hugo Proença, and Susana M. Vieira. 2016. Noise Versus Outliers. In *Secondary Analysis of Electronic Health Records*, MIT Critical Data (Ed.). Springer International Publishing, Cham, 163–183. https://doi.org/10.1007/978-3-319-43742-2_14
[176] Stefan Schönig, Claudio Di Ciccio, and Jan Mendling. 2019. Configuring SQL-based process mining for performance and storage optimisation. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19)*. Association for Computing Machinery, New York, NY, USA, 94–97. https://doi.org/10.1145/3297280.3297532
[177] William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *The Public Opinion Quarterly* 19, 3 (1955), 321–325. https://www.jstor.org/stable/2746450 Publisher: Oxford University Press, American Association for Public Opinion Research.
[178] Ronny Seiger, Francesca Zerbato, Andrea Burattin, Luciano García-Bañuelos, and Barbara Weber. 2020. Towards IoT-driven Process Event Log Generation for Conformance Checking in Smart Factories. In *2020 IEEE 24th International Enterprise Distributed Object Computing Workshop (EDOCW)*. IEEE, Eindhoven, Netherlands, 20–26. https://doi.org/10.1109/EDOCW49879.2020.00016 ISSN: 2325-6605.
[179] Arik Senderovich, Andreas Rogge-Solti, Avigdor Gal, Jan Mendling, and Avishai Mandelbaum. 2016. The ROAD from Sensor Data to Process Instances via Interaction Mining. In *Advanced Information Systems Engineering*, Selmin Nurcan, Pnina Soffer, Marko Bajec, and Johann Eder (Eds.). Springer International Publishing, Cham, 257–273. https://doi.org/10.1007/978-3-319-39696-5_16
[180] Sergey A. Shershakov. 2021. Multi-perspective Process Mining with Embedding Configurations into DB-Based Event Logs. In *Tools and Methods of Program Analysis*, Anna Kalenkova, Jose A. Lozano, and Rostislav Yavorskiy (Eds.). Springer International Publishing, Cham, 68–80. https://doi.org/10.1007/978-3-030-71472-7_5
[181] Sunghyun Sim, Hyerim Bae, and Ling Liu. 2021. Bagging Recurrent Event Imputation for Repair of Imperfect Event Log with Missing Categorical Events. *IEEE Transactions on Services Computing* 16, 1 (2021), 1–1. https://doi.org/10.1109/TSC.2021.3118381 Conference Name: IEEE Transactions on Services Computing.
[182] Mai Skjott Linneberg and Steffen Korsgaard. 2019. Coding qualitative data: a synthesis guiding the novice. *Qualitative Research Journal* 19, 3 (Jan. 2019), 259–270. https://doi.org/10.1108/QRJ-12-2018-0012 Publisher: Emerald Publishing Limited.
[183] V. Skydanienko, C. Di Francescomarino, C. Ghidini, and Fabrizio Maria Maggi. 2018. A Tool for Generating Event Logs from Multi-Perspective Declare Models. In *Proceedings of the Dissertation Award, Demonstration, and Industrial Track at BPM 2018*, Vol. 2196. CEUR Workshop Proceedings, 111–115. https://bia.unibz.it/esploro/outputs/991006186692701241 ISSN: 1613-0073.
[184] Wei Song, Xiaoxu Xia, Hans-Arno Jacobsen, Pengcheng Zhang, and Hao Hu. 2015. Heuristic Recovery of Missing Events in Process Logs. In *2015 IEEE International Conference on Web Services*. IEEE, New York, NY, USA, 105–112. https://doi.org/10.1109/ICWS.2015.24
[185] Wei Song, Xiaoxu Xia, Hans-Arno Jacobsen, Pengcheng Zhang, and Hao Hu. 2017. Efficient Alignment Between Event Logs and Process Models. *IEEE Transactions on Services Computing* 10, 1 (Jan. 2017), 136–149. https://doi.org/10.1109/TSC.2016.2601094 Conference Name: IEEE Transactions on Services Computing.
[186] Darko Stefanovic, Dusanka Dakic, Branislav Stevanov, and Teodora Lolic. 2020. Process Mining in Manufacturing: Goals, Techniques and Applications. In *Advances in Production Management Systems. The Path to Digital Transformation and Innovation of Production Management Systems (IFIP Advances in Information and Communication Technology)*, Bojan Lalic, Vidosav Majstorovic, Ugljesa Marjanovic, Gregor von Cieminski, and David Romero (Eds.). Springer International Publishing, Cham, 54–62. https://doi.org/10.1007/978-3-030-57993-7_7
[187] Vinicius Stein Dani, Henrik Leopold, Jan Martijn E. M. van der Werf, Xixi Lu, Iris Beerepoot, Jelmer J. Koorn, and Hajo A. Reijers. 2022. Towards Understanding the Role of the Human in Event Log Extraction. In *Business Process Management Workshops*, Andrea Marrella and Barbara Weber (Eds.). Springer International Publishing, Cham, 86–98. https://doi.org/10.1007/978-3-030-94343-1_7
[188] Xiaoxiao Sun, Wenjie Hou, Dongjin Yu, Jiaojiao Wang, and Jianliang Pan. 2019. Filtering Out Noise Logs for Process Modelling Based on Event Dependency. In *2019 IEEE International Conference on Web Services (ICWS)*. IEEE, Milan, Italy, 388–392. https://doi.org/10.1109/ICWS.2019.00069
[189] Yaguang Sun, Lyth AI-Khazrage, and Ömer Özümerzifon. 2021. Generating High Quality Samples of Process Cases in Internal Audit. In *Business Process Management Forum*, Artem Polyvyanyy, Moe Thandar Wynn, Amy Van Looy, and Manfred Reichert (Eds.). Springer International Publishing, Cham, 263–279. https://doi.org/10.1007/978-3-030-85440-9_16
[190] Yaguang Sun and Bernhard Bauer. 2015. A Novel Top-Down Approach for Clustering Traces. In *Advanced Information Systems Engineering*, Jelena Zdravkovic, Marite Kirikova, and Paul Johannesson (Eds.). Springer International Publishing, Cham, 331–345.

https://doi.org/10.1007/978-3-319-19069-3_21

[191] Yaguang Sun and Bernhard Bauer. 2016. A Graph and Trace Clustering-based Approach for Abstracting Mined Business Process Models. In *Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016)*. SCITEPRESS - Science and Technology Publications, Lda, Setubal, PRT, 63–74. https://doi.org/10.5220/0005833900630074

[192] S. Suriadi, R. Andrews, A. H. M. ter Hofstede, and M. T. Wynn. 2017. Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information Systems* 64 (March 2017), 132–150. https://doi.org/10.1016/j.is.2016.07.011

[193] Suriadi Suriadi, Chun Ouyang, Wil M. P. van der Aalst, and Arthur H. M. ter Hofstede. 2013. Root Cause Analysis with Enriched Process Logs. In *Business Process Management Workshops*, Marcello La Rosa and Pnina Soffer (Eds.). Springer, Berlin, Heidelberg, 174–186. https://doi.org/10.1007/978-3-642-36285-9_18

[194] Alifah Syamsiyah, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. 2018. DB-XES: Enabling Process Discovery in the Large. In *Data-Driven Process Discovery and Analysis*, Paolo Ceravolo, Christian Guetl, and Stefanie Rinderle-Ma (Eds.). Springer International Publishing, Cham, 53–77. https://doi.org/10.1007/978-3-319-74161-1_4

[195] Zeeshan Tariq, Darryl Charles, Sally McClean, Ian McChesney, and Paul Taylor. 2021. An Event-Level Clustering Framework for Process Mining Using Common Sequential Rules. In *Emerging Technologies in Computing*, Mahdi H. Miraz, Garfield Southall, Maaruf Ali, Andrew Ware, and Safeeullah Soomro (Eds.). Springer International Publishing, Cham, 147–160. https://doi.org/10.1007/978-3-030-90016-8_10

[196] Niek Tax, Emin Alasgarov, Natalia Sidorova, Reinder Haakma, and Wil M. P. van der Aalst. 2019. Generating time-based label refinements to discover more precise process models. *Journal of Ambient Intelligence and Smart Environments* 11, 2 (Jan. 2019), 165–182. https://doi.org/10.3233/AIS-190519 Publisher: IOS Press.

[197] Niek Tax, Natalia Sidorova, and Wil M. P. van der Aalst. 2019. Discovering more precise process models from event logs by filtering out chaotic activities. *Journal of Intelligent Information Systems* 52, 1 (Feb. 2019), 107–139. https://doi.org/10.1007/s10844-018-0507-6

[198] Ghalia Tello, Gabriele Gianini, Rabeb Mizouni, and Ernesto Damiani. 2019. Machine Learning-Based Framework for Log-Lifting in Business Process Mining Applications. In *Business Process Management*, Thomas Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling (Eds.). Springer International Publishing, Cham, 232–249. https://doi.org/10.1007/978-3-030-26619-6_16

[199] Álvaro Valencia Parra, Belén Ramos Gutiérrez, Ángel Jesús Varela Vaca, María Teresa Gómez López, and Antonio García Bernal. 2019. Enabling Process Mining in Aircraft Manufactures: Extracting Event Logs and Discovering Processes from Complex Data. *17th International Conference on Business Process Management 2019 Industry Forum (2019)* (2019), 166–177. https://idus.us.es/handle/11441/133387

[200] Ruud M.E. van Cruchten. 2019. Data Quality in Process Mining: A Rule-based Approach. In *Proceedings of the ICPM 2019 Doctoral Consortium*. Aachen, Germany, 53–79.

[201] Han van der Aa, Henrik Leopold, and Matthias Weidlich. 2020. Partial order resolution of event logs for process conformance checking. *Decision Support Systems* 136 (Sept. 2020), 113347. https://doi.org/10.1016/j.dss.2020.113347

[202] Han van der Aa, Adrian Rebmann, and Henrik Leopold. 2021. Natural language-based detection of semantic execution anomalies in event logs. *Information Systems* 102 (Dec. 2021), 101824. https://doi.org/10.1016/j.is.2021.101824

[203] Wil van der Aalst. 2012. Process Mining: Overview and Opportunities. *ACM Transactions on Management Information Systems* 3, 2 (July 2012), 7:1–7:17. https://doi.org/10.1145/2229156.2229157

[204] Wil van der Aalst, Arya Adriansyah, Ana Karla Alves de Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter van den Brand, Ronald Brandtjen, Joos Buijs, Andrea Burattin, Josep Carmona, Malu Castellanos, Jan Claes, Jonathan Cook, Nicola Costantini, Francisco Curbera, Ernesto Damiani, Massimiliano de Leoni, Pavlos Delias, Boudewijn F. van Dongen, Marlon Dumas, Schahram Dustdar, Dirk Fahland, Diogo R. Ferreira, Walid Gaaloul, Frank van Geffen, Sukriti Goel, Christian Günther, Antonella Guzzo, Paul Harmon, Arthur ter Hofstede, John Hoogland, Jon Espen Ingvaldsen, Koki Kato, Rudolf Kuhn, Akhil Kumar, Marcello La Rosa, Fabrizio Maggi, Donato Malerba, Ronny S. Mans, Alberto Manuel, Martin McCreesh, Paola Mello, Jan Mendling, Marco Montali, Hamid R. Motahari-Nezhad, Michael zur Muehlen, Jorge Munoz-Gama, Luigi Pontieri, Joel Ribeiro, Anne Rozinat, Hugo Seguel Pérez, Ricardo Seguel Pérez, Marcos Sepúlveda, Jim Sinur, Pnina Soffer, Minseok Song, Alessandro Sperduti, Giovanni Stilo, Casper Stoel, Keith Swenson, Maurizio Talamo, Wei Tan, Chris Turner, Jan Vanthienen, George Varvaressos, Eric Verbeek, Marc Verdonk, Roberto Vigo, Jianmin Wang, Barbara Weber, Matthias Weidlich, Ton Weijters, Lijie Wen, Michael Westergaard, and Moe Wynn. 2012. Process Mining Manifesto. In *Business Process Management Workshops (Lecture Notes in Business Information Processing)*, Florian Daniel, Kamel Barkaoui, and Schahram Dustdar (Eds.). Springer, Berlin, Heidelberg, 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

[205] Wil M.P. van der Aalst. 2011. Process mining: discovering and improving Spaghetti and Lasagna processes. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, Paris, France, 1–7. https://doi.org/10.1109/CIDM.2011.6129461

[206] Wil M.P. van der Aalst. 2013. A general divide and conquer approach for process mining. In *2013 Federated Conference on Computer Science and Information Systems*. IEEE, Krakow, Poland, 1–10. https://ieeexplore.ieee.org/abstract/document/6643968

[207] Wil M.P. van der Aalst. 2016. *Process mining: data science in action* (2 ed.). Springer, Heidelberg, Germany. https://doi.org/10.1007/978-3-662-49851-4

[208] Wil M. P. van der Aalst. 2019. Object-Centric Process Mining: Dealing with Divergence and Convergence in Event Data. In *Software Engineering and Formal Methods (Lecture Notes in Computer Science)*, Peter Csaba Ölveczky and Gwen Salaün (Eds.). Springer

International Publishing, Cham, 3–25. https://doi.org/10.1007/978-3-030-30446-1_1
[209] Wil M. P. van der Aalst. 2022. Process Mining: A 360 Degree Overview. In *Process Mining Handbook*. Springer, Aachen, Germany, 373–401.
[210] Wil M. P. van der Aalst and Alessandro Berti. 2020. Discovering Object-centric Petri Nets. *Fundamenta Informaticae* 175, 1-4 (Jan. 2020), 1–40. https://doi.org/10.3233/FI-2020-1946 Publisher: IOS Press.
[211] W. M. P. van der Aalst and A. J. M. M. Weijters. 2004. Process mining: a research agenda. *Computers in Industry* 53, 3 (April 2004), 231–244. https://doi.org/10.1016/j.compind.2003.10.001
[212] Boudewijn F. Van Dongen and Shiva Shabani. 2015. *Relational XES : data management for process mining.* Technical Report. BPMcenter.org. 8 pages. https://research.tue.nl/en/publications/relational-xes-data-management-for-process-mining
[213] Maikel L. van Eck, Xixi Lu, Sander J. J. Leemans, and Wil M. P. van der Aalst. 2015. PM^2: A Process Mining Project Methodology. In *Advanced Information Systems Engineering (Lecture Notes in Computer Science)*, Jelena Zdravkovic, Marite Kirikova, and Paul Johannesson (Eds.). Springer International Publishing, Cham, 297–313. https://doi.org/10.1007/978-3-319-19069-3_19
[214] Maikel L. van Eck, Natalia Sidorova, and Wil M. P. van der Aalst. 2016. Enabling process mining on sensor data from smart products. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*. IEEE, Grenoble, France, 1–12. https://doi.org/10.1109/RCIS.2016.7549355 ISSN: 2151-1357.
[215] Kees M. van Hee, Zheng Liu, and Natalia Sidorova. 2011. Is my event log complete? — A probabilistic approach to process mining. In *2011 Fifth International Conference on Research Challenges in Information Science*. IEEE, Gosier, France, 1–12. https://doi.org/10.1109/RCIS.2011.6006848 ISSN: 2151-1357.
[216] Sebastiaan J. van Zelst and Yukun Cao. 2020. A Generic Framework for Attribute-Driven Hierarchical Trace Clustering. In *Business Process Management Workshops*, Adela Del Río Ortega, Henrik Leopold, and Flávia Maria Santoro (Eds.). Springer International Publishing, Cham, 308–320. https://doi.org/10.1007/978-3-030-66498-5_23
[217] Sebastiaan J. van Zelst, Mohammadreza Fani Sani, Alireza Ostovar, Raffaele Conforti, and Marcello La Rosa. 2020. Detection and removal of infrequent behavior from event streams of business processes. *Information Systems* 90 (May 2020), 101451. https://doi.org/10.1016/j.is.2019.101451
[218] Sebastiaan J. van Zelst, Felix Mannhardt, Massimiliano de Leoni, and Agnes Koschmider. 2021. Event abstraction in process mining: literature review and taxonomy. *Granular Computing* 6, 3 (July 2021), 719–736. https://doi.org/10.1007/s41066-020-00226-2
[219] Seppe K. L. M. vanden Broucke, Jochen De Weerdt, Bart Baesens, and Jan Vanthienen. 2012. Improved Artificial Negative Event Generation to Enhance Process Event Logs. In *Advanced Information Systems Engineering*, Jolita Ralyté, Xavier Franch, Sjaak Brinkkemper, and Stanislaw Wrycza (Eds.). Springer, Berlin, Heidelberg, 254–269. https://doi.org/10.1007/978-3-642-31095-9_17
[220] Ágnes Vathy-Fogarassy, István Vassányi, and István Kósa. 2022. Multi-level process mining methodology for exploring disease-specific care processes. *Journal of Biomedical Informatics* 125 (Jan. 2022), 103979. https://doi.org/10.1016/j.jbi.2021.103979
[221] H.M.W. Verbeek. 2014. Decomposed process mining with DivideAndConquer: 12th International Conference on Business Process Management, BPM 2014. *BPM Demo Sessions 2014 (co-located with BPM 2014, Eindhoven, The Netherlands, September 20, 2014)* (2014), 86–90. Publisher: CEUR Workshop Proceedings.
[222] H.M.W. Verbeek, Wil M.P. van der Aalst, and Jorge Munoz-Gama. 2017. Divide and Conquer: A Tool Framework for Supporting Decomposed Discovery in Process Mining. *Comput. J.* 60, 11 (Nov. 2017), 1649–1674. https://doi.org/10.1093/comjnl/bxx040
[223] H. M. W. Verbeek, Joos C. A. M. Buijs, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. 2011. XES, XESame, and ProM 6. In *Information Systems Evolution*, Pnina Soffer and Erik Proper (Eds.). Springer, Berlin, Heidelberg, 60–75. https://doi.org/10.1007/978-3-642-17722-4_5
[224] Maxim Vidgof, Djordje Djurica, Saimir Bala, and Jan Mendling. 2020. Cherry-Picking from Spaghetti: Multi-range Filtering of Event Logs. In *Enterprise, Business-Process and Information Systems Modeling (Lecture Notes in Business Information Processing)*, Selmin Nurcan, Iris Reinhartz-Berger, Pnina Soffer, and Jelena Zdravkovic (Eds.). Springer International Publishing, Cham, 135–149. https://doi.org/10.1007/978-3-030-49418-6_9
[225] Maxim Vidgof, Djordje Djurica, Saimir Bala, and Jan Mendling. 2022. Interactive log-delta analysis using multi-range filtering. *Software and Systems Modeling* 21, 3 (June 2022), 847–868. https://doi.org/10.1007/s10270-021-00902-0
[226] M. Vijayakamal and D. Vasumathi. 2020. Unsupervised Learning Methods for Anomaly Detection and Log Quality Improvement Using Process Event Log. *International Journal of Advanced Science and Technology* 29, 1 (2020), 1109–1125.
[227] Michał Walicki and Diogo R. Ferreira. 2011. Sequence partitioning for process mining with unlabeled event logs. *Data & Knowledge Engineering* 70, 10 (Oct. 2011), 821–841. https://doi.org/10.1016/j.datak.2011.05.003
[228] Pan Wang, Wen'an Tan, Anqiong Tang, and Kai Hu. 2018. A Novel Trace Clustering Technique Based on Constrained Trace Alignment. In *Human Centered Computing*, Qiaohong Zu and Bo Hu (Eds.). Springer International Publishing, Cham, 53–63. https://doi.org/10.1007/978-3-319-74521-3_7
[229] Kong Xiangsheng. 2014. RFID Event Analysis Based on Complex Event Processing. *International Journal of Online and Biomedical Engineering (iJOE)* 10, 1 (Feb. 2014), 5–9. https://doi.org/10.3991/ijoe.v10i1.3049

[230] Yu Xiao and Maria Watson. 2019. Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research* 39, 1 (2019), 93–112. https://doi.org/10.1177/0739456X17723971

[231] Jiuyun Xu and Jie Liu. 2019. A Profile Clustering Based Event Logs Repairing Approach for Process Mining. *IEEE Access* 7 (2019), 17872–17881. https://doi.org/10.1109/ACCESS.2019.2894905 Conference Name: IEEE Access.

[232] Yang Xu, Qi Lin, and Martin Q. Zhao. 2016. Merging Event Logs for Process Mining with Hybrid Artificial Immune Algorithm. In *Proceedings of the International Conference on Data Mining (DMIN)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), Athens, United States, 10–16. https://www.proquest.com/docview/1806429015/abstract/683C96E96D0A4807PQ/1

[233] Hedong Yang, Lijie Wen, and Jianmin Wang. 2012. An Approach to Evaluate the Local Completeness of an Event Log. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, Brussels, Belgium, 1164–1169. https://doi.org/10.1109/ICDM.2012.66 ISSN: 2374-8486.

[234] Hedong Yang, Lijie Wen, and Jianmin Wang. 2013. An Approach to Identifying False Traces in Process Event Logs. In *Advances in Knowledge Discovery and Data Mining*, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (Eds.). Springer, Berlin, Heidelberg, 533–545. https://doi.org/10.1007/978-3-642-37456-2_45

[235] Hedong Yang, Lijie Wen, Jianmin Wang, and Raymond K. Wong. 2014. CPL+: An improved approach for evaluating the local completeness of event logs. *Inform. Process. Lett.* 114, 11 (Nov. 2014), 607–610. https://doi.org/10.1016/j.ipl.2014.06.001

[236] Rashid Zaman, Marwan Hassani, and Boudewijn F. Van Dongen. 2021. Prefix Imputation of Orphan Events in Event Stream Processing. *Frontiers in Big Data* 4 (Oct. 2021). https://doi.org/10.3389/fdata.2021.705243 Publisher: Frontiers.

[237] Fareed Zandkarimi, Jana-Rebecca Rehse, Pouya Soudmand, and Hartmut Hoehle. 2020. A Generic Framework for Trace Clustering in Process Mining. In *2020 2nd International Conference on Process Mining (ICPM)*. IEEE, Padua, Italy, 177–184. https://doi.org/10.1109/ICPM49681.2020.00034

[238] Zhenyu Zhang, Ryan Hildebrant, Fatemeh Asgarinejad, Nalini Venkatasubramanian, and Shangping Ren. 2021. Improving Process Discovery Results by Filtering Out Outliers from Event Logs with Hidden Markov Models. In *2021 IEEE 23rd Conference on Business Informatics (CBI)*, Vol. 01. IEEE, Bolzano, Italy, 171–180. https://doi.org/10.1109/CBI52690.2021.00028 ISSN: 2378-1971.

Received 17 May 2023; revised 19 September 2024; accepted 8 January 2025