Process-Data Quality: The True Frontier of Process Mining

ARTHUR H. M. TER HOFSTEDE, Queensland University of Technology, Australia
AGNES KOSCHMIDER, University of Bayreuth, Germany
ANDREA MARRELLA, Sapienza University of Rome, Italy
ROBERT ANDREWS, Queensland University of Technology, Australia
DOMINIK A. FISCHER, University of Bayreuth, Germany
SAREH SADEGHIANASL and MOE THANDAR WYNN, Queensland University of Technology, Australia
MARCO COMUZZI, Ulsan National Institute of Science and Technology, Korea
JOCHEN DE WEERDT, KU Leuven, Belgium
KANIKA GOEL, Queensland University of Technology, Australia
NIELS MARTIN, Hasselt University, Belgium
PNINA SOFFER, University of Haifa, Israel

Since its emergence over two decades ago, process mining has flourished as a discipline, with numerous contributions to its theory, widespread practical applications, and mature support by commercial tooling environments. However, its potential for significant organisational impact is hampered by poor quality event data. Process mining starts with the acquisition and preparation of event data coming from different data sources. These are then transformed into event logs, consisting of process execution traces including multiple events. In real-life scenarios, event logs suffer from significant data quality problems, which must be recognised and effectively resolved for obtaining meaningful insights from process mining analysis. Despite its importance, the topic of data quality in process mining has received limited attention. In this paper, we discuss the emerging challenges related to process-data quality from both a research and practical point of view. Additionally, we present a corresponding research agenda with key research directions.

CCS Concepts: • **Information systems → Data mining;**

Additional Key Words and Phrases: Event data quality, process mining, event log

29

Authors' addresses: A. H. M. ter Hofstede, R. Andrews, S. Sadeghianasl, M. T. Wynn, and K. Goel, Queensland University of Technology, School of Information Systems, Faculty of Science, 2 George St, Brisbane, QLD 4000, Australia; emails: {a.terhofstede, r.andrews, s.sadeghianasl, m.wynn}@qut.edu.au, goelk1988@gmail.com; A. Koschmider and D. A. Fischer, University of Bayreuth, Faculty of Law, Business and Economics, Business Informatics and Process Analytics, Wittelsbacherring 10, DE-95444 Bayreuth, Germany; emails: Agnes.Koschmider@uni-bayreuth.de, dominik.fischer@fim-rc.de; A. Marrella, Sapienza University of Rome, Department of Computer, Control and Management Engineering, Via Ariosto 25, 00185 Rome, Italy; email: marrella@diag.uniroma1.it; M. Comuzzi, Ulsan National Institute of Science and Technology, Department of Industrial Engineering, 50 UNIST-gil, Ulju-gun, Ulsan 44919, Republic of Korea; email: mcomuzzi@unist.ac.kr; J. De Weerdt, KU Leuven, Research Centre on Information Systems Engineering, Naamsestraat 69, 3000 Leuven, Belgium; email: jochen.deweerdt@kuleuven.be; N. Martin, UHasselt - Hasselt University, Research Group Business Informatics, Martelarenlaan 42, 3500 Hasselt, Belgium; email: niels.martin@uhasselt.be; P. Soffer, University of Haifa, Department of Information Systems, Hanamal 65, Haifa 3303221 Israel; email: spnina@is.haifa.ac.il.

1936-1955/2023/09-ART29
https://doi.org/10.1145/3613247

ACM Reference format:
Arthur H. M. ter Hofstede, Agnes Koschmider, Andrea Marrella, Robert Andrews, Dominik A. Fischer, Sareh Sadeghianasl, Moe Thandar Wynn, Marco Comuzzi, Jochen De Weerdt, Kanika Goel, Niels Martin, and Pnina Soffer. 2023. Process-Data Quality: The True Frontier of Process Mining. *ACM J. Data Inform. Quality* 15, 3, Article 29 (September 2023), 21 pages.
https://doi.org/10.1145/3613247

1 INTRODUCTION

In our era of data, great emphasis is placed on increasingly sophisticated techniques for data analysis. While this is undoubtedly important, it is imperative to ensure that the data used as input by these techniques is of sufficient quality, otherwise the maxim *garbage in, garbage out* will rear its ugly head. It could be argued that the focus on the development of analysis techniques has overshadowed our attention to managing data quality properly. In this regard, it is for example interesting to observe the relatively late emergence of *data governance*, which according to Ladley – quoting himself in [36] – is "a required business capability if you want to get value from your data". Data quality is considered one of the dimensions of data governance and high-quality data is sometimes defined as data that is "fit for purpose" [59].

In the field of process mining [55], which focuses on deriving process-related insights from event log data, a similar situation is manifesting itself. Process data contains historical events from process executions where each event, in its simplest form, refers to a case, an activity, a point in time, and (optionally) a resource. Process data is different from other data as it has well-defined semantics (e.g., cases and events and their relation), a well-formed structure (e.g., IEEE XES log format), and it is subject to a temporal ordering (e.g., through timestamps). Although process mining emerged as a field of research in the late 1990s, the topic of process-data quality began to receive serious attention only from early 2010 onwards, with the work of people such as J. C. Bose, Ronny Mans, and Wil van der Aalst (e.g., [9]). This is somewhat surprising given the typical substantial problems with process-data quality in real-life event logs and the disproportionate amount of time that is spent on resolving them [61]. While research in process-data quality has made progress in the last decade, it still lags behind other developments in the field, presenting a major obstacle to broader acceptance and application of process mining in practice.

In this paper, we identify the main challenges that need to be addressed in order to increase the maturity of data quality assessment and improvement in process mining. Based on these challenges, we outline an associated research agenda for the field of process-data quality in the form of key directions that should be followed by future research. We argue that realising this agenda will be instrumental in advancing the field of process mining.

The paper is organised as follows. Section 2 provides general background on data quality and process mining, while Section 3 specifically introduces the typical data quality issues in event logs and contextualises data quality detection and repair within a general process mining framework. The challenges are presented in Section 4, the future research directions in Section 5, and conclusions are finally drawn in Section 6.

2 BACKGROUND

2.1 Data Quality and its Dimensions

Data quality is a long-standing and multifaceted concept that has been addressed in different contexts, including statistics [24], management [5], and computer science [60]. In the first decade of the 2000s, data quality has been investigated with a focus on relational data, traditionally adopted in Database Management Systems (DBMSs [7]). The reasons for this were the growing need to inte-

grate information across disparate data sources and the tremendous impact of poor quality data on data integration efforts. More recently, social networks and the Web have made other types of data arising from linguistic and visual information ubiquitous, challenging researchers to investigate how data quality concepts can be modified or extended to fit such data [6], e.g., semi-structured texts, maps, images, linked open data, and so on.

To fully understand and characterize the data quality concept, researchers have identified a number of quality dimensions that capture specific facets of quality. The most commonly referenced dimensions are accuracy, completeness, consistency, and reliability [7]:

- *Accuracy* focuses on the adherence of data to a given reality of interest.
- *Completeness* refers to the capability of representing all relevant aspects of the reality of interest.
- *Consistency* refers to the capability of data to comply with all properties of the reality of interest, as specified in terms of integrity constraints and business rules.
- *Reliability* measures if data can be trusted and used for making informed decisions.

A long list of additional dimensions of data quality can be defined – some examples include appropriateness, credibility, conformity, currency, relevance, and usability [6].

2.2 Process Mining

With the recent developments of the Internet of Things (IoT) and cloud-based technologies, massive amounts of data are generated by heterogeneous sources and stored through dedicated cloud solutions. A broad spectrum of data science techniques is available to derive actionable business insights from the recorded historical data. One such family of data analysis techniques is process mining [55]. It encompasses several sub-disciplines, such as process discovery [4] concerned with the discovery of process models from event data, conformance checking [14] concerned with alignment between logs and models or rules, deviance analysis [42] concerned with identifying why certain process instances perform better or worse than others, performance analysis concerned with metrics such as waiting times and throughput times [62], organisational mining [45, 50, 64] concerned with patterns of collaboration among the (human) resources involved in the execution of a process, concept drift [10] concerned with detecting model or rule changes over time, and predictive process monitoring [39] concerned with creating predictive models of process execution based on historical data.

Process mining concentrates on the actual execution of processes, as reflected by the footprint of reality logged by the information systems of an organisation. The main input of process mining is an *event log*, which is analysed to extract insights and recurrent patterns about how processes are executed. Event logs consist of *traces*. A trace consists of the sequence of *events* logged during the execution of an individual instance of a process, i.e., a process *case*. Irrespective of the type of process mining analysis undertaken, events are related to a particular step in a process. Each event is minimally characterised by a *case identifier*, which informs the case to which the event relates, an *activity label* describing the related action, and a *timestamp* describing when the event occurred. Many types of process mining analysis (e.g., organisational mining) require that the log contains relevant supporting attributes. For instance, it is only possible to discover the social network of resources contributing to the process if event data is enriched with resource information. Figure 1 shows a hospital log fragment comprising patient visits, treatment, and personal details.

To enable the exchange of event logs between different information systems, the process mining community has developed an interchange standard that defines the structure and general contents of event logs. Since 2016, the official IEEE standard for storing and analysing event logs is XES[1]

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

CaseID	Attributes			
	Activity	Timestamp	Resource	Location
1	Present at ED	2022-10-03 07:54	Jason	Emergency
1	Triage request	2022-10-03 07:57	Jason	Emergency
1	Triage	2022-10-03 08:03	Susan	Emergency
1	Medical assign	2022-10-03 08:10	Jason	Emergency
1	Blood tests	2022-10-03 09:34	Sarah	Laboratory
1	Admit to hospital	2022-10-03 10:02	George	Ward
2	Present at ED	2022-10-03 08:12	Jason	Emergency
2	Triage request	2022-10-03 08:16	Jason	Emergency
2	Triage	2022-10-03 08:20	Ross	Emergency
2	Medical assign	2022-10-03 08:30	Jason	Emergency
2	Discharge home	2022-10-03 08:50	Sarah	Emergency

Fig. 1. Fragment of a hospital event log.

(eXtensible Event Stream). But log formats evolve over time when new insights emerge. A notable example is the recent object-centric paradigm for event logs [56].

3 DATA QUALITY IN PROCESS MINING

Section 3.1 presents the typical data quality issues in event logs, while Section 3.2 introduces a general process mining framework within which we position the research challenges that are discussed in Section 4.

3.1 Data Quality Issues in Event Logs

Process mining techniques crucially rely on historical data as the single source of truth. Input data of low and dubious quality poses significant hurdles not only to successfully translating historical data into business value, but even simply to applying process mining techniques in the first place. Early work in the area of process-data quality raised awareness of this issue, introducing various levels of quality of event logs [54] and manifestations of typical quality problems [9]. The latter work highlighted problems specific to process mining, but only at a fairly high level of abstraction. The chronic and profound data quality problems of event logs encountered in real-life situations resulted in a more systematic and in-depth approach to the detection and repair of data quality problems in event logs through the introduction of the so-called *event log imperfection patterns* [51].

Patterns have proven to be a powerful mechanism to address problems that are not well-defined. For example, the workflow patterns [57] provided an indication of the kind of control-flow dependencies one may need to capture when specifying workflows. This occurred at a time when there was no consensus around these needs and there was a surfeit of languages and tools. Patterns provide an abstraction that is independent of any technological solution and they introduce a terminology that facilitates discussion around problems and their solutions. Patterns have also proven to be an excellent starting point for tool development. The event log imperfection patterns exemplify the aforementioned benefits. They consist of 11 patterns that each name and define a specific data quality problem that can be observed in event data. Ways of detecting these problems are captured as well as possible repair mechanisms and potential side effects of these repairs. Pattern collections can never be argued to be complete, but they provide a framework that makes addition over time of new patterns, based on new insights, easier.

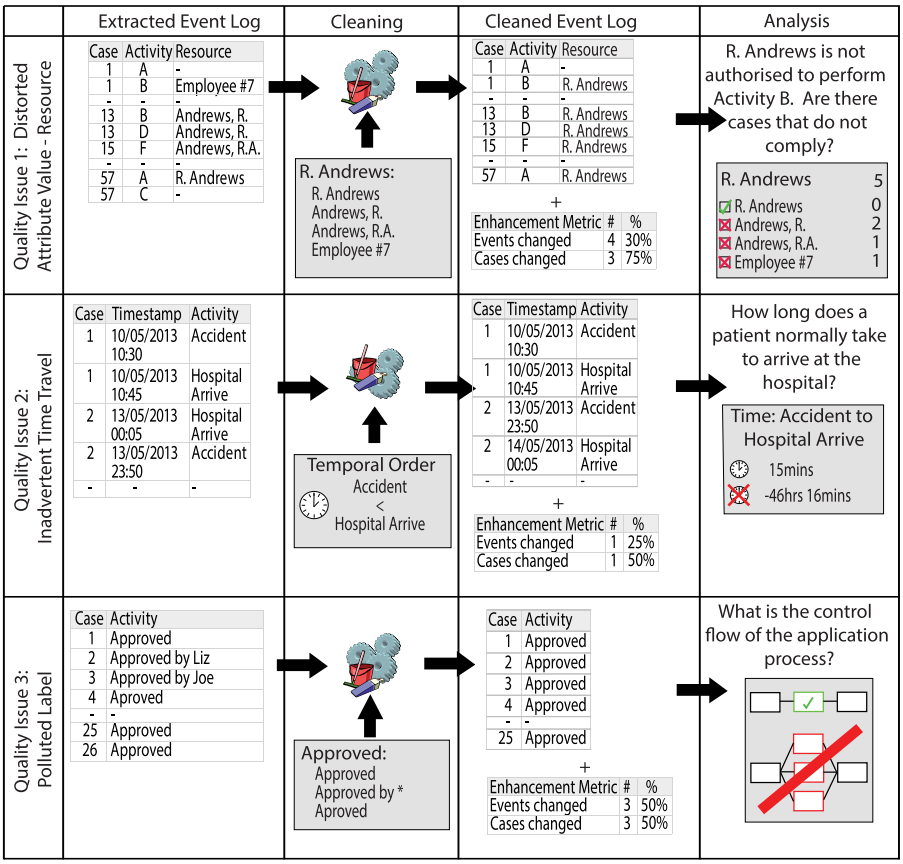

Fig. 2. An illustrated example of how data quality issues can be detected and repaired prior to the analysis phase to increase the accuracy of insights.

Figure 2 depicts three examples of how the occurrence of a particular event log imperfection pattern can be detected and (automatically) cleaned such that the final analysis is of high(er) quality. In the first example, we can detect the issue of a *distorted* label where a resource (i.e., Andrews, R) is referred to in several ways in an event log. Using the *cleaned* event log (i.e., the resource R. Andrews is unified) instead of the original event log, the analysis can correctly answer a compliance related question about this resource. In this example, the analysis detected that R.Andrews is not authorised to perform Activity B. In the second example, we depict a well-known example of an *inadvertent time travel* scenario where a wrong temporal order was recorded in an event log. In this case, the log seems to indicate that case #2 arrives at the hospital before s/he is in an accident. Again, without first detecting and cleaning such temporal issues in a log, the time-based performance analysis will be inaccurate. The third example shows a case of *polluted* activity labels in an event log. Without cleaning the log first, the process mining analysis will discover an *incorrect* process model (the bottom one) rather than the top one. All these three examples combined highlight the diversity of data quality issues and the importance of detecting and cleaning data quality issues to ensure accurate process mining insights.

Recent research has highlighted that process-data quality problems are best understood in the context of three worlds and the interactions among them [22]: the *personal* world (e.g.,

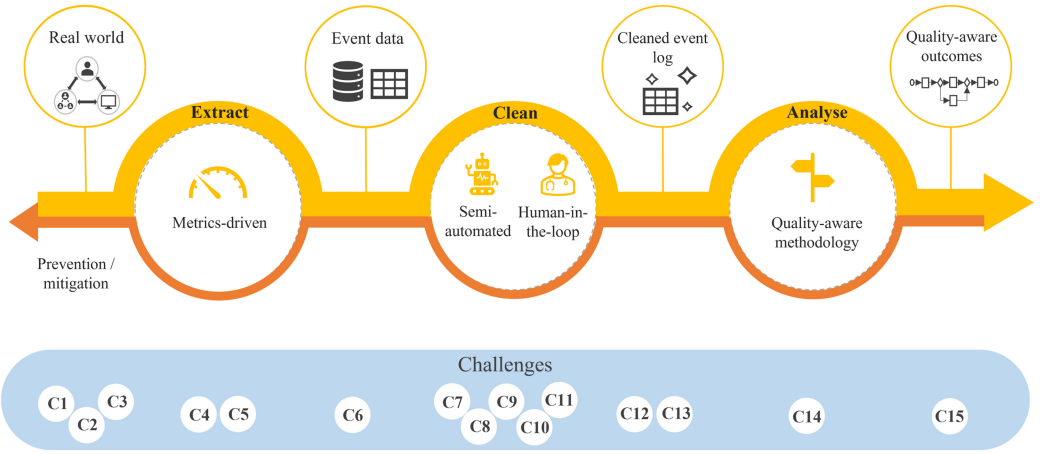

Fig. 3. A view on future treatment of data quality problems in process mining.

capturing people's habits and idiosyncrasies), the *social* world (e.g., capturing organisational rules and conventions), and the *material* world (e.g., capturing systems and their interfaces and configurations). For example, an organisational incentive (social world) may cause people to take spurious action – e.g., imagine that a premium is placed on client interaction – or a free-text field in an interface (material world) may result in a wide range of names for the same action. It has been documented how specific issues in, and interactions between, the three worlds may result in specific event log imperfections [2]. This highlights how important it is to have a deep understanding of the business context in which process mining takes place. It also provides a first insight into how data quality problems can be prevented or mitigated in the first place.

3.2 A General Process Mining Framework

There exist several process mining methodologies to guide the execution of process mining projects (e.g., the staged approach described in [55] or the PM^2 methodology [58]). Any process mining approach, at its core, typically involves a process-data *extraction* and a *cleaning* phase, before an *analysis* can be conducted, and the results reported to the stakeholders. The event log extraction and cleaning phases have not yet been integrated or are not yet well-supported by existing commercial process mining tools. The success of these two activities is therefore largely determined by the experience of the analysts involved. It is well-known that the process data extraction and cleaning tasks (a.k.a. pre-processing) take a significant amount of time and their effectiveness has a significant impact on the ultimate outcome of process mining analysis. A recent survey of 289 process mining users across the four roles of practitioners, researchers, software vendors, and end-users further highlights the need for a systematic approach to process data pre-processing [61]. The data pre-processing task is recognised to be one of the most time-consuming aspects of a process mining project, with many projects spending 60-80% of their efforts on this task, while some up to 90%.

Figure 3 provides a view on the treatment of data quality problems incorporating the various stages mentioned earlier. The challenges of the subsequent section are organized in terms of this treatment. Extraction of data is ideally driven by metrics (see, e.g., Andrews et al. [3] or Ziolkowski et al. [65]) and its cleaning guided and performed automatically as much as possible. The analysis should be methodologically guided and the artefacts created should be "quality-aware", i.e.,

their dependence on cleaning operations should be clear and how certain we can be about them. Throughout the process, data quality problems are traced back to the organisation to help with their prevention and mitigation.

4 CHALLENGES

In this section, we will postulate a number of challenges that need to be addressed to advance the area of process-data quality. They will give rise to a number of future directions for the field (discussed in the subsequent section).

Challenge 1. How do we methodically assess that the right data is being collected for the organisational needs? □

The collection of the 'right' data depends on the objectives of the analysis. Thus, it has to be assessed if the data is right to find the answers needed for the objectives one has in mind within the time frame. Considering the objectives, the choice of data depends on three central questions: (1) What is the goal or the purpose of the analysis? (2) What types of data are planned to be gathered and (3) What methods and procedures will be used to collect, store, and process the information? However, as highlighted by [53], it may not be known what data is relevant for answering a certain question until after some initial data analysis has been conducted. It turns out that data analysis can be needed for determining what data should be gathered, and relying on partial data that has been gathered may miss relevant information. For example, the delay in delivery may be associated with the shipping company involved, but if this data has not been included in an event log (e.g., in the Order Management Log, available at https://ocel-standard.org/) it cannot be revealed. Facing this challenge, it has been suggested that multiple data sources can be used in combination with the event log (e.g., database records, database transaction log [53] or as a direct replacement of the log [19]). Additionally, objectives can change over time. This means that the initial objective of the analysis could change and other objectives could become more relevant. For example, an initial objective may be concerned with finding bottlenecks in a process, but it could also be interesting to make predictions and forecasts during the analysis. In this sense, this challenge would require flexibility in adding or removing information sources that complement the information available in a log, and the methods for assessing the relevance of features (or attributes) for a given analysis task.

Considering the time frame, since processes can change over time, it might be useful to identify concept drifts [1, 10]. Then, depending on the objective of the analysis, irrelevant data that concerns the obsolete behavior, can be removed in order to get (more) meaningful results.

Challenge 2. How can process-data quality problems be prevented or, if that is not possible, mitigated? □

It seems evident that it is preferable to prevent process-data quality problems from occurring in the first place than detecting and repairing them subsequently. If these cannot be fully prevented altogether, then at least it would be desirable to mitigate them as much as possible. In the context of collecting process data to monitor process performance indicators, Cappiello et al. [13] distinguish between improving the quality of event logs already collected and modifying the process execution systems to improve the quality of process data collected in the future. Regarding the latter, some modifications may involve only individual systems, e.g., updating the configuration of one system to log the data currently missing in a log, or of multiple systems, e.g., resolving timestamp inconsistencies. However, mitigating process-data quality issues goes beyond updated system configurations.

In order for prevention and mitigation to be effective, the root causes of process-data quality problems need to be understood. As mentioned already in Section 3, the Odigos framework [22] shows how three worlds and their interactions – the *social world* (e.g., organisational policies), the *material world* (e.g., computer interfaces), and the *personal world* (e.g., behavioural traits of individuals) – can help explain why process-data quality problems emerge. The work has been elaborated and initial links have been made between the framework and the event log imperfection patterns [2]. This is a first attempt at a structured approach to root-cause analysis, but more work is needed to elaborate on possible root causes for the various patterns and for emerging patterns. The personal world and its relation to data quality in particular is ill-understood and it would be interesting to investigate how soft factors such as personal traits, external pressures (e.g., time), and group dynamics can influence process-data quality.

Challenge 3. How can prevention and mitigation solutions be assessed in terms of their cost to the organisation(s) involved? □

Preventing and mitigating process-data quality problems can be costly. For example, it may involve changing computer interfaces, guiding people's behaviour, or improving organisational best practices. Organisational resources are not infinite and choices may need to be made as to what measures to take to improve process-data quality. In order to do this optimally, it needs to be understood how to assess the cost of various prevention and mitigation actions. This cost may be financial (e.g., the implementation of a new computer system or the reconfiguration of an existing one), but can also be psychological (e.g., lower morale as certain freedoms around data entry and reporting have been taken away) or reputational (e.g., improved quality measures may lead to increased trust by customers). To understand the full implications of a set of process-data quality prevention and mitigation measures a cost model needs to be developed that can take both quantitative and qualitative input into account.

By improving the quality of the insights obtained from process mining, process-data quality improvement actions also yield benefits to organisations. Depending on the domain, type of organisation, and relevant process mining use cases, certain improvement actions may be more beneficial than others from an economic standpoint [32]. For instance, an action targeted at maximising the accuracy of event timestamps may be optimal for bottleneck analysis, but it may be too costly if process mining is used only for the discovery of high-level process maps. A cost model of process-data quality improvement actions, therefore, may be extended into a cost-benefit model for identifying optimal process-data improvement scenarios.

Challenge 4. How can data quality problems be minimised when extracting information from a potentially diverse and heterogeneous collection of data sources? □

As data can come from multiple sources with some sources more reliable than others – e.g., patient data collected at the emergency department may be less reliable than patient data in medicare payment systems – it is essential to understand the origin of data quality problems. When multiple attributes with similar concepts are available from different data sources, it is necessary to assess the overall quality of an attribute before deciding to include it in the generation of an event log. For instance, Andrews et al. [3] propose a technique to assess the suitability of various columns in relational tables for inclusion in an event log. It may also be necessary to undertake sanity checking of data values by cross-checking values from one source against the other. For example, a patient's gender can be recorded as male in one data source and as female in another data source. Furthermore, it is possible to introduce new data quality challenges during the

curation/transformation process - for example due to human errors (e.g., incorrect conversion of date fields, incorrect duplication of attributes).

To improve the quality of input data, there is a need to keep track of where the data originated from and how the data has been transformed. Several approaches exist to capture the provenance of event log data. For instance, data quality annotations might be used that map event logs with their current data quality, as described in [29]. In [29], data quality annotations at event, trace, and log levels are proposed to track the data quality issues found in an event log and also to record the extent of the repairs applied to the event log. Such metadata about data quality can assist in undertaking quality-aware process mining. Another solution could be to rank event log attributes according to their sensitivity from most to least impacted by quality issues and fix the issues based on their relative importance.

Challenge 5. How can we monitor the occurrence of data-quality problems as new event log data is being generated? How can we quickly detect the manifestation of a new type of problem or the recurrence of well-known issues? □

Data problems can occur in different phases, e.g., during extraction or repair (see Figure 3) and data monitoring aims to prevent data problems. Data monitoring checks the event log data against quality rules, i.e., if the data still meets the quality. Data monitoring depends on the data quality metrics for accuracy, completeness or how timely the data is. Data visualisation techniques also allow to explore data in order to identify data that is incorrect, incomplete or irrelevant as event log data. Appropriate metrics must be defined identifying changes in the data in time. How do we know that the data cleaning pipeline works appropriately? How can we check that the right event data is getting through? How do we keep track of event data uptime? How can we continually validate the data cleaning pipeline?

Challenge 6. How is data provenance best addressed? In other words, how can we keep track of the origins of data and how it has been transformed over time? □

Data provenance tells us how data was sourced and how it was transformed over time. In other words, it provides us with the DNA of data [12]. It is imperative that we capture this information in order to understand how a questionable data source may have affected data or how repairs may have changed the nature of potential data quality issues [11]. A provenance scheme is needed that records data sources and transformations to which data has been subjected. Such a scheme needs to be conceptually clear, so that we may easily understand the various transformations, and somewhat space efficient given the potentially large volumes of data affected. One way of capturing provenance information is by using annotations. Data quality and data transformation annotations have been considered by Goel et al. [29]. The authors demonstrate how consideration of annotations in automated techniques can result in reliable insights. These may be considered an alternative to recording full provenance and thus sacrifice more detailed insights into the history of quality issues.

Challenge 7. Can we provide a precise characterisation of the typical process-data quality problems that may occur in event logs? □

Process-data quality problems may take a variety of forms, but it certainly seems that a number of them are frequently occurring and have particular relevance. Identifying and characterising these issues and implementing appropriate detection and repair approaches can greatly enhance the pre-processing of event logs. In particular, the use of event log imperfection patterns [51] is a feasible and desirable approach to capturing process-data quality problems.

Although the event log imperfection patterns have provided valuable insights into common process-data quality problems, they are not exhaustive. New types of event data, such as sensor data, may reveal additional patterns, as recently shown in [8], a study describing six specific patterns of poor sensor data quality leading to event log data quality issues. Similarly, other sources of event data, such as unstructured text and images, may offer opportunities to identify novel patterns. Furthermore, changes in event log formats, such as the object-centric approach [56], may also give rise to new patterns. Moreover, domain-specific data quality patterns may exist, which may require context-aware repair algorithms that rely on business logic rules. Recent research on data imperfection patterns in digital health systems has demonstrated this, where six common data imperfection patterns and their root causes were described [30]. To develop appropriate prevention and mitigation strategies, it is crucial to first characterise typical data quality problems at a higher level of abstraction and for specific domains.

Challenge 8. Given that domain expertise may be pivotal in solving certain types of quality problems, how can domain experts be best engaged to help with process-data quality improvement? □

Given the complexity of some domains, it is not realistic to expect that automated approaches for detecting and/or solving certain process-data quality problems will yield fully satisfactory results. Consistent with the "human-in-the-loop" debate in artificial intelligence, it is clear that the involvement of one or more experts can significantly improve the results. Domain experts tend to have extensive contextual knowledge, which for instance enables them to assess whether a potential data quality issue is indeed problematic. This highlights the need for developing novel approaches in which the domain expert interactively detects and/or rectifies process-data quality issues [40, 41]. One approach to involve domain experts is gamification, which has been trialled within the context of activity label correction [47]. But there are several remaining issues to be addressed, such as how to design the right type of game for a given process-data quality problem for a certain domain. Other than gamification and crowdsourcing there may be other ways to systematically and effectively involve domain experts, but such approaches must take their limited time into account as well as, typically, their unfamiliarity with technical aspects of data cleaning and preparation. Hence, future approaches should be accessible for experts with potentially limited data skills and intelligence such that they only request minimal input to algorithmically identify and/or rectify a broader class of data quality issues. Also, approaches should be studied where domain knowledge is injected in machine learning-based data cleaning as suggested in [34]. At what stage should human interaction be sought and to resolve what kinds of issues?

Challenge 9. How can process-data quality problems be detected and repaired, preferably automatically as much as possible? □

Detection and repair of process-data quality problems are essential tasks of data cleaning. Particularly, manually detecting and repairing process-data quality problems can be both time-consuming and error-prone. While human judgement (see Challenge 8) should ideally be used for resolving complex and subjective problems where domain knowledge is necessary, a mixed approach is recommended for resolving process-data quality issues. This also involves automatically detecting and repairing issues that can be identified without requiring human intervention. Various techniques have been identified from different fields such as statistics (e.g., [21]), machine learning (e.g., [1, 34, 43]), data mining (e.g., [48]), automata theory (e.g., [16]), and integer linear programming (e.g., [17, 20]), to deal with process data-quality problems. A link between visual analytics and process mining in terms of the importance of data quality for both was suggested

by Gschwandtner [31], though not elaborated upon. The work by Lu et al. [38], though not specifically focused on data quality, demonstrates the potential of visual techniques for pattern detection in event logs. In short, there are a rich set of techniques from a variety of fields that can be used and adapted for process-data quality detection and repair with a minimum of human involvement. It should be noted that coverage of not only the event log imperfection patterns would be desirable but also emerging patterns. In addition, it cannot be expected that patterns will have perfect solutions that work under all circumstances. Each pattern will require a collection of approaches for detection and repair that may be more or less suitable depending on log characteristics (e.g., time granularity, availability of certain attributes) or availability of other input (e.g., ontologies).

Challenge 10. How do we deal with process-data quality problems that are practically intractable? □

Intractable process-data quality problems refer to intolerable defects in an event log that are inherently unsolvable using traditional data cleaning and pre-processing approaches. Such issues may manifest if the event log results from (manual) data entry activities performed by employees keeping track of process events with (for example) Excel sheets. Employees may accidentally enter data into the wrong field or use different labels to represent the same concept (e.g., activities, resources, etc.) in distinct log traces. Employees may also be distracted and make quick corrections, leading to further errors that render impossible the use of data matching and record linkage technologies [33]. Acknowledging the potential limitations caused by such issues and minimising their impact on the analysis is crucial. Data governance procedures can be implemented to prevent such mistakes from occurring in the first place, including appropriate training sessions for employees. In some cases, validating forms in real-time using specific lists to restrict what employees are allowed to input may be necessary to (partially) address the problem. Another challenge is to factor in at an early stage what limitations such problems will impose on the application of process mining techniques.

Challenge 11. How can process-data quality problems that have been detected be best repaired? And how can these repairs be characterised at a suitable level of abstraction? □

There may be multiple approaches to repair process-data quality problems once they are identified. Which one to choose may depend on a variety of factors and guidance is ideally provided. For example, log or process characteristics may guide the choice for specific repairs. The order of repairs will also need to be considered as repairing one problem may have consequences for other identified problems. Repairs should be characterised at a suitable level of abstraction so that it can be understood which problems they aim to tackle. This will help with understanding the fundamental problems of the event log (which can be beneficial for future root-cause analysis) and may help with future treatment of similar event logs. A proper characterisation of repairs may also help their specification, for instance when specifying provenance meta-data in repaired event logs. It will also facilitate comparing repairs of different logs and, possibly, mining repair information collected from several logs to learn the features of effective repair actions.

Challenge 12. How can we determine that data has become obsolete and no longer needs to be cleansed? □

Generally, the relevance of data may diminish over time. In this way, data becomes less relevant, maybe even become irrelevant or no longer accurate. Obsolete data could also be data that is no longer used by the organisation. The relevance of data depends on what goal or purpose one tries

to achieve. It is possible that the data is not currently relevant but could become relevant again in the future. Outdated data is a problem when making decisions based on that data. The classification of data into obsolete data depends on which questions need to be answered in the present and in the future. That means that data cleaning efforts do not need to be expended on *obsolete* data. The challenge arises as to how to recognise which data should still be taken into account for analysis purposes, which data can be safely ignored, appropriately downplayed or even should be stored for future purposes. Another challenge is that data does not become a data graveyard with repositories of unused data. Solutions to this challenge may require a systematic way of aligning decisions with organisational analytical needs, the design of a taxonomy of data understanding what data is used and needed for which purpose and how to factor in data that has diminished relevance but is not completely irrelevant yet.

Challenge 13. How can we quantify the extent of various data quality problems in an event log? Can metrics be developed that help quantify such problems meaningfully? Can visualisations help pinpoint these problems? And if so, how? □

Once process-data quality problems have been detected, their nature and their extent need to be communicated to analysts and in some cases to stakeholders as well. This communication should serve several purposes: first, to form a basis for cleaning and repairing the data – both assessing the need for such actions and indicating what kind of repair operations are needed and where; second, to assess the feasibility of the intended analysis; and third, to assess possible implications on the reliability of the analysis.

Characterising the extent of data quality problems can be achieved through the definition of appropriate quality dimensions and related metrics. Preliminary work in this space has been done in the area of event log generation [3], but in order to fully cater for the substantial variety of process-data quality problems, additional metrics, relating to the different quality dimensions [63] are needed. These metrics can largely serve to quantify the extent of the problems in an event log. Assessment of the feasibility of the intended analysis may refer to quality metrics attached to certain attributes (e.g., granularity level of timestamps [25]). Assessment of the implications on the reliability of the analysis may lead to probabilistic process mining [44]. To form an actionable basis for cleaning and repairing operations, the communicated information needs not only be comprehensive, but also somewhat intuitive. Often, metrics may not suffice for truly understanding the nature and extent of some process-data quality problems. For this, appropriate visualisations need to be developed. This may be quite complex. Consider, for example, the visualisation of the manifestation of occurrences of the Scattered Event [51] pattern, where each occurrence may affect multiple columns and rows and the relationships between the attributes affected are not always straightforward.

Furthermore, how do we deal with quality problems that interfere with each other? It may be the case that certain problems only manifest when other problems have been resolved. For example, after the resolution of Scattered Events, synonyms that were previously hidden in text attributes suddenly come to the fore. To address this, quality metrics and visualisations should be dynamically adjusted as problems are resolved and ideally there is guidance which ones to tackle first.

Challenge 14. Which dimensions of data quality are particularly important for process mining? Are certain dimensions more important for specific subareas of process mining? □

The framework of Bose et al. [9] states that missing data, incorrect data, imprecise data, and irrelevant data, affect the quality of an event log. Although these issues concern event log

entries,[2] they have a different impact on process mining results. For example, the timestamp is an essential construct of process mining. Missing and incorrect timestamps have a significant impact on the outcomes of process mining. However, when discovering a process model, timestamps must be accurate only to the extent that a correct order of events in a log is established [40]. On the other hand, a missing resource, for example, is less significant for process discovery than for process enhancement. Basically, the meaning of an event log quality issue depends on the objective and the application area of process mining (e.g., process discovery, conformance checking, process performance analysis, deviance analysis, drift analysis).

Published research in the field tends to focus on event log data quality in general or, often implicitly, considers the issue of event log quality only in the context of process discovery [3, 27]. When considering data quality more broadly, additional dimensions have an impact on it. For example, *accuracy* – in terms of the degree that event data reflects reality – also plays a role. Depending on the objective, the dimensions have different significance for different types of process mining.

Challenge 15. How can we best show how process-data quality issues may have impacted analytical outcomes? □

The impact of process-data quality issues should be documented and shown effectively to the stakeholders of process mining analysis. This can facilitate understanding the repairs that are needed and prioritising them based on their impact on the results. One way to show the impact of data quality issues on analytical outcomes is to perform a sensitivity analysis, where different levels of data quality are simulated and the impact on the analysis is measured. Another approach could be to compare results obtained from a "clean" event log to those obtained from a log known to have quality issues. This way the impact of these issues can be clearly demonstrated. Additionally, a root-cause analysis may be conducted. This involves identifying and investigating the specific issues that led to the poor data quality, and then demonstrating the impact of those issues on the analytical results. Another approach is to create control groups with known good data and compare the results of the analysis with the control group results, which would help to identify any significant differences caused by poor data quality. Finally, documenting and describing the data quality issues and their potential impact using visualisations such as charts or graphs to illustrate the impact can also be an effective way to convey the message.

5 RESEARCH AGENDA AND FUTURE WORK

In this section, we focus on where we believe the field of process-data quality should be heading. We will present a few of the key directions that will move the field forward in the years to come and argue why we believe this to be the case.

Direction 1. The use of domain knowledge in solving process-data quality problems is essential and should be facilitated. □
Related challenges: 2, 3, 4, 8, 9, 11, 12

Exploiting domain knowledge in the development of solutions to detect and repair process-data quality issues will lead to superior solutions. Domain knowledge may serve many purposes, such as helping to improve the accuracy of data (e.g., fixing labels and terminology issues) [47], assisting in determining what features are relevant for inclusion in an event log [3], or preventing future data quality issues through root-cause analysis [2]. Future research should consider how

[2] According to Bose et al. [9] (Table 1), imprecise data only applies to *relationships* (between events and the cases they belong to), *case attributes*, *position* (of events in a case), *activity names*, *timestamps*, *resources*, and *event attributes*; and irrelevant data only applies to *cases* and *events*.

and when domain knowledge should be used. Regarding the "how" question, domain knowledge may be required in algorithmic solutions with human-in-the-loop inputs, e.g., detection and repair approaches that are trained based on the input of domain experts [1]. Developing visualisation techniques that assist humans in these tasks [31] is also an important part of this research direction. Domain experts can be involved directly or indirectly. Directly involving domain experts poses a challenge as their time is costly and limited [37, 49]. This can be addressed through gamification [47] as well as crowdsourcing (see, e.g., [15]). An ontology, in its turn, provides a means to access domain expert knowledge indirectly. Here the challenge lies in the creation of a high-quality ontology. An incentive for domain experts to contribute high-quality knowledge is the longer-term pay-off determined by the reusability of an ontology across process mining projects. It has been shown that the use of gamification techniques can be beneficial to engage experts in creating an ontology [46].

Regarding the "when" question, domain knowledge can clearly contribute to curating, inspecting, and repairing an event log in preparation for process mining analysis. It may also become crucial even before an event log is created, preventing data quality issues through a root-cause analysis and identifying relevant features to be logged. Additionally, domain knowledge can contribute to identifying the impacts of poor data quality when assessing the results after process mining analysis.

Direction 2. Develop and investigate the implementation and effectiveness of process-data governance, with a focus on ensuring that the necessary data for solving business problems is always available and of high quality. □
Related challenges: 1, 2, 5, 12

To effectively answer the most pressing questions for the business, it is crucial to have access to the appropriate data, which should always be easily accessible and of high quality. This principle applies to process mining and process data, and a process-data governance framework has been developed as a first step towards achieving this goal [28]. However, extensive validation of this framework is still required in a variety of settings to ensure its effectiveness. Furthermore, detailed methodological guidance and tool support are necessary to facilitate the operationalisation of process-data governance. Eventually, synthetic data might also be beneficial for this purpose [26, 66]. Generally, it has been shown that synthetic data not only provides a substitution for real data, but can even enhance insight into domain-specific research. The challenge is to generate synthetic data that is very close to what one could encounter in the real environment.

Direction 3. Guidance is essential when solving process-data quality problems. □
Related challenges: 1, 2, 3, 4, 5, 11, 14

Data quality problems can be overwhelming in scale and complexity. Stakeholders need to be supported when they are trying to find such problems and when they are considering possible repairs for problems identified. As an example, visual analytics is a field of study that helps analysts, or stakeholders more generally, recognise issues that are noteworthy through visual means. This makes visual analytics eminently suitable for stakeholder guidance when detecting and repairing data quality problems.

Another form of guidance can be in the form of methodological support. As an example, the stakeholders may need guidance to ensure that the right data is being collected for analysis purposes. Furthermore, stakeholders can be given best practice guidelines to prevent and mitigate data quality problems at source systems as well as a systematic way to undertake the data correlation and pre-processing steps, e.g., taking quality considerations explicitly into account while conducting a process mining project (see, e.g., the Signpost methodology [23]).

Direction 4. Data cleaning should be able to handle a wide range of types of data and log formats. □
Related challenges: 4, 7, 9, 11, 14

Benefits of object-centric multi-event logs over the standard "flat" XES logs have been highlighted by the process mining community [61] and data cleaning should be able to clean these types of logs effectively. In addition, event logs may contain not only structured data but also unstructured data, notably text, and multi-media data. Furthermore, new variants of event data, such as sensor data, have their own unique characteristics that data cleaning techniques should be able to handle [8]. Data cleaning should evolve to deal with this more complex type of data. Data cleaning also needs to account for a wide range of data quality dimensions and metrics to address the huge variety of process-data quality problems. Moreover, data cleaning should leverage techniques from different fields such as statistics, machine learning, and data mining to reduce the associated time and to make it less error-prone.

Direction 5. The consequences of data cleaning should be quantified and scoped when presenting process mining artefacts. □
Related challenges: 6, 15

Cleaning a log is not an exact science and implies making context-specific choices about data elements that should be imputed, changed, or removed. It is important to realise that data cleaning operations will impact the generated process mining outcomes. The more cleaning operations we perform and the wider the scope of these operations, the more likely it is that we have affected the final analysis results in some fundamental manner [35]. Rather than presenting the analysis results as if they originated from the original log, it is imperative that the potential impact of cleaning operations on process mining outcomes is properly quantified. This will allow stakeholders to understand what the ramifications of the cleaning operations are and how carefully (various parts of) the analysis should be interpreted and used. In this respect, process mining outcomes could, e.g., take the form of confidence intervals or parts of process models can be marked such that it becomes clear how much their presence depends on outliers in the log. This way, process mining outcomes can be interpreted against the background of the data cleaning operations that have been performed.

Direction 6. Maintaining provenance of process-data throughout the cleaning process is key. □
Related challenges: 6, 11, 15

In order to be able to fully understand the data that is ultimately subjected to analysis and the analysis results, we need to understand the provenance of the data. When cleaning an event log, in fact, we make choices in terms of what we clean and how we clean. This means that we need to record all operations that have been performed on the data. How we can best keep track of these operations is an open problem. And it is intriguing to think how we can exploit historic operations for the purpose of future data cleaning efforts. Can we learn that certain approaches work better than others in a certain context? This requires insights into how well past cleaning operations worked, based on analyst and stakeholder feedback. Processes and the systems supporting their execution may be modified to collect better process data in the future based on the insights extracted from data provenance information.

Direction 7. Emphasis should shift from the *detection* and *repair* of process-data quality problems to their *prevention* and *mitigation*. □
Related challenges: 2, 5

As argued in the context of the Odigos framework [2], process-data quality problems are reflections of the business context in which this data is created and updated. This context can be seen

as three worlds – the material, personal, and social worlds – and their interactions. It has been shown that well-known process-data quality problems can be traced back to these worlds and their interactions [2]. This provides an opportunity to shift the problem of process-data quality from detection and repair to prevention and mitigation as that tends to be more effective both in terms of the cost involved and the level of quality achieved. Future research should investigate in more depth how interactions in the three worlds may lead to process-data quality problems and how these problems can best be prevented, and if not prevented altogether, how they can be best mitigated. Note that this shifts the problem from a purely IT-problem to one that also involves business considerations.

Direction 8. The whole lifecycle of data needs to be considered and monitored. □
Related challenges: 1, 2, 4, 5, 6, 9, 12

It is helpful to think of data as being part of a lifecycle, from its creation, through to its storage and analysis (there are many variations of data lifecycles in the literature, but see, e.g., the data value chain described in Figure 1 of [52] or the one described in Figure 3.1 of [18]; note also that Figure 3 contains part of such a lifecycle). The various stages of this lifecycle need to be supported and particular emphasis needs to be placed on catching errors as soon as possible in this lifecycle (as also stressed in Direction 7). Monitoring data throughout the lifecycle is important in order to detect whether the quality of existing (types of) data is deteriorating or new problems are emerging. New analysis needs give rise to new types of data having to be collected. This needs to be pickup up sooner rather than later. New types of data may require an understanding of the data quality issues the new data is subject to and may require the enhancement of existing techniques, or even the introduction of new techniques, to resolve them. Techniques are to be developed that can detect *quality drift* and help alert stakeholders to the reasons for this drift in terms understandable to them. Overall, there is a need to be proactive in regard to the use of data rather than considering data and its quality as an afterthought.

Direction 9. Repairs need to be specified at the right level of abstraction and also need to be logged throughout the lifecycle of data. □
Related challenges: 6, 9, 11, 15

To better comprehend systematic issues related to the creation and manipulation of data, it is essential to have a clear understanding of the repairs performed on the data. These repairs should be specific at an appropriate level of abstraction to facilitate their interpretation, enabling root cause analysis. Furthermore, repairs performed on certain data sets over time may help us understand potential treatment of similar data sets. To ensure proper comprehension, the repairs should be represented at a reasonably high level of abstraction. In principle, a mining exercise can be conducted on all the repair operations that have been performed on data sets to analyse the order of repairs, detect quality drifts, and understand the effectiveness of specific repair approaches.

6 CONCLUSIONS

Process mining research historically has focused primarily on process-data analysis techniques, often neglecting the issue of the quality of the input data. Process-data quality, however, is recognised as a crucial concern in real-life projects, where up to 90% can be spent on process-data extraction and cleaning [61].

In this paper, we have identified the main challenges to be addressed to improve the maturity of process-data quality management approaches. Based on the identified challenges, we have proposed several key directions for future research in this field. These directions address

managerial concerns, like designing data governance and data provenance policies, as well as technical concerns, like incorporating the knowledge of domain experts effectively in repair methods or improving process mining artefacts with data quality provenance.

We acknowledge that the proposed challenges and directions reflect the view of the authors and have not been empirically validated. Nevertheless, we argue that their rationale is deeply grounded in the recent literature on process-data quality. As researchers often involved in projects with industry, we also witness the relevance of the identified challenges in our daily involvement in process mining projects with practitioners. As such, we believe that the proposed challenges and directions will help researchers identify crucial research avenues to pursue in the future, as well as practitioners in identifying and possibly anticipating issues related to process-data quality in their daily work.

REFERENCES

[1] Jan Niklas Adams, Sebastiaan J. van Zelst, Lara Quack, Kathrin Hausmann, Wil M. P. van der Aalst, and Thomas Rose. 2021. A framework for explainable concept drift detection in process mining. In *Business Process Management - 19th International Conference, BPM 2021, Rome, Italy, September 06–10, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 12875)*, Artem Polyvyanyy, Moe Thandar Wynn, Amy Van Looy, and Manfred Reichert (Eds.). Springer, 400–416. https://doi.org/10.1007/978-3-030-85469-0_25

[2] Robert Andrews, Fahame Emamjome, Arthur H. M. ter Hofstede, and Hajo A. Reijers. 2022. Root-cause analysis of process-data quality problems. *Journal of Business Analytics* (2022), 51–75. https://doi.org/10.1080/2573234X.2021.1947751

[3] Robert Andrews, Christopher G. J. van Dun, Moe T. Wynn, Wolfgang Kratsch, Maximilian Röglinger, and Arthur H. M. ter Hofstede. 2020. Quality-informed semi-automated event log generation for process mining. *Decision Support Systems* (2020), 113265. https://doi.org/10.1016/j.dss.2020.113265

[4] Adriano Augusto, Raffaele Conforti, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, Andrea Marrella, Massimo Mecella, and Allar Soo. 2018. Automated discovery of process models from event logs: Review and benchmark. *IEEE TKDE* (2018), 686–705.

[5] Donald P. Ballou and Harold L. Pazer. 1985. Modeling data and process quality in multi-input, multi-output information systems. *Management Science* (1985), 150–162.

[6] Carlo Batini, Anisa Rula, Monica Scannapieco, and Gianluigi Viscusi. 2015. From data quality to big data quality. *Journal of Database Management (JDM)* (2015), 60–82.

[7] Carlo Batini and Monica Scannapieco. 2016. *Data and Information Quality - Dimensions, Principles and Techniques*. Springer. https://doi.org/10.1007/978-3-319-24106-7

[8] Yannis Bertrand, Rafaël Van Belle, Jochen De Weerdt, and Estefanía Serral. 2023. Defining data quality issues in process mining with IoT data. In *Process Mining Workshops: ICPM 2022 International Workshops, Bozen-Bolzano, Italy, October 23–28, 2022, Revised Selected Papers*. Springer, 422–434.

[9] Jagadeesh Chandra J. C. Bose, R. S. Mans, and Wil M. P. van der Aalst. 2013. Wanna improve process mining results?. In *IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013, Singapore, 16–19 April, 2013*. IEEE, 127–134. https://doi.org/10.1109/CIDM.2013.6597227

[10] R. P. Jagadeesh Chandra Bose, Wil M. P. van der Aalst, Indre Zliobaite, and Mykola Pechenizkiy. 2014. Dealing with concept drifts in process mining. *IEEE Trans. Neural Networks Learn. Syst.* (2014), 154–171. https://doi.org/10.1109/TNNLS.2013.2278313

[11] Peter Buneman and Susan B. Davidson. 2010. Data provenance–the foundation of data quality. In *Workshop: Issues and Opportunities for Improving the Quality and Use of Data within the DoD, Arlington, USA*. 26–28.

[12] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *Database Theory–ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings*. Springer, 316–330.

[13] Cinzia Cappiello, Marco Comuzzi, Pierluigi Plebani, and Matheus Fim. 2022. Assessing and improving measurability of process performance indicators based on quality of logs. *Information Systems* (2022), 101874.

[14] Josep Carmona, Boudewijn F. van Dongen, Andreas Solti, and Matthias Weidlich. 2018. *Conformance Checking - Relating Processes and Models*. Springer. https://doi.org/10.1007/978-3-319-99414-7

[15] Tianwa Chen, Lei Han, Gianluca Demartini, Marta Indulska, and Shazia W. Sadiq. 2020. Building data curation processes with crowd intelligence. In *Advanced Information Systems Engineering - CAiSE Forum 2020, Grenoble,*

France, June 8–12, 2020, Proceedings (Lecture Notes in Business Information Processing, Vol. 386), Nicolas Herbaut and Marcello La Rosa (Eds.). Springer, 29–42. https://doi.org/10.1007/978-3-030-58135-0_3

[16] 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* (2017), 300–314. https://doi.org/10.1109/TKDE.2016.2614680

[17] Raffaele Conforti, Marcello La Rosa, Arthur H. M. ter Hofstede, and Adriano Augusto. 2020. Automatic repair of same-timestamp errors in business process event logs. In *Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12168)*, Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas (Eds.). Springer, 327–345. https://doi.org/10.1007/978-3-030-58666-9_19

[18] Edward Curry. 2016. The big data value chain: Definitions, concepts, and theoretical approaches. In *New Horizons for a Data-Driven Economy - A Roadmap for Usage and Exploitation of Big Data in Europe*, José María Cavanillas, Edward Curry, and Wolfgang Wahlster (Eds.). Springer, 29–37. https://doi.org/10.1007/978-3-319-21569-3_3

[19] Eduardo González López de Murillas, Hajo A. Reijers, and Wil M. P. van der Aalst. 2018. Connecting databases with process mining: A meta model and toolset. *Software & Systems Modeling* (2018), 1209–1247.

[20] 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 - 41st International Conference, PETRI NETS 2020, Paris, France, June 24–25, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12152)*, Ryszard Janicki, Natalia Sidorova, and Thomas Chatain (Eds.). Springer, 239–259. https://doi.org/10.1007/978-3-030-51831-8_12

[21] Prabhakar M. Dixit, Suriadi Suriadi, Robert Andrews, Moe Thandar 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 - 30th International Conference, CAiSE 2018, Tallinn, Estonia, June 11–15, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10816)*, John Krogstie and Hajo A. Reijers (Eds.). Springer, 274–290. https://doi.org/10.1007/978-3-319-91563-0_17

[22] Fahame Emamjome, Robert Andrews, Arthur H. M. ter Hofstede, and Hajo Reijers. 2020. Alohomora: Unlocking data quality causes through event log context. In *Proceedings of the 28th European Conference on Information Systems*. Association for Information Systems, 1–16. https://aisel.aisnet.org/ecis2020_rp/80

[23] Fahame Emamjome, Robert Andrews, Arthur H. M. ter Hofstede, and Hajo A. Reijers. 2020. Signpost - a semiotics-based process mining methodology. In *ECIS 2020 Research-in-Progress Papers*. Association for Information Systems, 1–10. https://aisel.aisnet.org/ecis2020_rip/50

[24] Ivan P. Fellegi and Alan B. Sunter. 1969. A theory for record linkage. *J. Amer. Statist. Assoc.* (1969), 1183–1210.

[25] Dominik Andreas Fischer, Kanika Goel, Robert Andrews, Christopher G. J. van Dun, Moe Thandar Wynn, and Maximilian Röglinger. 2020. Enhancing event log quality: Detecting and quantifying timestamp imperfections. In *Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12168)*, Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas (Eds.). Springer, 309–326. https://doi.org/10.1007/978-3-030-58666-9_18

[26] Frederik Fonger, Milda Aleknonyte-Resch, and Agnes Koschmider. 2023. Mapping Time-Series Data on Process Patterns to Generate Synthetic Data. In *Advanced Information Systems Engineering Workshops - CAiSE 2023 International Workshops, Zaragoza, Spain, June 12-16, 2023, Proceedings (Lecture Notes in Business Information Processing, Vol. 482)*, Marcela Ruiz and Pnina Soffer (Eds.). Springer, 50–61. https://doi.org/10.1007/978-3-031-34985-0_6

[27] Frank Fox, Vishal R. Aggarwal, Helen Whelton, and Owen Johnson. 2018. A data quality framework for process mining of electronic health record data. In *2018 IEEE International Conference on Healthcare Informatics*. IEEE, 12–21. https://doi.org/10.1109/ICHI.2018.00009

[28] Kanika Goel, Fahame Emamjome, and Arthur H. M. ter Hofstede. 2021. Data governance for managing data quality in process mining. In *Proceedings of the 42nd International Conference on Information Systems*. Association for Information Systems. https://aisel.aisnet.org/icis2021/governance/governance/9

[29] Kanika Goel, Sander J. J. Leemans, Niels Martin, and Moe Thandar Wynn. 2022. Quality-informed process mining: A case for standardised data quality annotations. *ACM Trans. Knowl. Discov. Data* (2022), 97:1–97:47. https://doi.org/10.1145/3511707

[30] Kanika Goel, Sareh Sadeghianasl, Robert Andrews, Arthur H. M. ter Hofstede, Moe Wynn, Dakshi Kapugama Geeganage, Sander J. J. Leemans, James M. McGree, Rebekah Eden, Andrew Staib, Rob Eley, and Raelene Donovan. 2023. Digital health data imperfection patterns and their manifestations in an Australian digital hospital. In *56th Hawaii International Conference on System Sciences, HICSS 2023, Maui, Hawaii, USA, January 3–6, 2023*, Tung X. Bui (Ed.). ScholarSpace, 3235–3244. https://hdl.handle.net/10125/103029

[31] Theresia Gschwandtner. 2015. Visual analytics meets process mining: Challenges and opportunities. In *Data-Driven Process Discovery and Analysis - 5th IFIP WG 2.6 International Symposium, SIMPDA 2015, Vienna, Austria, December*

9–11, 2015, Revised Selected Papers (Lecture Notes in Business Information Processing, Vol. 244), Paolo Ceravolo and Stefanie Rinderle-Ma (Eds.). Springer, 142–154. https://doi.org/10.1007/978-3-319-53435-0_7

[32] Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. 2018. Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)* (2018), 1–32.

[33] Thomas N. Herzog, Fritz J. Scheuren, and William E. Winkler. 2007. *Data Quality and Record Linkage Techniques*. Vol. 1. Springer.

[34] Ihab F. Ilyas and Theodoros Rekatsinas. 2022. Machine learning and data cleaning: Which serves the other? *J. Data and Information Quality*, Article 13 (Jul. 2022), 11 pages. https://doi.org/10.1145/3506712

[35] Agnes Koschmider, Kay Kaczmarek, Mathias Krause, and Sebastiaan J. van Zelst. 2021. Demystifying noise and outliers in event logs: Review and future directions. In *Business Process Management Workshops - BPM 2021 International Workshops, Rome, Italy, September 6–10, 2021, Revised Selected Papers (Lecture Notes in Business Information Processing, Vol. 436)*, Andrea Marrella and Barbara Weber (Eds.). Springer, 123–135. https://doi.org/10.1007/978-3-030-94343-1_10

[36] John Ladley. 2019. *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Academic Press. https://doi.org/10.1016/C2017-0-03353-0

[37] Chaoqun Li, Liangxiao Jiang, and Wenqiang Xu. 2019. Noise correction to improve data and model quality for crowdsourcing. *Engineering Applications of Artificial Intelligence* (2019), 184–191. https://doi.org/10.1016/j.engappai.2019.04.004

[38] Xixi Lu, Dirk Fahland, Robert Andrews, Suriadi Suriadi, Moe Thandar Wynn, Arthur H. M. ter Hofstede, and Wil M. P. van der Aalst. 2017. Semi-supervised log pattern detection and exploration using event concurrence and contextual information. In *On the Move to Meaningful Internet Systems. OTM 2017 Conferences - Confederated International Conferences: CoopIS, C&TC, and ODBASE 2017, Rhodes, Greece, October 23–27, 2017, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 10573)*, Hervé Panetto, Christophe Debruyne, Walid Gaaloul, Mike P. Papazoglou, Adrian Paschke, Claudio Agostino Ardagna, and Robert Meersman (Eds.). Springer, 154–174. https://doi.org/10.1007/978-3-319-69462-7_11

[39] Alfonso Eduardo Márquez-Chamorro, Manuel Resinas, and Antonio Ruiz-Cortés. 2017. Predictive monitoring of business processes: A survey. *IEEE Transactions on Services Computing* (2017), 962–977.

[40] Niels Martin, Antonio Martinez-Millana, Bernardo Valdivieso, and Carlos Fernández-Llatas. 2019. Interactive data cleaning for process mining: A case study of an outpatient clinic's appointment system. In *Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1–6, 2019, Revised Selected Papers (LNBIP, Vol. 362)*, Chiara Di Francescomarino, Remco M. Dijkman, and Uwe Zdun (Eds.). Springer, 532–544. https://doi.org/10.1007/978-3-030-37453-2_43

[41] Niels Martin, Greg van Houdt, and Gert Janssenswillen. 2022. DaQAPO: Supporting flexible and fine-grained event log quality assessment. *Expert Systems with Applications* (2022), 116274. https://doi.org/10.1016/j.eswa.2021.116274

[42] Hoang Nguyen, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Suriadi Suriadi. 2014. Mining business process deviance: A quest for accuracy. In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences - Confederated International Conferences: CoopIS, and ODBASE 2014, Amantea, Italy, October 27–31, 2014, Proceedings (Lecture Notes in Computer Science, Vol. 8841)*, Robert Meersman, Hervé Panetto, Tharam S. Dillon, Michele Missikoff, Lin Liu, Oscar Pastor, Alfredo Cuzzocrea, and Timos K. Sellis (Eds.). Springer, 436–445. https://doi.org/10.1007/978-3-662-45563-0_25

[43] Hoang T. C. Nguyen, Suhwan Lee, Jongchan Kim, Jonghyeon Ko, and Marco Comuzzi. 2019. Autoencoders for improving quality of process event logs. *Expert Systems with Applications* (2019), 132–147.

[44] Marco Pegoraro and Wil M. P. van der Aalst. 2019. Mining uncertain event data in process mining. In *2019 International Conference on Process Mining (ICPM'19)*. IEEE, 89–96.

[45] Anastasiia Pika, Michael Leyer, Moe Thandar Wynn, Colin J. Fidge, Arthur H. M. ter Hofstede, and Wil M. P. van der Aalst. 2017. Mining resource profiles from event logs. *ACM Trans. Manag. Inf. Syst.* (2017), 1:1–1:30. https://doi.org/10.1145/3041218

[46] Sareh Sadeghianasl, Arthur H. M. ter Hofstede, Moe T. Wynn, Selen Turkay, and Trina Myers. 2021. Process activity ontology learning from event logs through gamification. *IEEE Access* (2021), 165865–165880. https://doi.org/10.1109/ACCESS.2021.3134915

[47] 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 *2nd International Conference on Process Mining, ICPM 2020, Padua, Italy, October 4–9, 2020*, Boudewijn F. van Dongen, Marco Montali, and Moe Thandar Wynn (Eds.). IEEE, 41–48. https://doi.org/10.1109/ICPM49681.2020.00017

[48] Sareh Sadeghianasl, Arthur H. M. ter Hofstede, Moe Thandar 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 - Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece,*

October 21–25, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11877), Hervé Panetto, Christophe Debruyne, Martin Hepp, Dave Lewis, Claudio Agostino Ardagna, and Robert Meersman (Eds.). Springer, 76–94. https://doi.org/10.1007/978-3-030-33246-4_5

[49] Edoardo Scibona. 2018. *Cost-effective and Scalable Activity Matching using Crowdsourcing*. Master's thesis. Politecnico di Milano, Milan, Italy.

[50] Minseok Song and Wil M. P. van der Aalst. 2008. Towards comprehensive support for organizational mining. *Decis. Support Syst.* (2008), 300–317. https://doi.org/10.1016/j.dss.2008.07.002

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

[52] Ikbal Taleb, Mohamed Adel Serhani, Chafik Bouhaddioui, and Rachida Dssouli. 2021. Big data quality framework: A holistic approach to continuous quality management. *Journal of Big Data* (2021), 76. DOI: https://doi.org/10.1186/s40537-021-00468-0

[53] Arava Tsoury, Pnina Soffer, and Iris Reinhartz-Berger. 2018. A conceptual framework for supporting deep exploration of business process behavior. In *Conceptual Modeling: 37th International Conference, ER 2018 (Lecture Notes in Computer Science)*. Springer, 58–71.

[54] Wil M. P. van der Aalst, Arya Adriansyah, Ana Karla Alves de Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, R. P. Jagadeesh Chandra Bose, Peter van den Brand, Ronald Brandtjen, Joos C. A. M. Buijs, Andrea Burattin, Josep Carmona, Malú Castellanos, Jan Claes, Jonathan E. 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 W. Günther, Antonella Guzzo, Paul Harmon, Arthur H. M. ter Hofstede, John Hoogland, Jon Espen Ingvaldsen, Koki Kato, Rudolf Kuhn, Akhil Kumar, Marcello La Rosa, Fabrizio Maria Maggi, Donato Malerba, R. 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 D. 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 Thandar Wynn. 2011. Process mining manifesto. In *Business Process Management Workshops (LNBIP, Vol. 99)*, Florian Daniel, Kamel Barkaoui, and Schahram Dustdar (Eds.). Springer, 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

[55] Wil M. P. van der Aalst. 2016. *Process Mining - Data Science in Action, Second Edition*. Springer. https://doi.org/10.1007/978-3-662-49851-4

[56] 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 - 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11724)*, Peter Csaba Ölveczky and Gwen Salaün (Eds.). Springer, 3–25. https://doi.org/10.1007/978-3-030-30446-1_1

[57] Wil M. P. van der Aalst, Arthur H. M. ter Hofstede, Bartek Kiepuszewski, and Alistair P. Barros. 2003. Workflow patterns. *Distributed and Parallel Databases* (2003), 5–51. https://doi.org/10.1023/A:1022883727209

[58] 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 - 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8–12, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9097)*, Jelena Zdravkovic, Marite Kirikova, and Paul Johannesson (Eds.). Springer, 297–313. https://doi.org/10.1007/978-3-319-19069-3_19

[59] R. Wang and D. M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *JMIS* (1996).

[60] Richard Y. Wang. 1998. A product perspective on total data quality management. *Commun. ACM* (1998), 58–65.

[61] Moe Thandar Wynn, Julian Lebherz, Wil M. P. van der Aalst, Rafael Accorsi, Claudio Di Ciccio, Lakmali Jayarathna, and H. M. W. Verbeek. 2021. Rethinking the input for process mining: Insights from the XES survey and workshop. In *Process Mining Workshops - ICPM 2021 International Workshops, Eindhoven, The Netherlands, October 31–November 4, 2021, Revised Selected Papers (Lecture Notes in Business Information Processing, Vol. 433)*, Jorge Munoz-Gama and Xixi Lu (Eds.). Springer, 3–16. https://doi.org/10.1007/978-3-030-98581-3_1

[62] Moe Thandar Wynn, Erik Poppe, J. Xu, Arthur H. M. ter Hofstede, Ross Brown, Azzurra Pini, and Wil M. P. van der Aalst. 2017. ProcessProfiler3D: A visualisation framework for log-based process performance comparison. *Decis. Support Syst.* (2017), 93–108. https://doi.org/10.1016/j.dss.2017.04.004

[63] Moe Thandar Wynn and Shazia W. Sadiq. 2019. Responsible process mining - A data quality perspective. In *Business Process Management - 17th International Conference, BPM 2019, Vienna, Austria, September 1–6, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11675)*, Thomas T. Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling (Eds.). Springer, 10–15. https://doi.org/10.1007/978-3-030-26619-6_2

[64] Jing Yang, Chun Ouyang, Wil M. P. van der Aalst, Arthur H. M. ter Hofstede, and Yang Yu. 2022. *OrdinoR*: A framework for discovering, evaluating, and analyzing organizational models using event logs. *Decis. Support Syst.* (2022), 113771. https://doi.org/10.1016/j.dss.2022.113771

[65] Tobias Ziolkowski, Lennart Brandt, and Agnes Koschmider. 2021. ElogQP: An event log quality pointer. In *Proceedings of the 13th European Workshop on Services and their Composition (ZEUS 2021), Bamberg, Germany, February 25–26, 2021 (CEUR Workshop Proceedings, Vol. 2839)*, Johannes Manner, Stephan Haarmann, Stefan Kolb, Nico Herzberg, and Oliver Kopp (Eds.). CEUR-WS.org, 42–45. https://ceur-ws.org/Vol-2839/paper8.pdf

[66] Yorck Zisgen, Dominik Janssen, and Agnes Koschmider. 2022. Generating synthetic sensor event logs for process mining. In *Intelligent Information Systems - CAiSE Forum 2022, Leuven, Belgium, June 6–10, 2022, Proceedings (Lecture Notes in Business Information Processing, Vol. 452)*, Jochen De Weerdt and Artem Polyvyanyy (Eds.). Springer, 130–137. https://doi.org/10.1007/978-3-031-07481-3_15

Received 28 July 2023; accepted 30 July 2023