A Collection of Publicly Available Event Logs Enhanced by Metadata

Ana Costa[1], Selin Y. Eroglu[1], Kerstin Andree[1] and Luise Pufahl[1]

[1]Technical University of Munich, School of Computation, Information and Technology, Heilbronn, Germany

Abstract

With the rising importance of process mining and business process management research, access to suitable event logs is critical for research artifact development and evaluation. However, the current landscape of publicly available data lacks in metadata. This poses a challenge for researchers to identify relevant event logs for their research objectives. We address this gap by introducing a metadata structure for event logs and describing a collection of publicly available event data. 98 event logs were analyzed and categorized based on 37 criteria relevant to process mining research. A collection containing these logs and categorization is provided and analyzed with two use cases.

Keywords

Event data, Publicly Available Event logs, Collection of Event Logs

1. Introduction

Event logs are fundamental for discovering, monitoring, and improving business processes [1], and the extraction of valuable information is achieved using process mining techniques [2]. When developing new process mining artifacts, publicly available, real-world event logs from various domains are crucial–not only for identifying relevant requirements but also for evaluating the artifacts in realistic settings. Over the years, the process mining community has compiled a rich collection of event logs. These include datasets published through the Business Process Intelligence (BPI) Challenges (eg. [3]), logs shared as supplementary material to research papers, and logs extracted from public sources, such as MIMIC-IV [4] or the Ethereum blockchain [5]. However, these logs are distributed across various platforms and are often shared with limited metadata. As a result, effectively utilizing them requires significant manual effort. Researchers must conduct time-consuming preliminary assessments to determine whether a log fits their needs, due to the lack of structured descriptions and inconsistent metadata annotations [6].

This paper presents a curated collection of publicly available event logs that have been enriched with an enhanced metadata structure. The structure was developed based on requirements formulated in interviews with process mining researchers and subsequently validated and refined in a second round of interviews and a focus group discussion. Based on this metadata schema, pre-selected event logs were assessed and annotated in detail. For event log selection, we followed a systematic review methodology adapted from the PRISMA guidelines [7]. Datasets from 4TU, Kaggle, UC Irvine, and IEEE were assessed against predefined inclusion and exclusion criteria. Inclusion criteria required public accessibility, compatibility with common process mining formats (e.g., XES, CSV), and the presence of mandatory event attributes. Datasets were excluded if they exceeded 600 MB, required payment or preprocessing, or lacked English documentation. Researchers can now explore the collection using search and filter capabilities and download logs that match their specific requirements.

The remainder of this paper is structured as follows: Section 2 introduces the event log resource and

Proceedings of the Best BPM Dissertation Award, Doctoral Consortium, and Demonstrations & Resources Forum co-located with 23rd International Conference on Business Process Management (BPM 2025), Seville, Spain, August 31st to September 5th, 2025.
a.costa@tum.de (A. Costa); selinyagmureroglu@gmail.com (S. Y. Eroglu); kerstin.andree@tum.de (K. Andree); luise.pufahl@tum.de (L. Pufahl)
0000-0001-9241-5614 (A. Costa); 0000-0001-6711-8458 (S. Y. Eroglu); 0009-0007-6360-8661 (K. Andree); 0000-0002-5182-2587 (L. Pufahl)

metadata structure; Section 3 discusses its preliminary usage; and Section 4 outlines future applications and directions for this work.

2. Description of the Resource

The resource contains a set of 98 publicly available event logs, whose metadata was enriched based on domains and data features relevant to process mining research. The metadata structure for event logs includes 37 attributes with (1) basic process information, (2) domain context, (3) academic context, (4) data characteristics, and (5) resource context. The pre-selective event logs contain the three mandatory attributes of process mining [8] and can be accessed in 4TU, Kaggle, UC Irvine, or IEEE. They are provided in either eXtensible Event Stream (XES) or Comma-separated values (CSV) format. The collection does not include data files larger than 600 MB, with restricted or paid access, or not in English. Figure 1 shows metadata structure with the attributes that give context to the event logs, as well as a short description of these.

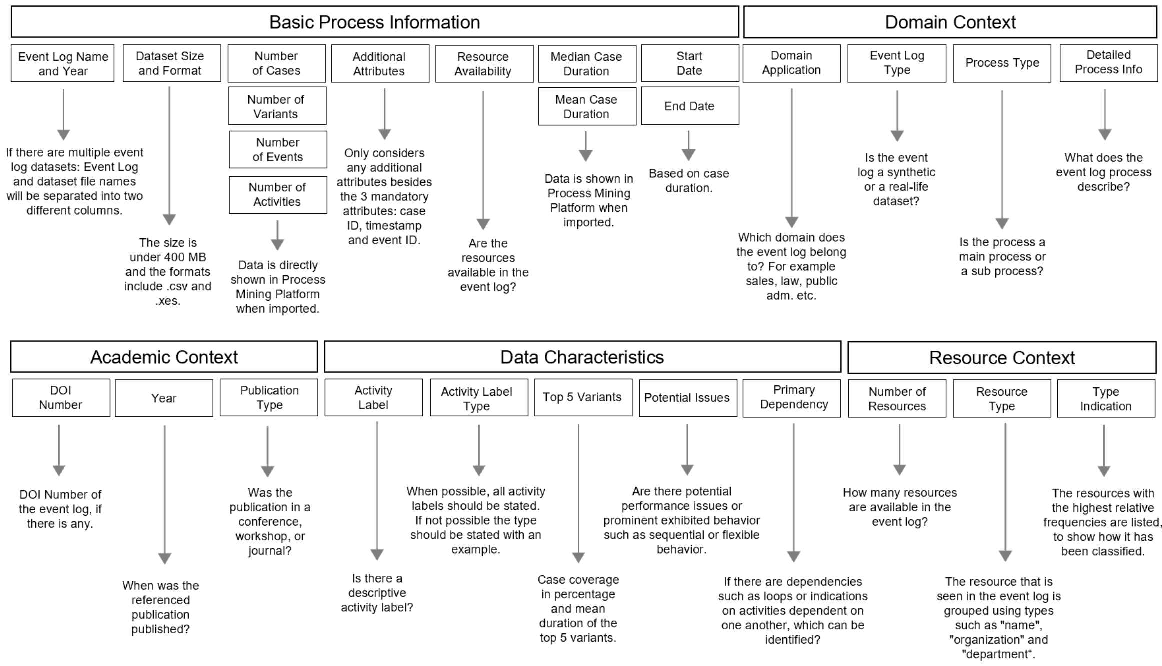

Figure 1: Metadata structure for event logs with 37 attributes categorized in five attribute dimensions.

We developed the metadata structure for event logs through the following procedure: First, we conducted five structured interviews with experts to identify relevant metadata attributes from a research perspective. The experts shared their requirements and highlighted current challenges in process mining research stemming from the limited availability and accessibility of event log data. We synthesized common requirements from the interviews and organized them into attributes and corresponding attribute dimensions. To validate our extraction, we followed a two-step approach: (1) we confirmed with the same experts in a second interview round that all their requirements were addressed, and (2) we conducted a focus group discussion with three additional process mining researchers to reflect on the terminology and completeness of the structure. Based on the finalized set of metadata attributes, we assessed each selected event log and compiled the information into the final metadata table. The following subsections present a short description of the metadata attributes and logs included in the collection.

2.1. Basic Process Information, Domain Context, and Academic Context.

The metadata structure for event logs offers filtering capabilities based on fundamental information about the event logs, such as their names, publication years, dataset size, format of the log, and summary statistics, including the number of cases, variants, events, and activities. The final collection includes event logs ranging from 2010 to 2026, with 28 logs published or updated in 2017. The dataset comprises 61 event logs in XES format and 37 in CSV format, with file sizes ranging from 1.1 KB to 551 MB. The highest number of recorded cases is 251,734, and the number of variants ranges from 1 to 22,632.

Case duration provides information on the time between the start and end of the event log traces given as mean and median. With the help of this, event logs can be identified that describe short or long running business processes. Our identified event logs show periods ranging from hours to months. The additional available start and end timestamps indicate that events are typically recorded shortly before the year of publication of the event log; however, there are some outliers in timestamps that exhibit start dates such as 1948 or 1970. Other criteria, such as the additional attributes (e.g., costs, status, descriptions) and resource availability, vary across format types. XES logs contain up to 16 non-mandatory attributes, and resources are available in half of the event logs.

The domain context classifies the real-world domain or industry from which the event log originates (e.g., healthcare, manufacturing, finance) and contextualizes the recorded process with information such as whether it is the main or sub-process, a real-world or synthetic log, and offers a brief process description. The most frequent domain applications for real-world event logs are public administration, IT service management, and healthcare, while synthetic logs are provided mostly for process mining applications and administrative processes. Furthermore, the academic context encompasses information related to the research environment, including the DOI, publication time, and type.

2.2. Data Characteristics and Resource Context.

The data characteristics analysis covers factors that describe the activity label, information on the top five variants (e.g., number of cases in the five most frequent variants, mean duration), or factors that affect data analysis, such as potential performance issues, prominent exhibited behavior of the process, or dependencies between activities. Almost all logs have activity labels, and the descriptions vary from interpretable names to randomized characters requiring additional context (33% of the logs name their activities with an alphabetical letter, and the others with the verb of the action being performed).

The top five variants are observed in relation to case coverage and mean duration. With that, it is possible to analyze whether a process is standardized or if the event log is exhibiting behavior where the obtained case coverage is low. This attribute shows how many cases, the percentage of case coverage in comparison to the total number of cases, and the mean duration when only the five most frequent variants are considered. Potential issues are also being observed since we want to offer the possibility to identify processes with sequential behavior and event logs containing more flexible behavior. With this attribute, it is possible to filter logs that exhibit a sequential behavior, a behavior with parallel structures, or a more flexible behavior. Furthermore, primary dependency shows loops between activities or indications of activities that are dependent on one another.

Finally, resource context provides some detailed information regarding resources. Although 55% of the logs contain resources, the resources range from 0 to 1440 across all logs. While some resources are represented numerically, others contain the organization name or profession. The attribute resource type indicates how the resources for each log are named. Furthermore, the type indication parameter shows two or three examples of how the resources are named in each log.

3. Preliminary Analysis

The collection of event logs has been applied in two distinct use cases, a project in the field of process discovery and a process in process prediction, which are demonstrated in this section. Each use case is presented with a description of its specific objectives and associated requirements. For each requirement, we discuss its coverage by the proposed framework and name the exact filtering option. This is followed by an explanation of how the proposed resource was utilized, along with a summary of the outcomes, specifically whether relevant and use case-specific event log data could be identified.

3.1. Use Case: Process Discovery and Feature Extraction.

This project focuses on the discovery of relevant features of decision tasks in processes. The research team aimed to analyze specific attributes of activities that contribute to or precede particular decisions. While the project was situated within the context of financial processes (R2.1), its scope also extends to processes that involve decision points more generally (R2.2). A central objective was to understand which aspects of process execution are taken into account when decisions are made. Event log documentation was considered a key requirement (R2.3). Related publications ensure clarity regarding the structure, semantics, and context of the data, thereby supporting a more accurate interpretation of the decision-related features within the processes. Table 1 summarizes the requirements and shows the corresponding filtering applied to the categorization framework.

Table 1
Requirements of process discovery use case and their coverage in the categorization framework

No.	Requirement	Framework Coverage	Filter
R1.1	Financial Domain	✓	Domain Application = Financial, Banking
R1.2	Decision Logic	×	
R1.3	Documentation Available	✓	Publication Type $\neq N/A$

R1.1 covers the contextual requirement and is covered by the categorization framework. The attribute *Domain Application* provides filtering functionality with regard to the domain of the process. It was set to *Financial* and *Banking* since both domains are interesting for the overall requirement of having process data of financial processes. The number of decision points within processes is not covered by the categorization framework. Even though the number of variants is given for every event log, it is not made clear whether these variants are due to decision points or other behavior patterns, such as parallelization. R1.3, however, is covered. The *academic context* offers possibilities to filter for publication-based event logs or event logs to which a publication can be associated. This filtering option was set to not match N/A so that it is ensured the filtered datasets are well-documented. In total, the applied filtering resulted in six event logs, each fulfilling the requirements and objectives of the project.

3.2. Use Case: Event Log Sampling for Next Activity Prediction.

This project focused on finding suitable samples of event logs for training a next-step activity prediction model. For that, it is necessary to foresee undesired execution of activities (R2.1) and to obtain a considerable representation of the process through frequent traces (R2.2). The distribution of data attributes should be pre-analyzed, e.g., by computing the frequency of categorical data values or the mean of numerical attributes (R2.3). With that, traces of each variant can be sorted and given a priority to traces that have more resources of each variant. Finally, a sampling function is applied that returns traces with higher priorities. Table 2 shows the corresponding requirements and filtering criteria.

All requirements are covered by the attributes of the metadata and could be selected additionally with different filter settings. R2.1, for example, was fulfilled by filtering out sequential or parallel behavior from the logs with *prominent exhibit behavior*, but using the *primary dependency* filter to recognize loops or indications of dependency between activities could also be relevant for this requirement. In order to have a representative trace frequency (R2.2), the *number of variants* was filtered between 100 and 5000 variants, but other filtering could have been included, such as the percentage of case coverage and mean duration of the *top five variants*. R2.3 was fulfilled by filtering the *mean case duration* with a specific desired time range, but it could also have been covered by computing the frequency of categorical data,

Table 2
Requirements of event log sampling for next activity prediction

No.	Requirement	Framework Coverage	Filter
R2.1	Foresee Undesired Execution	✓	Prominent Exhibit Behavior $\neq$ Sequential, Parallel Behavior
R2.2	Representative Trace Frequecy	✓	Number of variants $\between$ 100 and 5000
R2.3	Mean of Numerical Attributes	✓	Mean Case Duration $\between$ 100 minutes and 100 hours

since the activity label description shows the three most frequent activities of each log. As a final result, the framework covered all requirements and resulted in 12 possible event logs.

4. Possible Usage and Outlook

A collection of 98 publicly available event logs categorized by 37 metadata attributes serves as a valuable resource for researchers in the field of process mining. The collection of event logs enhanced by metadata is available in Zenodo[1] and is licensed under the Creative Commons Attribution 4.0 International[2]. The availability and license of each log is provided together with the resource, and all logs are free to copy and redistribute for research purposes. The ability to filter logs based on domains or data features accelerates the search process, supports reproducibility, and ensures efficient selection of logs for research. By organizing event logs according to key characteristics, the metadata structure offers a clear overview that facilitates informed dataset selection. We encourage researchers to extend the framework by continuously adding new logs along with relevant metadata, fostering a growing and structured repository. The provided resource, thus, democratizes process data handling by increasing data accessibility. Event logs can be searched and found efficiently and purposefully.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] W. M. P. Van Der Aalst, Process Mining: Data Science in Action, Springer Publishing Company, Incorporated, 2018.

[2] J. De Weerdt, M. De Backer, J. Vanthienen, B. Baesens, A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs, IS 37 (2012) 654–676.

[3] B. van Dongen, Bpi challenge 2019, 2019. URL: https://doi.org/10.4121/UUID:D06AFF4B-79F0-45E6-8EC8-E19730C248F1, data set.

[4] J. Cremerius, L. Pufahl, F. Klessascheck, M. Weske, Event log generation in MIMIC-IV research paper, in: Process Mining Workshops - ICPM, Bozen-Bolzano, Italy, Springer, 2022, pp. 302–314.

[5] H. D. Bandara, H. Bockrath, R. Hobeck, C. Klinkmüller, L. Pufahl, M. Rebesky, W. van der Aalst, I. Weber, Event logs of ethereum-based applications, in: BPM'21, Rome, Italy, 2021.

[6] A. Berti, G. Park, M. Rafiei, W. Aalst, A generic approach to extract object-centric event data from databases supporting sap erp, Journal of Intelligent Information Systems 61 (2023) 1–23.

[7] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the prisma statement, BMJ 339 (2009) b2535–b2535. doi:10.1136/bmj.b2535.

[8] W. M. van der Aalst, Process mining: a 360 degree overview, in: J. C. Wil M. P. van der Aalst (Ed.), Process Mining Handbook, Springer, 2022, pp. 3–34.

[1] The resource is available in Zenodo with the link https://zenodo.org/records/16268743
[2] The license is described at https://creativecommons.org/licenses/by/4.0/deed.en