Rethinking the Input for Process Mining: Insights from the XES Survey and Workshop

Moe Thandar Wynn[1], Julian Lebherz[2], Wil van der Aalst[3], Rafael Accorsi[4], Claudio Di Ciccio[5], Lakmali Jayarathna[1], and H.M.W. Verbeek[6]

[1] Queensland University of Technology, Brisbane, Australia, {m.wynn,lakmali.herathjayarathna}@qut.edu.au
[2] A.P. Møller-Mærsk, Denmark, IEEE_TFPM_SC@lebherz.me
[3] RWTH Aachen University, Germany, wvdaalst@pads.rwth-aachen.de
[4] Accenture, Switzerland, rafael.accorsi@accenture.com
[5] Sapienza University of Rome, Italy, claudio.diciccio@uniroma1.it
[6] Eindhoven University of Technology, The Netherlands, h.m.w.verbeek@tue.nl

Abstract. Although the popularity and adoption of process mining techniques grew rapidly in recent years, a large portion of effort invested in process mining initiatives is still consumed by event data extraction and transformation rather than process analysis. The IEEE Task Force on Process Mining conducted a study focused on the challenges faced during event data preparation (from source data to event log). This paper presents findings from the online survey with 289 participants spanning the roles of practitioners, researchers, software vendors, and end-users. These findings were presented at the XES 2.0 workshop co-located with the 3rd International Conference on Process Mining. The workshop also hosted presentations from various stakeholder groups and a discussion panel on the future of XES and the input needed for process mining. This paper summarises the main findings of both the survey and the workshop. These outcomes help us to accelerate and improve the standardisation process, hopefully leading to a new standard widely adopted by both academia and industry.

Key words: Process Mining, XES, Event Data, Data Transformation

1 Introduction

It is well known that data pre-processing is the most time-consuming task of a process mining project. The XES workshop, organised by the IEEE Task force on Process Mining XES working group, aims to seek contributions from process mining vendors and researchers on the challenges faced in curating data input for process mining projects. The scope of the workshop covers the different aspects of the data input pipeline, starting from the raw event data to generating an event log (e.g., data curation, data cleaning, data standardisation). The intended outcome is a collection of data-related challenges and potential solutions to address these challenges. This paper summarises the main findings from this initiative.

The rest of the paper is organised as follows: Section 2 provides an overview of the current IEEE standard for eXtensible Event Stream (XES). Section 3 describes the key insights from the online survey, while Section 4 synthesises the discussion on the day of the XES workshop. Section 5 concludes the paper.

2 XES Standard: A Brief Overview

MXML (Mining eXtensible Markup Language), defined in 2003, was the first process mining standard to exchange event data [1]. Due to its limitations, the standardisation for new format called XES started in 2009 supported by the IEEE Task Force on Process Mining. Already in the first meeting of the Task Force on September 15th 2010 at the Stevens Institute of Technology in Hoboken USA there was consensus to establish XES as an official standard. The XES standard was adopted by the IEEE Standards Association (SA) as the "IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams" [2] in 2016.

After the adoption of the XES standard by the IEEE, work was done on creating new extensions to the XES standard. The conceptual model of XES introduces components (logs, traces, events, and attributes) that may all contain attributes. Every such attribute is represented as a key-value mapping, where the value is assigned according to the attribute's type (string, timestamp, integer, real, boolean, ID, or list).

The purpose of the extensions was, and still is, to provide semantics to the attribute keys. A typical example for this is the "`concept:name`" key, which is generally considered to be the name of the corresponding activity (for an event) or the name of the corresponding case (for a trace). However, to provide this key with semantics, the `Concept` extension needs to be included in the XES log, as, by default, keys have no fixed semantics. To provide semantics to some basic attributes, the XES standard comes with a collection of standard extensions[1]. The `Concept` extension is a typical example thereof, and the standard additionally includes the `Lifecycle`, `Time`, `Organizational`, and `Cost`. In the end, this work led to the adoption of a number of additional extensions by the XES Working Group (WG), like `Micro` in 2016, `Software` in 2017, and `Artifact Lifecycle` in 2018.

However, the adoption of the XES standard by the different software tools in the process mining community remained low. Also, whenever a tool claimed to support the XES standard it was often unclear to what extent it supported the XES standard. To provide a better overview of this support of the XES standard, the XES Working Group initiated a XES certification process in 2017. As a result, at the time of writing **twelve** process mining tools[2] have been certified by the XES WG as supporting the XES standard. The XES standard helped to progress the field of process mining. It led to consensus about core concepts [1] and many publicly available event logs were made available for competitions and benchmarks. However, adoption in industry is limited, mostly due to the verbosity of the XML serialisation of XES. Moreover, the extraction and pre-processing of event data is still seen as a limiting factor for process mining.

3 Survey Design and Insights

To investigate the challenges faced during event data preparation for process mining, we conducted an online survey collecting the insights from the process mining community from various roles (i.e., academia, professional services, software vendors, and commercial end users).

[1] `www.tf-pm.org/resources/xes-standard/about-xes/standard-extensions`
[2] `www.tf-pm.org/resources/xes-standard/for-vendors/certification/tools`

Survey Design. The survey instrument was developed by the XES WG through several review iterations. The survey contained **12** questions and captured the participants' insights on the suggestions for speeding up the data pre-processing, particularly to understand what enhancement can be made to an industry-wide process mining data standard such as XES.

1. How much experience do you have with Process Mining?
2. Which area and role best describe how you have interacted with PM?
3. What share of effort is typically spent on data pre-processing?
4. Which process mining solutions have you used?
5. Which technologies have you used in data pre-processing for process mining?
6. In which format(s) is your source data available in?
7. Which source systems have you analysed with process mining?
8. How big was the largest data set you worked with in process mining?
9. To what extent did you encounter the following data-related challenges while undertaking PM projects in terms of sourcing data, processing data, analysing process data?
10. Which data-related challenges have you encountered beyond the ones listed in question 9?
11. There is general consensus amongst practitioners that data pre-processing tasks still consume most of the effort put into process mining initiatives. How could we speed up the data pre-processing to focus on analysis?
12. How could a re-imagined industry-wide process mining data standard help you excel in your role?

The XES online survey was distributed to the international process mining community (through LinkedIn posts, email lists and website announcements) and was opened from June to July 2021. In total, 290 responses were received. A duplicate response was detected and removed, thus the total number of responses used for the analysis is **289**.

Survey Insights. The responses for Questions 1 to 9 were quantitatively analysed using the descriptive and frequency analysis. In addition, the responses are grouped based on a participant's role. Free-text responses provided in Questions 10, 11, and 12 were analysed by a research assistant to identify the emerging themes and then reviewed by two XES WG members. This led to the final grouping of common themes presented later in the section.

Out of the 289 responses, the highest response rate is from the professional service role ($n = 112$, 39 %), followed by academia ($n = 97$, 33 %), software vendors ($n = 46$, 16 %), and commercial end users ($n = 34$, 12 %), as depicted in Fig. 1. The highest range of experience reported was 2-5 years (38 %), followed by 5-10 years (24 %), 1-2 years (18 %), 10+ years (10 %), and less than one year (9 %). Participants with no experience are less than 1 %.
Next, we present individual key findings for Questions 3–12.

Q3: What share of effort is typically spent on data pre-processing? Figure 2 shows that 61 % to 80 % of the effort of share for data pre-processing is the highest reported response by participants (36 %) across all roles. The maximum percentage reported was 90 % for the academic role and the professional service role. These results confirm that a significant amount of effort is being spent to pre-process event data for process mining. It is also interesting to notice that most of the participants with less than one year of experience did not respond to this question. This may indicate that process mining novices are more focused on the novel techniques and tool development than on the input data.

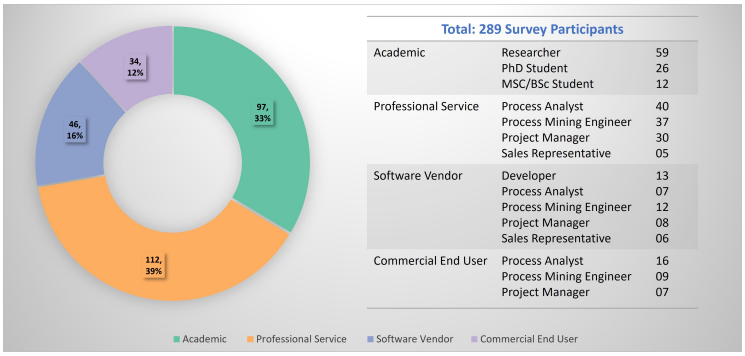


Fig. 1. XES survey participants: demographics

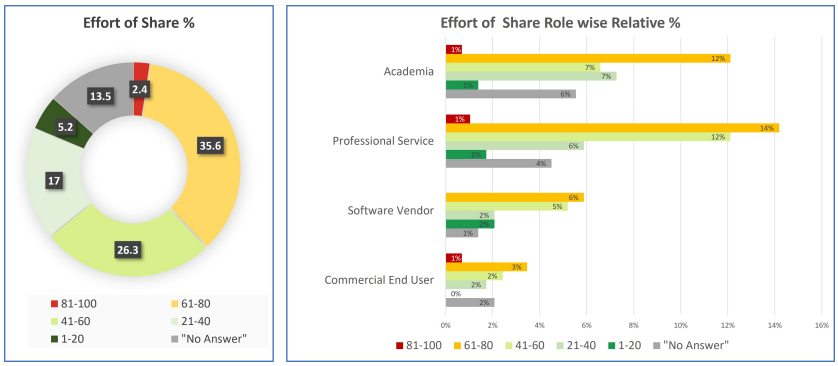
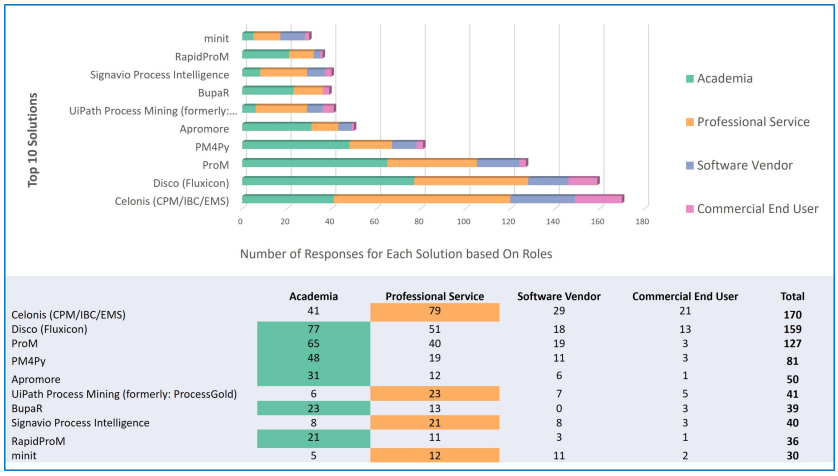


Fig. 2. Q3: Share of effort on data pre-processing

Q4: Which process mining solutions have you used? Celonis is the overall highest selection ($n = 170$), with Disco ($n = 159$) and ProM ($n = 127$) rounding off the top three process mining solutions reported by the participants (see Fig. 3). Note that it is possible for participants to select multiple solutions, and many opted for this. The role-wise comparison for the top ten process mining solutions, where variations can be observed among the four roles. For example, Disco (Fluxicon) is the most selected option for academics ($n = 77$), closely followed by ProM ($n = 65$).

Q5: Which technologies have you used in data pre-processing for process mining? Microsoft SQL server is the highest selected response for database management and data storage systems ($n = 125$). Figure 4 shows a slightly different perspective among the four roles, with academia selecting MySQL ($n = 45$) ahead of Microsoft SQL server and the software vendors preferring PostgresSQL ($n = 26$). PowerBI ($n = 122$) has been the most selected response as a data visualisation tool (see Fig. 5). Python ($n = 177$) turned out to be the most used custom data transformation language (see Fig. 6).

Q6: In which formats is your source data available in? A plain text file (e.g., `txt` or `csv`) is the most commonly available source data format ($n = 229$), with the relational format



	Academia	Professional Service	Software Vendor	Commercial End User	Total
Celonis (CPM/IBC/EMS)	41	79	29	21	**170**
Disco (Fluxicon)	77	51	18	13	**159**
ProM	65	40	19	3	**127**
PM4Py	48	19	11	3	**81**
Apromore	31	12	6	1	**50**
UiPath Process Mining (formerly: ProcessGold)	6	23	7	5	**41**
BupaR	23	13	0	3	**39**
Signavio Process Intelligence	8	21	8	3	**40**
RapidProM	21	11	3	1	**36**
minit	5	12	11	2	**30**

Fig. 3. Q4: 10 most used process mining solutions

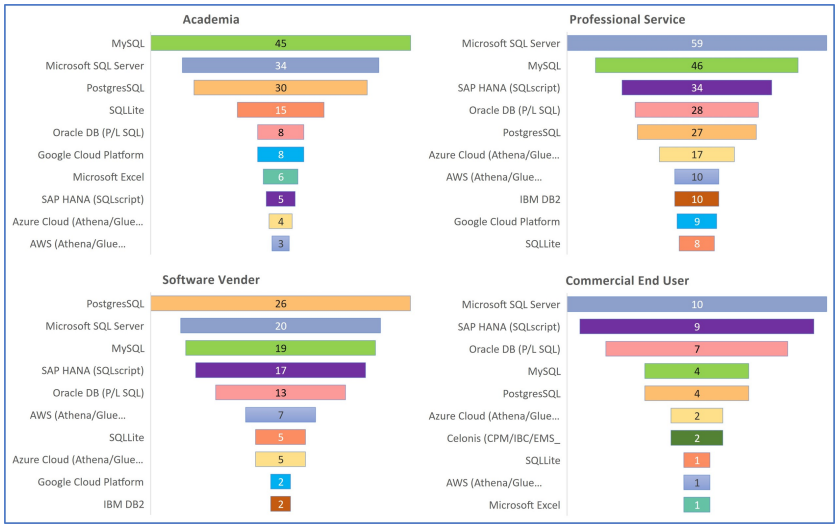


Fig. 4. Q5: 10 most used database management and data storage systems (role-wise)

access ($n = 168$), and the XML format such as XES ($n = 112$) being selected as the second and the third most common ones (see Fig. 7). Please notice that participants could select more than one source data format. The responses also confirm that XML (e.g., XES) is not widely used in the community with only 39 % ($n = 112$) selecting this option.

The frequencies and their relative order among the top five source formats are also different across the different roles (as shown in Fig. 8). For example, XML is the second, third, fourth, and third choice for academia, professional services, software vendors, and commercial end users, respectively.

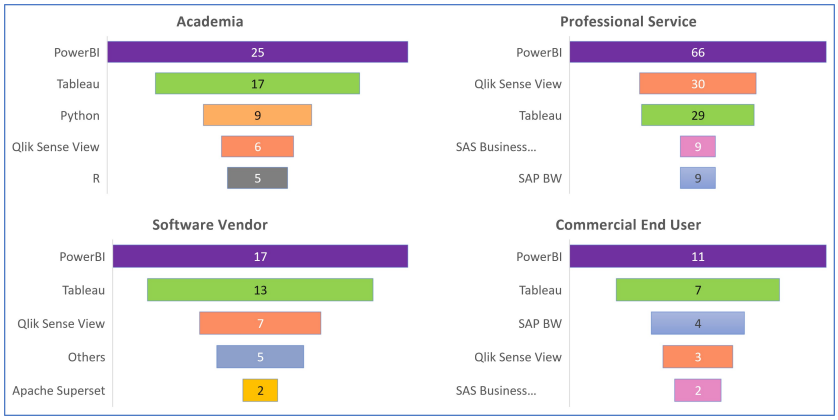


Fig. 5. Q5: 10 most used data visualisation technologies (role-wise)

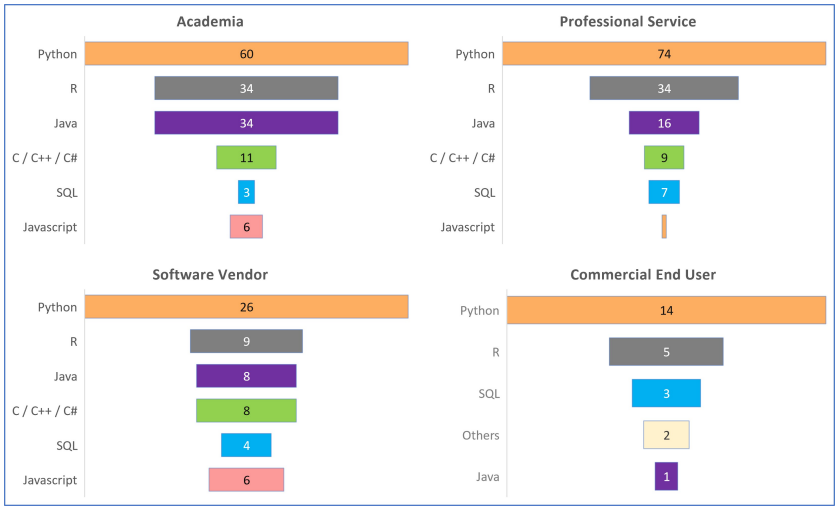


Fig. 6. Q5: 10 most adopted custom data transformation languages (role-wise)

Q7: Which source systems have you analysed with process mining? SAP ECC (R/3) ($n=114$), SAP S/4 HANA ($n=101$), and Salesforce ($n=71$) are the top three most analysed source systems (see Fig. 9). Interestingly, 35 % of academics (34 out of 97) selected "I don't know" as their response for this question. This is probably due to the fact that they primarily work with publicly available data sets such as those provided by the BPI challenges.

Q8: How big was the largest data set you worked with in process mining? Around 16 % of participants ($n=45$) have mentioned that they have worked with less than 1000 events and 0.05 % participants have mentioned that they have worked with more than 1,000,000,000 (1 billion) events. Moreover, around 20 % of the participants ($n=58$) have worked with less than 1000 process cases or instances and around 4 % of the participants ($n=12$) mentioned that the highest number of process cases or instances they have worked with is larger than 1 billion.

	Academia	Professional Service	Software Vendor	Commercial End User	Total
Plain text files (e.g. *.txt, *.csv)	84	86	39	18	**229**
Relational format access (e.g., to MS SQL, Oracle DB or SAP HANA)	37	73	37	18	**168**
XML (e.g. XES)	58	32	13	7	**112**
JSON	24	22	17	6	**70**
Relational format file (e.g. SQL bak, SQLite file)	15	21	11	1	**48**
Parquet	1	15	13	4	**34**
Database re-do logfiles	8	10	6	4	**29**
ORC (Optimized Row Columnar)	0	6	4	1	**11**
Avro	0	3	3	2	**9**

Fig. 7. Q6: 10 most used source data formats

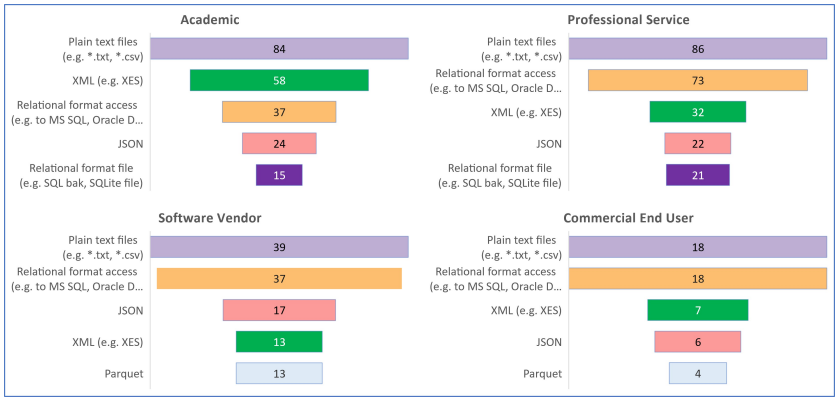
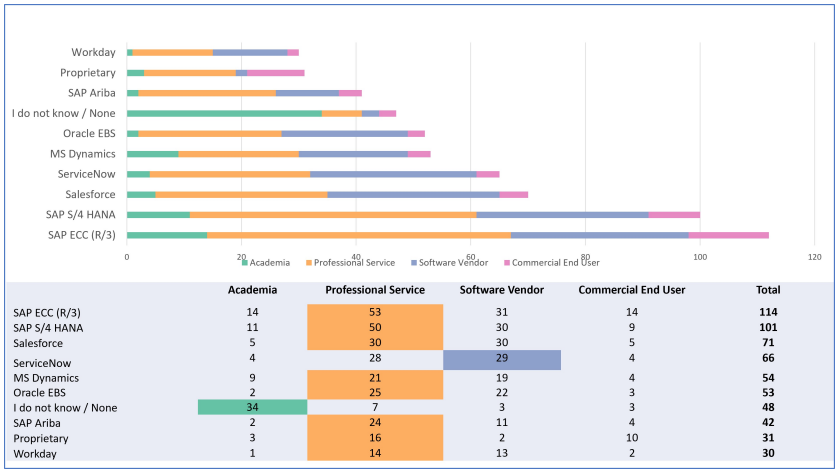


Fig. 8. Q6: 10 most used source data formats (role-wise)

Q9: To what extent did you encounter data-related challenges while undertaking PM projects in terms of sourcing data, processing data, and analysing process data? Figure 10 depicts an overview of sixteen data-related challenges identified across three categories: sourcing data, process data, and analysing data. The participants were asked to select a single option, ranging from none to very significant, for each data challenge.

Among the six challenges linked to the sourcing of process data, the challenge of complex data structures stands out as the most problematic, with $64\,\%$ of the participants ($n = 185$) selecting either significant or very significant. Moreover, $54\,\%$ of them selected the undocumented data structures as a key challenge (significant or very significant). On the other hand, $61\,\%$ indicated that the challenge of identifying the required data in the source systems as either moderate, minor or none, while $49\,\%$ felt the same about the challenge of exporting data from source systems.



	Academia	Professional Service	Software Vendor	Commercial End User	Total
SAP ECC (R/3)	14	53	31	14	**114**
SAP S/4 HANA	11	50	30	9	**101**
Salesforce	5	30	30	5	**71**
ServiceNow	4	28	29	4	**66**
MS Dynamics	9	21	19	4	**54**
Oracle EBS	2	25	22	3	**53**
I do not know / None	34	7	3	3	**48**
SAP Ariba	2	24	11	4	**42**
Proprietary	3	16	2	10	**31**
Workday	1	14	13	2	**30**

Fig. 9. Q7: 10 most used source systems

Among the five processing data-related challenges, $45\,\%$ ($n=140$) identified inconsistent data as being a relevant challenge (significant or very significant) while $42\,\%$ identified incomplete data as being a significant challenge. However, $75\,\%$ of all participants ($n=217$) indicated that the performance issues are not very significant by selecting either moderate, minor or none for that challenge.

Among the six data-related challenges linked to the analysis, the limitation related to analysing one-to-many and many-to-many relationships has been identified as a crucial challenge ($48\,\%$ selecting either significant or very significant) while $76\,\%$ indicated that exporting data from a process mining tool is the least significant challenge by selecting either moderate, minor, or none.

Q10: Which data-related challenges have you encountered beyond the ones listed in question 9? Figure 11 shows the frequency of the new challenges proposed by the participants. Among them, lack of documentation and data quality feature as the two top themes.

Q11: How could we speed up the data pre-processing to focus on analysis? The main themes identified from the responses ($n=199$) relate to the standardisation of data formats as well as data transformation pipelines, suggestions for better tool support, scaling up of domain and process mining expertise, and suggestions to improve data quality. Figure 12 captures the main themes with exemplar comments received from the participants.

Q12: How could a re-imagined industry-wide process mining data standard help you excel in your role? The participants foresaw a variety of potential benefits ranging from the acceleration of data pre-processing to commodisation of analysis ($n=156$) (see Fig. 13).

Discussion. The survey results reconfirm the common belief that the data pre-processing task is highly time consuming (with the maximum amount of effort estimated to be $90\,\%$) while $36\,\%$ estimated their efforts to be within the range of $60\,\%$ to $80\,\%$ (cf. Q3). The responses also confirm that the XML format (i.e., the one of XES) is not widely used in the community

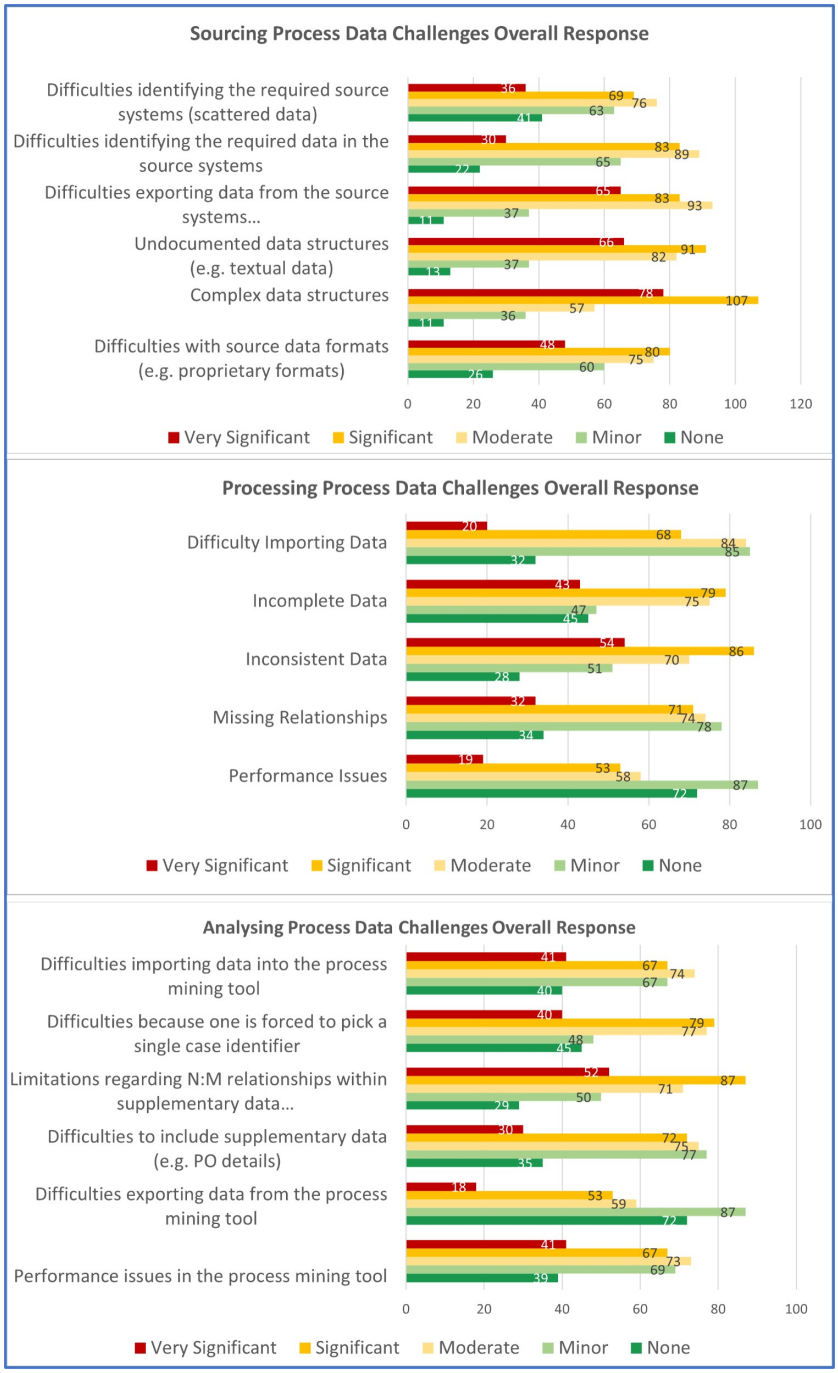


Fig. 10. Q9: Ranking the significance of data-related challenges

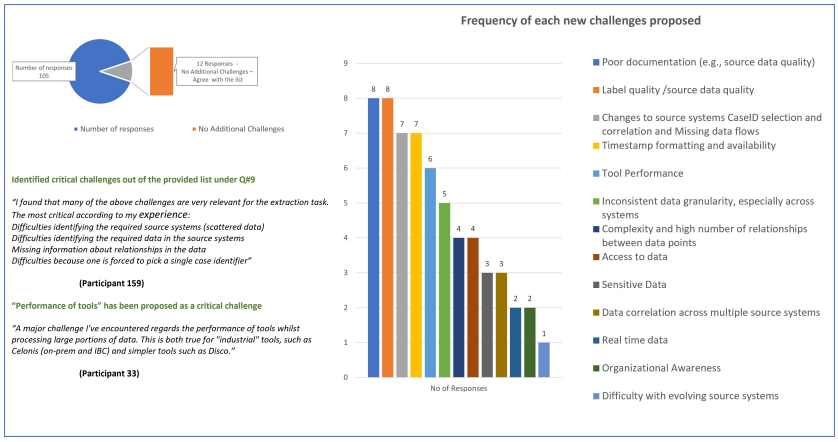
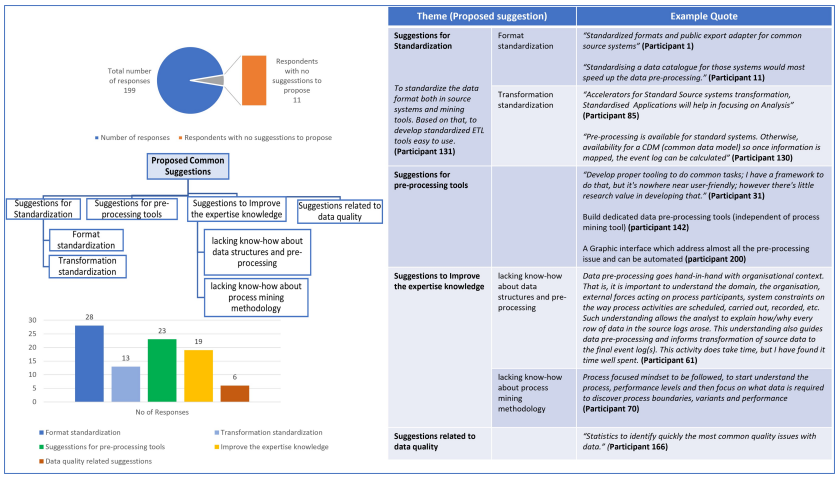


Fig. 11. Q10: Qualitative insights of other data-related challenges

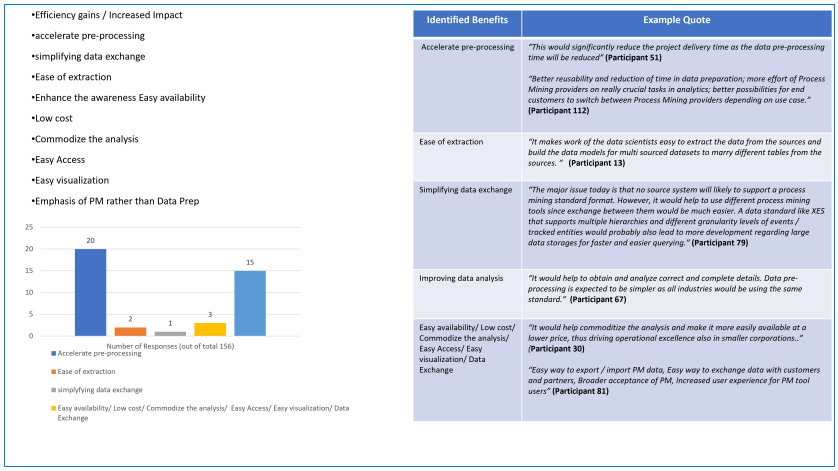


Theme (Proposed suggestion)		Example Quote
Suggestions for Standardization	Format standardization	*"Standardized formats and public export adapter for common source systems"* **(Participant 1)** *"Standardising a data catalogue for those systems would most speed up the data pre-processing."* **(Participant 11)**
To standardize the data format both in source systems and mining tools. Based on that, to develop standardized ETL tools easy to use. **(Participant 131)**	Transformation standardization	*"Accelerators for Standard Source systems transformation, Standardised Applications will help in focusing on Analysis"* **(Participant 85)** *"Pre-processing is available for standard systems. Otherwise, availability for a CDM (common data model) so once information is mapped, the event log can be calculated"* **(Participant 130)**
Suggestions for pre-processing tools		*"Develop proper tooling to do common tasks; I have a framework to do that, but it's nowhere near user-friendly; however there's little research value in developing that."* **(Participant 31)** Build dedicated data pre-processing tools (independent of process mining tool) **(participant 142)** A Graphic interface which address almost all the pre-processing issue and can be automated **(participant 200)**
Suggestions to Improve the expertise knowledge	lacking know-how about data structures and pre-processing	*Data pre-processing goes hand-in-hand with organisational context. That is, it is important to understand the domain, the organisation, external forces acting on process participants, system constraints on the way process activities are scheduled, carried out, recorded, etc. Such understanding allows the analyst to explain how/why every row of data in the source logs arose. This understanding also guides data pre-processing and informs transformation of source data to the final event log(s). This activity does take time, but I have found it time well spent.* **(Participant 61)**
	lacking know-how about process mining methodology	*Process focused mindset to be followed, to start understand the process, performance levels and then focus on what data is required to discover process boundaries, variants and performance* **(Participant 70)**
Suggestions related to data quality		*"Statistics to identify quickly the most common quality issues with data."* **(Participant 166)**

Fig. 12. Q11: Qualitative insights for suggestions to speed up the data pre-processing

to store event logs, with only 39 % selecting this option (cf. Q6). There seems to be consensus among the process mining community that there are significant data-related challenges associated with complex data structures, complex one-to-many and many-to-many relationships, inconsistent data, incomplete data and missing relationships (cf. Q9). These data challenges should be carefully considered and addressed when a new standard is being prepared.

The participants also indicated a need for systematic and automated data pre-processing techniques for efficient and reproducible data transformations for process mining. A dedicated methodology for data pre-processing to support a structured approach to PM methodology (*Stage 0*) seems definitely desirable, with the ability to create templates that capture best practices to ultimately speed up the data pre-processing task (cf. Q11). Approaches to assess and improve the data quality issues identified in the survey (e.g., inconsistent data or incom-



Identified Benefits	Example Quote
Accelerate pre-processing	*"This would significantly reduce the project delivery time as the data pre-processing time will be reduced"* **(Participant 51)** *"Better reusability and reduction of time in data preparation; more effort of Process Mining providers on really crucial tasks in analytics; better possibilities for end customers to switch between Process Mining providers depending on use case."* **(Participant 112)**
Ease of extraction	*"It makes work of the data scientists easy to extract the data from the sources and build the data models for multi sourced datasets to marry different tables from the sources. "* **(Participant 13)**
Simplifying data exchange	*"The major issue today is that no source system will likely to support a process mining standard format. However, it would help to use different process mining tools since exchange between them would be much easier. A data standard like XES that supports multiple hierarchies and different granularity levels of events / tracked entities would probably also lead to more development regarding large data storages for faster and easier querying."* **(Participant 79)**
Improving data analysis	*"It would help to obtain and analyze correct and complete details. Data pre-processing is expected to be simpler as all industries would be using the same standard."* **(Participant 67)**
Easy availability/ Low cost/ Commodize the analysis/ Easy Access/ Easy visualization/ Data Exchange	*"It would help commoditize the analysis and make it more easily available at a lower price, thus driving operational excellence also in smaller corporations.."* **(Participant 30)** *"Easy way to export / import PM data, Easy way to exchange data with customers and partners, Broader acceptance of PM, Increased user experience for PM tool users"* **(Participant 81)**

Fig. 13. Q12: Qualitative insights for the expected benefits of an industry-wide standard

plete data) could be beneficial. Furthermore, a new event log standard should leave room for various mechanisms to import/export the event data, not only using XML.

4 Adding context: reflections from the XES 2.0 workshop

In order to challenge and validate the survey's takeaways presented above, the XES WG hosted a workshop co-located with the Third Int. Conference on Process Mining in Eindhoven (Netherlands) on November 2, 2021. A session on survey results set the scene, followed by contributions from software vendors (represented by Celonis and Signavio), academia (represented by RWTH Aachen and the Free University of Bozen-Bolzano) and professional services (represented by KPMG). Concluding with a panel discussion centred around select findings from the survey, the workshop not only offered well-balanced viewpoints from different players in the discipline, but most notably revealed an unexpected homogeneity concerning the most relevant levers for a successful evolution of the XES standard.

First, rethink the core concepts of an event log. Numerous participants raised questions about the fundamental scope of what information is captured in event logs. The support of extensions render the current XES standard extremely flexible – even towards future, non-anticipated requirements – but at the cost of complexity. With limited awareness and use of existing extensions, this split needs to be revisited. It also became apparent that even though XES itself does not stipulate any storage format, most participants equate XES with its XML schema definition and call out its misfit with data volume and velocity of current, practical use cases. This showcases the need to strictly focus on storage-agnostic core concepts first and to later create multiple relevant reference implementations.

In addition, recent trends in industry and academia (e.g., object-centric event logs like OCEL, multi-event logs, and knowledge graphs) point to the need for complex data structures and relationships to be captured in an event log. A consensus has been reached to revisit the

core concepts in an event log and propose a conceptual data model alongside a metadata schema that can support complex data structures (including many-to-many relationships between multiple objects, cases, and events).

Event logs as a semantic layer. The current standard focuses mainly on syntactic interoperability and, to a lesser extent, on semantic aspects. However, enriching event logs with semantics would open up a whole array of possibilities across academia and industry (e.g., novel algorithms, autonomous data transformation, dynamic perspective change, real-time data extraction). Additionally, domain-specificity could tailor the semantics extensions to selective industries and thereby mimic real-world domain ontologies.

Taking this concept one step further, domain ontologies linked with event data could support process analytics without case identifiers. Different event logs could be generated as views over the same event data store. This intermediate layer would also hide the ultimate sources of the event data (let them be single or multiple, homogeneous or heterogeneous, legacy or newly implemented).

Generating momentum amongst industry players. Contributions, Q&A and panel discussion also evidenced an intrinsic challenge of generating momentum around an industry standard for interoperability. It is acknowledged that the current XES standard is hardly used in industry or professional services. Vendors see themselves in a balancing act with true interoperability on the one side, arguably a catalyst for the whole industry, and proprietary solutions on the other side, often attributed with preventing customer churn. In the end, the community needs to find ways to present all sides with compelling cases to not only support, but jointly design the next evolution of XES. Not only vendors of process analytics tools should be involved, but also those implementing systems for process execution. Their support could become the linchpin to propel the industry.

5 Conclusion

This paper presents a summary of findings from an online survey with 289 participants, who span across the roles of practitioners, researchers, software vendors and end-users. It also provides a synthesis of the discussion among the participants during the XES workshop at the International Conference on Process Mining 2021 and sketches the next steps for the XES WG.

Acknowledgements. The authors would like to thank all survey respondents who contributed their ideas presented in this paper. We would also like to thank the following presenters and panelists who participated during the XES workshop: Marco Montali, Philipp Hoch, Elham Ramezani, Björn Wagner, Sven Wagner-Boysen, Constantin Wehmschulte, and Sebastiaan van Zelst. The presentation slides are available at `www.tf-pm.org/resources/xes-standard/xes-2-0-workshop`.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016)
2. IEEE 1849 (XES) WG: IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. IEEE Std 1849-2016 (Nov 2016) 1–50