Business process mining in Banking Operations at the Swiss National Bank – A theoretical and empirical analysis of data and event log quality

Dr. Stefan Michel

stefan.michel@snb.ch

Zurich, January 2021

Contents

Abstract

The Swiss National Bank (SNB), as a commercially operated company, is also affected by digitisation. Chief among them are the numerous processes that are gradually being automated, which ultimately help to reduce costs and operational risks as well as increase efficiency in process management. Business process mining is an effective means to reconstruct and analyse process models based on log data. Such a procedure may thus also become important for the SNB when implementing future operational initiatives that aim to improve and automate existing tasks.

The author reviews the relevant literature, which shows that business process mining has been widely discussed since the start of the century. A key factor for the successful mining of reliable process models is high data quality. The main objective of this study is to use expert interviews to develop practical solutions to data-specific problems shown in an SNB case study analysis using real payment data. Such solutions to address data-specific problems have not yet been sufficiently covered in the academic literature. Furthermore, preventive measures to improve the quality of data and event logs are also discussed.

The results of the analysis show that the quality of the available data is appropriate for applying business process mining to operational processes in the core banking platform, Avaloq Banking System (ABS). The paper identifies a number of data-specific issues, e.g. the timestamp recorded in the event log does not reflect the actual time of the activity or the characteristics of individual attributes are only partly available in free-text format. These issues can be addressed by changing the system logging logic or by using robotics combined with artificial intelligence to create attribute designations and define categories. In addition, the paper points out that data quality for the mining application can also be approached preventively, for instance by performing history-based error analyses at system field level.

This study contributes to the existing literature by providing both theoretical and practical insights and including a solid understanding of the analysed business process, and by providing an overview of the literature regarding data quality problems and solutions.

Keywords: Business process mining, data and event log quality

+ This study has been carried out within the context of a doctoral thesis. The views expressed in this study are those of the author and do not necessarily reflect those of the SNB.

1. Introduction

Owing to rapid developments in the field of digitisation, greater market transparency and ever-growing customer requirements, flexibly and efficiently designed business processes are gaining in importance (cf. Schmiedel and Jessensky, 2015, p. 56). In its white paper, entitled *Data Age 2025*, the US market research and advisory services company, International Data Corporation (IDC), confirms the transition of analogue to digital data and predicts that the global datasphere will increase to a total of 163 zettabytes (1 zettabyte = 10,007 bytes) by 2025 (cf. IDC, 2017, pp. 2–3). That is equivalent to ten times the data generated in 2016 (cf. IDC, 2017, p. 3). While the amount of data generated globally has largely been consumer-driven up to now, IDC predicts that, in 2025, companies will account for nearly 60% of the global datasphere (cf. IDC, 2017, p. 21). According to Bose et al. (2013, p. 1), this increasing availability of data, in particular event data, is also the reason for the growing interest in business process mining, which can be used to derive process models based on event logs from information systems.

In their case study research paper, Mans et al. (2013, p. 7) conclude that, in addition to project-specific factors, mining-related aspects, such as data and event log quality, contribute to the success of a project. The authors go on to say that the model quality resulting from a process mining analysis depends greatly on data and event log quality and that this could increase the trustworthiness of the results (cf. 2013, p. 10). The success of a business process mining project thus hinges to a large extent on the quality of the input data, in other words on data and event log quality.

The following chapter addresses the topic relevance (chapter 1.1.), the positioning of business process mining in the context of other, existing technologies and management approaches (chapter 1.2.), as well as the potential structure and content of an event log (chapter 1.3.).

1.1. Relevance of the topic

1.1.1. Relevance for the SNB

As a commercially operated company, the SNB is just as affected by digitisation as the rest of the Swiss financial centre. Chief among them are the numerous processes that are gradually being automated, which ultimately help to reduce costs and operational risks, as well as increase efficiency in process management. Digitisation also presents the SNB with new challenges. Not only as a company, but also in the context of its objective to fulfil its statutory mandate and as an infrastructure manager. Take for instance the Banking Operations division, which is responsible for fulfilling said mandate in the area of cashless payment systems, for providing banking services to the Confederation and for ensuring the smooth settlement of all transactions conducted as part of the SNB's monetary and investment policy. Operational excellence is central to the fulfilment of operational tasks. Its purpose is to ensure an effective, relevant, efficient and risk-oriented conduct of business. By using business process mining in the future, the division could achieve additional transparency with regard to services, processes and resources to guarantee an effective and efficient fulfilment of tasks and operational management.

1.1.2. Relevance for other companies

It is evident from the examination of the topic that business process mining is also relevant for other companies. Take for instance a financial services company whose strategic decision to bring its organisation closer in line with the value chain for the customer also brought the processes into the focus of its analysis. According to said company, aligning an organisation with its processes calls for a process governance framework with clear roles, tasks, competencies and responsibilities, a common understanding of the process landscape within the organisational structure, clear and defined processes, as well as measurable procedures

for management and ongoing improvement. To achieve the latter two elements, the company employs business process mining. This facilitates a comparison of documented and as-is processes, the identification of process variances and process-specific bottlenecks, as well as the determination of indicators for process measurement and management. It also helps save time when identifying areas for improvement and reduces the need for maintenance in connection with process documentation.

Furthermore, with regard to procurement processes, the mining-analysis shows that a reduction in maverick buying, a higher degree of automation and fewer adjustments can be achieved. With a view to sales processes, it is evident that business process mining can reduce manual interventions, accelerate invoicing and generate higher e-business rates.

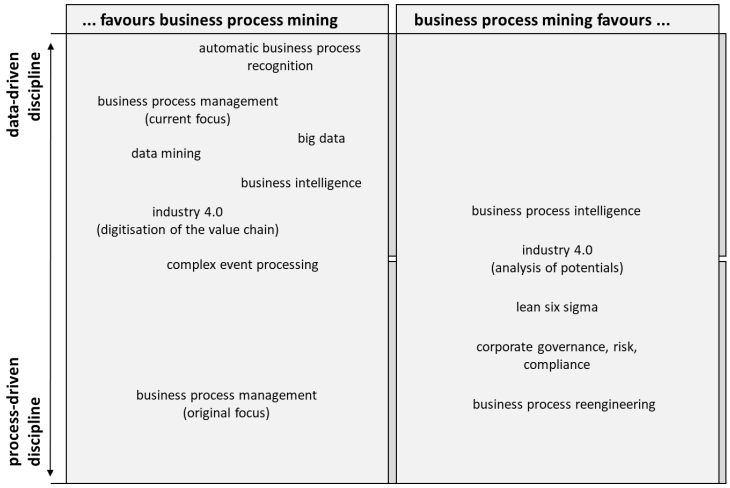
These observations indicate that business process mining, and hence process recognition and analysis, lends itself to being used in a business environment characterised by efficiency and effectiveness, irrespective of industry.

1.2. Positioning of business process mining

According to Klevers (2018, p. 2), business process mining is suitable for the analysis of both operational and administrative processes, for example in production, logistics, procurement, sales, finances and administration. Klevers (2018a, p. 7) further states that the idea behind business process mining is particularly relevant in cases where conventional process mapping, such as workshops with process experts, have been exhausted. This standard method of process mapping means that the current state of a process is recorded and considered as given (static mapping), that the survey is dependent on the compiler's interpretation, that correlations between the processes can usually only be regarded as general conditions and that, among other things, merely process states are recorded, which precludes any analysis of the development over time (cf. Klevers, 2018a, p. 7).

In his seminal book on process mining, van der Aalst states that the business process mining spectrum is broad and extends far beyond process discovery and conformance checking (cf. 2016, p. 44). Consequently, he notes that overlaps between business process mining objectives with those of other approaches, methodologies and principles are inevitable (cf. 2016, p. 44). Figure 1 thus shows – separately for process and data-driven disciplines – the thematic positioning of business process mining in the context of other, existing technologies and management approaches.

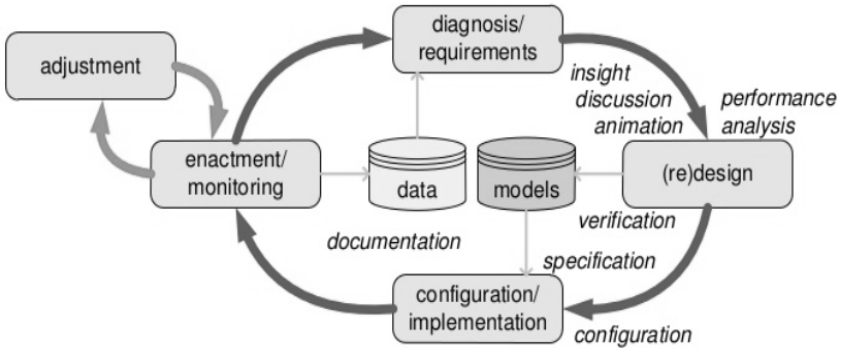
Figure 1: Positioning of business process mining in the context of other, existing technologies and management approaches



Source: Own illustration

The following section deals with the comparison between business process mining and business process management, which is particularly relevant for a general understanding of the subject. Business process management refers to the discipline that combines approaches for the design, execution, control, measurement and optimisation of business processes (cf. van der Aalst, 2016, p. 44). The relationship between business process mining and business process management is best illustrated by considering the business process management life cycle shown below (cf. figure 2). The main focus of business process management was initially on process design and implementation (cf. van der Aalst, 2013, pp. 2 et seq.); the discipline tended to be model-driven in its approach and disregarded the evidence hidden in the data. Domain experts are now increasingly focusing on the enactment, monitoring, adjustment, diagnosis and requirements phases, which are more data-driven and thus allow for the more frequent use of mining-techniques (cf. van der Aalst, 2016, p. 44). Even though van der Aalst (2016, pp. 44–45) maintains that business process mining should not be limited to business process management, since every process for which events can be recorded constitutes a mining candidate, the two disciplines are nonetheless closely related.

Figure 2: Life cycle of business process management



Source: van der Aalst, 2016, p. 31

1.3. Structure and content of an event log

The following notes on figure 3 are based on considerations by van der Aalst (2016, p. 35). In the figure below, the fragment of an exemplary event log shows how a compensation request is handled. Each line references one event – grouped by case (case id) – and each event is logged using a unique event id. The event id is used solely for identification purposes so as to distinguish it from other events that refer to the same activity.

The event log shows in condensed form that there are five events (event id: 35654423 to 35654427) associated with the first case (case id 1). The first event in the case – with the id 35654423 – corresponds to the execution of the 'register request' activity by Pete on 30 December 2010. The figure also shows that each event in a log has both a date and a timestamp, with the latter being more coarse or fine-grained depending on the log. Some logs also include information about the start and end times of an activity. The timestamps in the illustrated event log indicate the time of completion; the duration of an activity is not specified, however. Each event in this log is associated with a resource. This information is not available in every event log, or is not recorded with the same degree of granularity in every log. The figure below also provides the cost related to an event. Cost is just one example of a data attribute, of which there can be many. Other examples might include the amount of compensation requested or the outcome of the different types of checks. It should be possible to link events, identified using event id's for instance, to individual cases as well as to activities, and also to order them numerically within a case (cf. van der Aalst, 2016, p. 35). Thus, for the application of business process mining, activity designations, associated timestamps and case id's are required as a minimum.

Figure 3: Fragment of an exemplary event log

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...

Source: Own illustration, based on van der Aalst, 2016, p. 36

With regard to event logs, van der Aalst (2011, p. 99) makes the following assumptions:

- A process consists of cases.
- A case consists of events, with each event relating to precisely one case.
- Events within a case are ordered.
- Events can have attributes (e.g. timestamp, resources, cost).

2. Objectives and methodological approach

This study, which appears in the *SNB Economic Studies* publication series, examines the data and event log quality relevant for business process mining, with the aim of determining which empirically validated solutions exist to address those data-specific problems that have not yet been sufficiently considered in the academic literature. A case study analysis developed at the SNB's Banking Operations division shows, among other things, which data-specific challenges the division would face if business process mining were to be used in the future to settle operational processes, both in terms of data and model quality. This study also discusses preventive measures aimed at increasing the quality of data and event logs. Data and event log quality is established on the basis of the quality criteria for determining the maturity level of an event log as defined by the IEEE Task Force on Process Mining (2011, p. 7) (cf. Appendix I). It should be noted that data quality requirements depend largely on the data users, which is why it is not possible to generalise on what constitutes optimum data quality.

Drawing on the findings of a systematic, extensive study of the literature (primarily in the aggregator databases EBSCO and ProQuest), conclusions on data and model quality were made in the context of a case study analysis in the SNB's Banking Operations division, using real settlement data from a payment transaction process. Guided expert interviews were conducted in order to develop solutions for addressing data-specific problems and measures to improve the quality of data and event logs. Because business process mining is a relatively young and unexplored research discipline to be recognised, a predominantly qualitative and explorative approach was necessary to achieve the aforementioned objectives. This business-related, applied, snapshot study follows a theory-oriented, predominantly inductive approach and is based on the scientific theory of empiricism[1]. The results draw on both secondary and primary data sources and adopt a qualitative as well as a quantitative research approach.

[1] The author holds the view that sensory experience is the primary source of knowledge, as opposed to rational thinking or reason (theory of rationalism).

3. Literature about business process mining: A summarising overview

Despite intense research activity in the subject area, certain aspects of data quality issues remain neglected. The literature, most of which comes from Europe, clearly shows that business process mining is an emerging research discipline compared to other data analysis technologies, such as statistics or data mining. From 2000 onwards, the domain was explored in a comprehensive, largely consistent and academically sound manner, particularly with regard to the development of mining algorithms. As noted by Erdogan and Tarhan (2018, p. 1), business process mining is being used in many areas, including in manufacturing, healthcare, government and software development. However, the focus is primarily on an individual system and/or organisational unit. In healthcare, in particular, where most processes are complex, variable, dynamic and multidisciplinary in nature, the use of this approach is becoming increasingly challenging (cf. Erdogan and Tarhan, 2018, p. 1).

Business process mining is not a new concept. Cook and Wolf (1996, p. 3 and 1998, p. 35) already explored the topic more than twenty years ago in connection with software engineering processes, using an industrial case study. Mans et al. (2013a, p. 146) note that business process mining can be employed for a wide range of systems, such as administrative and logistics systems. Much of the research and development in this field is concentrated at Eindhoven University of Technology. In addition, the research is methodically approached primarily by means of case study analyses, although these are not limited to specific company processes. Moreover, it is clearly evident that data quality – and thus this research topic – has a direct impact on the quality of the analysis results and therefore plays a decisive role in making work processes more efficient.

According to Alves de Medeiros et al. (2003, p. 391), research in the business process mining domain is focused on mining heuristics, which are based on ordering relations of the events in the process log. Data-specific problems (frequently in connection with noise and flawed timestamp configurations) are generally only mentioned as an aside. Rozinat (2011, n.p.) addresses the key issues that can arise in mining projects. She mentions incorrect logging (*wrong data, or noise*), insufficient logging (*missing data, such as missing case id's or timestamps*), semantics (*finding and understanding the right information*), correlation (*processes often span several IT systems, and normally each system has its own local id's; these local process id's need to be correlated in order to combine the log fragments from the different systems*), and timing (*if timestamps are incorrect or not precise enough it can be difficult to create the correct order of events in the history*). Rozinat (2016, n.p.) also notes that there are a number of challenges in connection with data quality specific to business process mining, many of which relate to problems with timestamps. One could even say that timestamps are the 'Achilles heel' of data quality in business process mining (cf. Rozinat, 2016, n.p.).

According to Tiwari et al. (2008, p. 5), several studies address common mining problems such as noise and mining loops. That said, the majority of these papers are of a technical, mathematical nature and therefore only of limited use in practice. To resolve data-specific problems, the relevant literature often suggests reducing the data, in other words filtering out flawed data points. However, such an approach is regarded by the author as problematic, since it could give rise to new issues and potentially to misinterpretations. These kinds of implications, which solutions (like the exclusion of incomplete cases) can have on the meaningfulness of the generated models, are mostly ignored in the academic literature.

The data-centric solutions discussed in the literature are generally quite technical, such as using statistical inference to detect noise in event logs or employing a genetic algorithm to derive process models. The ones that are considered comprehensible and feasible for practitioners are not suitable because their implications are often unclear. Worth noting in this regard is, for instance, the removal of cases with poor data quality, which affects the representativeness of the dataset and thus the analysis options. Additional examples include the concepts of activity sequencing and activity hierarchy as posited by Dunkl (2013). It can be assumed that a significant amount of manual work is involved in implementing these concepts, which may no longer be

economically justified once a certain number of events need to be processed. Furthermore, it remains to be seen whether these concepts do not merely reduce data quality in favour of a readily understandable process map. The majority of the solutions identified are perceived to be abstract and barely comprehensible. One of the reasons for this may be that the proposed, data-specific solutions are often not put into practice.

Current research shows that business process mining and associated application techniques and algorithms have been widely discussed in the scientific community in recent years, with a high level of research activity and, judging by the latest findings, largely consistent. What emerges is that the discipline has not yet found widespread application in practice. Although the quality of the data used for mining projects is considered crucial for success, it has not been possible to find any academically sound literature that offers solutions for a broad set of data-specific problems.

4. Case study analysis

This chapter presents, in tabular form, the metadata deemed relevant for the case study analysis and describes the procedure for implementing the mining project.

Table 1: Metadata for case study analysis

Element	Element value
Case study design used	Single-case design with single unit of analysis (holistic design)
Duration of case study analysis	28.12.2017 to 08.11.2018
Business process to be analysed	Processing of outgoing payments on behalf of legal entity (third party)
Process type	Routine process
Modelling standard used for recording as-is processes	Business Process Model and Notation (BPMN) 2.0
Process scope	From receipt of payment message in ABS to when it is technically ready to be processed in payment system
Information systems used in relevant business process and its sub-processes	• ABS • SwiftRef • Microsoft Outlook • Skype for Business
Scope of application of data extraction	• Time dimension of extracted event data: Time dimension with hierarchy of year, month, day, hour, minute and second • Data extraction time period: 30.07.2018 (time: 00:00:01) to 07.09.2018 (time: 23:59:59): All recorded activities were considered for this time period • Data-link attribute Order number serving as case id
Data collection methods	• Workshops with project members • Document analyses • System documentation • Interviews with specialist and IT information providers

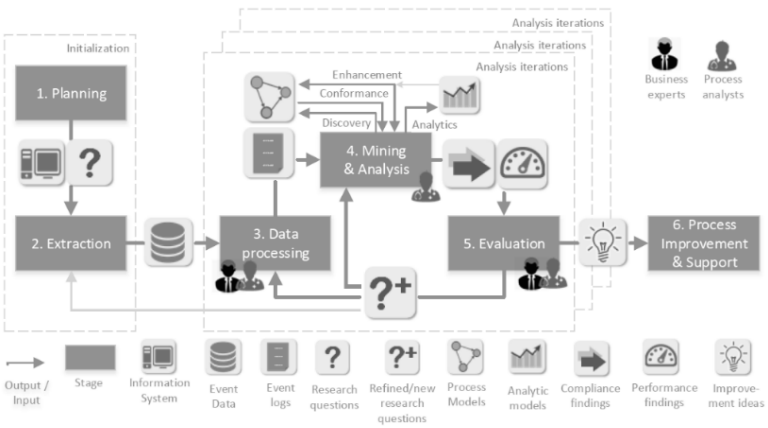
Steps used in data analysis phase (cf. Fluxicon, 2018, n.p.)	• Inspect data • Import data (into mining software) • Inspect process model • Inspect process statistics • Inspect cases • Visualise bottlenecks • Animate process
Mining toolkit used	Disco version 2.2.0 from Fluxicon (licence version: Academic)
Treatment of working days, weekends, public holidays in mining evaluation	Relevant for analysis were working days (Monday to Friday), 08.00–18.00; thus excluding weekends (Saturday and Sunday), without adjustment for public holidays (01.08.2018, Switzerland's national holiday)
Generated model type	Fuzzy model
Mining algorithm used	Based on fuzzy miner algorithm
Coding used for unicode sign	UTF-8

Source: Own illustration

The present findings are based on the PM^2 methodology, developed by van Eck et al. (2015) at Eindhoven University of Technology.

According to van Eck et al. (2015, pp. 1 et seq.), companies about to embark on a mining project can use PM^2 to plan and structure their procedure for improving process performance and compliance, for instance. The methodology consists of six stages (cf. figure 4). In stage 1 (planning), initial research questions are defined, while in stage 2 (extraction), event data are extracted. Once the first two stages have been completed, one or more analysis iterations are conducted, sometimes in parallel, with each iteration performing the following stages one or more times: data processing (stage 3), mining & analysis (stage 4) and evaluation (stage 5) (cf. van Eck et al. 2015, p. 3). An analysis iteration aims to answer a specific research question by applying process mining-related activities and evaluating the derived process model and, in some cases, other findings. Depending on the complexity of the mining & analysis phase, such an iteration may take anywhere from minutes to several days to complete. If the analysis results are satisfactory, the findings can then be used in stage 6 (process improvement & support) (cf. van Eck et al. 2015, p. 3). This study focused primarily on stages 1 through 4.

Figure 4: Overview of PM^2 methodology



Source: van Eck et al, 2015, p. 3

Upon studying the various methodologies, the author found that the description of the data preparation step was often too brief. The aforementioned stages 2 (extraction) and 3 (data processing) of the PM[2] methodology have therefore been extended to include the data preparation steps proposed by Aguirre et al. (2017, pp. 106 et seq.): data localisation, data extraction, data quality analysis, data cleansing and data transformation.

In order to identify the information systems and data points considered relevant for the business process analysed, the data architecture model illustrated in Appendix II was used. Starting at the process map level, it was thus possible to determine the application functions (AF) – such as entry masks, reports and order books in ABS – and sub-business functions (SBF/SGF) – such as system tasks and workflow steps – of the business processes being considered, so that the extract, transform, load (ETL) process could be carried out in a targeted manner.

5. Guided expert interviews

Based on theoretical sampling approach[2], the interview sample used for this study comprised nine interview participants from an equivalent number of organisations. Interview requests were sent to eleven institutions in total. The interviews were held between 28 December 2018 and 21 February 2019. The combined length of time required for all of the personally conducted expert interviews was around five hours, with the duration of the individual talks lasting between 39 and 70 minutes. Owing to the range of industries considered (including medical technology and financial services) as well as the executives interviewed (including senior auditor and senior business analyst), the sample is particularly meaningful. This heterogeneous sampling provided a detailed insight into the research area. The interview participants were selected according to Creswell's (2007, pp. 125 et seq.) approach of purposeful sampling. This is described as follows: "The concept of purposeful sampling is used in qualitative research. This means that the inquirer selects individuals and sites for study because they can purposefully inform an understanding of the research problem and central phenomenon in the study." (Creswell, 2007, p. 125)

The data evaluation approach roughly follows the steps of interview transcription and qualitative content analysis[3]. Using Creswell's (2013, pp. 184 et seq.) method of lean coding, a list with eight thematic groups (codes) was compiled after an initial review of the transcriptions so as to systematically categorise (code) the recordings from the expert interviews. These codes were then gradually expanded by repeatedly analysing the transcribed interviews. The final code system, created in MAXQDA, consists of eight codes and 143 associated codings.

6. Results

Chapter 6.1. deals with the descriptive statistics on the analysed event log data, while chapter 6.2.1. on data quality discusses whether the event data analysed here are relevant for the application of business process mining. The extent to which the execution graph – i.e. the process executed in the log-generating system – is able to depict the operational process (workflow graph) is addressed in chapter 6.2.2. on model quality. The purpose of dual quality evaluation (data and model quality) was to determine how closely event log data are based on operational processing and whether the use of business process mining is thus not only technically feasible, but also professionally sound. After concluding the quality evaluation of the workflow log (chapter

[2] As regards sampling, Merkens (2012, pp. 291–292) notes that samples can be selected using two different approaches. The sample can either be determined in advance of an analysis with regard to certain characteristics or be expanded and supplemented based on the findings obtained. The latter is referred to as theoretical sampling. Given that the sample for the expert interviews in this study was dependent on the findings of the case study analysis, the theoretical sampling approach was considered appropriate.

[3] Systematic interview transcriptions using the transcription rule system and audio transcription program, f4transkript. Transcription analysis with the qualitative data analysis software tool, MAXQDA Analytics Pro, based on Mayring's (2009, pp. 472–473) qualitative, summarising content analysis. This approach allowed the interviews to be efficiently categorised and further processed, and the main findings to be extracted from them.

6.2.3.), chapter 6.3. explains which practical solutions are available for tackling the data-specific problems identified. Practice-based measures that could improve data and event log quality are outlined in chapter 6.4., while chapter 6.5. closes with key findings regarding the implementation of a mining project.

6.1. Descriptive statistics on research-relevant event log data

The analysed event log (workflow log), which originated from ABS, contains the following descriptive parameters:

Table 2: Descriptive statistics for workflow log

Parameters	Value
Total number of events	23,317
Total number of cases	5,540
Total number of different activities	27
Median of all case times	13.4 hours
Average of all case times	14.9 hours
Extraction period	30.07.2018 (time: 00:00:01) to 07.09.2018 (time: 23:59:59)
Analysis period	30.07.2018 (time: 00:35:13) to 07.09.2018 (time: 15:13:51)

Source: Own illustration

The workflow log consists of the following, context-specific process properties. For analysis reasons, a distinction is made between mandatory data (*) and data that contain process-related context information (**):

Table 3: Logged process characteristics of workflow log

Classification	Characteristic	Example of characteristic
*	Order number	52208690 *(served as case id)*
**	Trade date	30.07.2018 *(order entry date)*
**	Value date	30.07.2018 *(value date)*
**	Order date	26.07.2018 *(date of creation of order number)*
*	Business type	Payment
*	Order type	Payment (incoming message) *(classification of order type)*
**	Payment method	Single customer transfer
**	Medium	Swift SCT
**	Payment channel	SIC
**	Post-it	:20 :Transaction reference number / :21 :Related reference / :79 :Narrative

**	Sequence number	13 *(systemically generated)*
*	Timestamp	30.07.2018 00:35:13
**	User	912102544 *(hash value, stands for a personal user and is used for protection of personal data)*
*	Activity	Approve (4-Eyes) (222180)
**	Old status	Ready to verify (2221)
**	New status	Ready to mail (80)

Source: Own illustration

In order to prepare the workflow log in its unprocessed form for import into the Disco mining toolkit, the following data manipulation steps were concluded:

Table 4: Data manipulation of workflow log

Element	Data manipulation
Order type	Data record filtering so that only data with order type 'Payment (incoming message)' was available (*this type indicates that an order was issued via SWIFT*)
Data anonymisation	Anonymisation of all customer data
Cell formatting	Formatting change of timestamp from type 'DD.MM.YYYY hh:mm' (example 07.09.2018 12:27) to type 'DD.MM.YYYY hh:mm:ss' (example 07.09.2018 12:27:52)
File save format	Format change from Excel workbook (*.xlsx) to Microsoft Excel CSV file (.csv)

Source: Own illustration

6.2. Evaluation of data and model quality of the workflow log

6.2.1. Evaluation of data quality of the workflow log

The following section aims to explain how the event log data in the workflow log can be used for the application of business process mining. It becomes evident that the analysed process leaves behind extensive data traces in ABS. For instance, table 2 shows clearly that in the process under review, 23,317 events, 5,540 cases and 27 different activities were completed over a period of only about six weeks.

The analysis of the data quality in the workflow log yields a predominantly positive result. This shows that the event log for the entire data extraction period comprises unique, time-stable case id's, which are structured in such a way that they can map a complete workflow. The time stability of order numbers, which contribute significantly to the validity of results and analyses, is ensured in ABS at all times, even in the event of potential order cancellations. Furthermore, it is not possible to manipulate these numbers, either by system users or by the system itself. In addition, the analysis reveals that none of the case id's in the dataset are able to falsify the analysis findings (such as test, development or dummy orders). This is guaranteed in accordance with the work environment instances stored in ABS. For system releases, the test procedure requires database administrators to first develop and test system functionalities on development instances. Subsequently, both automatic and user-controlled tests are conducted on the ABS test instance. Adjustments are only approved

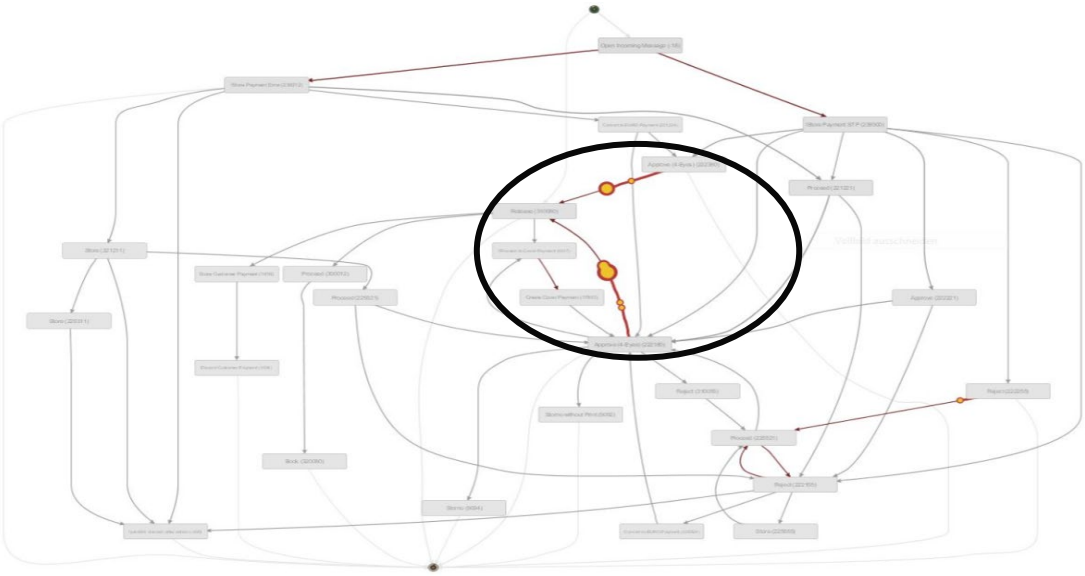
and transferred to the productive environment once development, system and user tests have been successfully completed. By means of this test procedure on different work environment instances, it can be technically ensured that the data extracted from the productive environment, as in this case, are also productive.

If the focus of the analysis is shifted to the timestamp elements, a similarly positive picture emerges. The timestamps in the workflow log are uniformly formatted and, in terms of their interpretation, uniquely assigned, in other words there is no mixing of activity start and end times. It is also evident that there are no missing or incomplete timestamps or dummy values (such as 01.01.1900 owing to the 1900 date system). What emerges, however, is that the timestamps have an inherent data-specific problem. This is discussed below, in the context of the data-specific issues observed in the case study analysis.

Regarding the quality of activity designations, it is noted that the designations logged in the workflow log (in this case, workflow activities) can be contextualised with the help of domain knowledge. This substantiates the quality of the log. However, it becomes clear during the analysis that, despite their consistency, activity or attribute designations need to be described as non-unique. This is also addressed in greater detail below.

The quality of the available data is, in principle, appropriate for the application of business process mining. On the basis of this finding and in the context of a self-assessment, it is noted that the event log used for this study can be assigned to the category ***, as per the quality criteria of the IEEE Task Force on Process Mining (2011, p. 7) (cf. Appendix I). In an event log of this quality, events are automatically recorded without a systematic approach. The IEEE Task Force states that, at this category level, there is some degree of guarantee that the logged events reflect reality, or in other words, that an event log is trustworthy, yet not necessarily complete (cf. 2011, p. 7). The reliability of the results and their proximity to reality are particularly evident in the process animation produced in Disco (cf. figure 5). According to the *Disco User Guide*, a process animation can be useful for communicating analysis results to process owners. Such animations demonstrate how cases move through the process (at their relative, actual speed) (cf. Rozinat, n.d., p. 67). Figure 5 illustrates, for example, the queue and wait times of payments with the ‘wait for trade date’ status that are necessary because the instructed payment execution date has not yet been reached and the workflow activity ‘release’ can therefore not yet be executed.

Figure 5: Fuzzy model: Process animation based on workflow log
(100% process activities and paths; model with absolute case frequencies)



Source: Own illustration

The analysis of the workflow log's data quality also reveals data-specific problems; these are described below and are supplemented with the implications for potential future log analyses.

Time recorded in event log does not reflect actual time of activity

- **Description of problem:** The timestamp stored in the event log for each activity shows when a field in the ABS input mask was saved. However, it does not provide any information about when an employee carried out an activity in the underlying process. In other words, it is not a timestamp that allows conclusions to be drawn about the actual activities of employees, it is merely one that shows when data were saved in the system.
- **Implications:** The event log analysed thus essentially documents steps that may have been performed at an earlier point in time. Needless to say, this complex, data-specific problem leads to distortions in log analyses and resulting key performance indicators (for example, with regard to lead and wait times).

Existence of attribute designations in free-text format

- **Description of problem:** The 'Post-it' functionality was included in the data extraction process. While not considered essential for the case study, this functionality can provide process-related context information for potential future analyses, thereby enhancing the settlement of a payment order with additional information. It could be used, for instance, to electronically document queries made in connection with a pending payment order, owing to incorrect debit or credit account details, for example.
- **Implications:** According to Brucker-Kley et al. (2018, p. 53), the inclusion of an optional free-text field offers scope for content to be recorded in an unstructured, unsystematic manner, with the result that the entries may be lacking in accuracy and/or completeness. Depending on the focus of the analysis, this could lead to too many variations for an analysis algorithm, thus making meaningful evaluations impossible (cf. Brucker-Kley et al., 2018, p. 53). By way of example, this implication is illustrated as follows: If one and the same property in a process event is initially logged once with and once without spaces, this would constitute two different process tasks should the field be classified as an activity in the mining software. An accumulation of such cases ultimately makes a mining analysis more complex and time consuming.

Workflow activities with identical main designations, but different add-ons

- **Description of problem:** The analysis shows that different workflow activities are recorded in the workflow log with the same main designation. It is also found that each designation contains additional descriptions, which are both specific and technically justified. For instance, the workflow activity 'Approve (4-Eyes)' was given both the activity designation 'Approve (4-Eyes) (222180)' and 'Approve (4-Eyes) (222380)'.
- **Implications:** An in-depth analysis by database administrators revealed that the additional technical descriptions vary depending on the processing stage of a payment order, and that different system commands are therefore executed. The main implication of this problem is that a raw dataset containing duplicate or very similar activity designations cannot be interpreted and contextualised without further analysis due to their potential ambiguity. It can be assumed that mining algorithms will have difficulty identifying duplicate activities in the first place. If such an identification is not possible or only possible with errors, a subsequent analysis would produce results that have a high potential for wrong conclusions to be drawn.

Events occurring in reality, but not recorded in event log

- **Description of problem:** This quality issue, as described by Bose et al. (2013, p. 8), refers to a scenario where one or more events are missing in the trace, even though they occurred in reality. This pattern can also be seen in the dataset underlying the workflow log. Based on the above, the event data in the workflow log that refer to status information pertain only to the technical workflows stored in ABS, meaning that actually occurring, technical process activities remain disregarded.
- **Implications:** As a result of this problem, it is not possible to determine which technical process steps need to be carried out by members of the Payments unit in order to initiate technical workflow activities. Bose et al. (2013, p. 8) note that the effect of missing events is that there may be problems with the results produced by a mining algorithm. In particular, relations may be inferred which hardly exist or do not exist at all in reality (cf. Bose et al., 2013, p. 8).

6.2.2. Evaluation of model quality of the workflow log

This section addresses the question of the extent to which the process model derived from event log data (execution graph) is able to represent the process model used in practice (workflow graph). From these findings, it will then be possible to determine how closely event log data are based on operational processing and whether the application of business process mining is not only technically possible, as already observed, but also professionally sound.

In order to obtain answers to the above, the workflow log first had to be thoroughly revised. It was found that the activities recorded in the log were too fine-grained to make an informed comparison between the execution and workflow graphs. For this reason, all the designations of thematically related activities were standardised, thus reducing the complexity of the analysis. After this data manipulation, the event log only comprised workflow activities that were semantically unique.

The first analysis scenario (analysis level 1) involved examining whether process activities can be transferred thematically from the execution graph to the workflow graph (and vice versa). This scenario represents the ideal situation in that it allows the conclusion to be drawn that the processes documented and practised in an organisation are congruent. However, given that this scenario did not materialise, two further analysis levels were introduced. While analysis level 2 assumes that the execution graph does not consider all the process activities documented in the workflow graph, analysis level 3 focuses on the contrasting situation, namely that process activities are missing in the workflow graph.

Hypothesis examined by analysis level 2:

The execution graph does not consider all process activities documented in the workflow graph

A previously discussed finding emerges from the analysis regarding the completeness of the execution graph. It appears that the event data referring to status information pertain only to the technical workflows stored in ABS, meaning that technical process activities – and thus the business context – remain disregarded. This leads to the realisation that, on the basis of the execution graph, it is not possible to determine which technical process steps need to be carried out, for example to change the technical workflow status of a payment transaction from ‘Reject’ to ‘Proceed’.

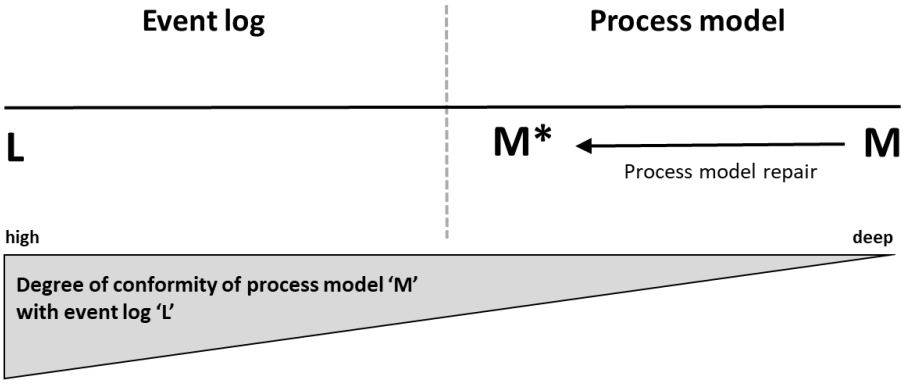
Hypothesis examined by analysis level 3:

Process activities are missing in the workflow graph

The findings from the execution graph and analysis of the model quality on analysis level 2 show that even the operational process to be used in practice (workflow graph) was incomplete in its modelling. There were four process activities missing in this process that had become transparent in the execution graph.

Carmona et al. (2018, p. 229) note in this respect that if deviations emerge between a process model and event log, they may be due to the fact that the current process model failed to represent the real process as recorded in the event data. The authors go on to say that process models should be modified so as to better reflect reality (cf. 2018, p. 229). Conceptually, an associated process model repair – as illustrated in figure 6 – attempts to bring process model 'M' closer to event log 'L'. It is hoped that, after the repair, the resulting process model 'M*' is more accurately aligned with respect to event log 'L' than the original model 'M' (cf. Carmona et al., 2018, p. 229).

Figure 6: Process model repair to bring model 'M' closer to log 'L'



Source: Own illustration, based on Carmona et al., 2018, p. 230

Based on the model quality analyses, it is noted that the process model derived from event log data, which takes into account all process activities and paths for the data extraction period in question, can only partially represent the process model used in practice. As previously mentioned, this is attributable to the fact that the event data referring to status information pertain only to the technical workflows stored in ABS. Even if the event data used here cannot fully reflect the technical process, the execution graph still provides a valuable basis for discussion. This allowed the process owner to improve their understanding of the process and to integrate the findings of the analysis into the workflow graph.

6.2.3. Conclusion regarding the quality evaluation of the workflow log

The quality of the extracted event data in the workflow log is, in principle, considered appropriate for applying business process mining in the future to ABS-based operational processing at the SNB, for instance to create process transparency. This notwithstanding, data-specific problems were identified in the case study analysis. The analysis of the model quality, in turn, shows that the process model derived from event log data can only partially represent the model used in practice. This is attributable to the fact that the event data referring to status information pertain only to the technical workflows stored in ABS. Technical process steps thus remain disregarded in the execution graph. It is therefore noted that, in this instance, business process mining is considered as a complement and not a substitute for conventional process recordings.

6.3. Solutions for case study-specific data quality problems

In accordance with the above findings, the SNB's Banking Operations division would face four data-specific problems in the event of business process mining being used in the future for ABS-based payment transaction processes: incorrect timestamps, duplicate activities/imprecise activity designations, missing events, and attribute designations in free-text format. It also becomes evident that the current academic literature

provides very little in the way of applicable solutions to these problems. Figure 7 below presents practical approaches to solving the aforementioned problems; these were collected from the expert interviews.

Figure 7: Practice-based approaches to solving data-specific problems

Some practical solutions for addressing incorrect timestamps:

- Assessment of whether data-specific problems are relevant for a mining analysis (if yes, acceptance of the situation and abandonment of a mining analysis)
- Use of specific logging software (e.g. open source web analytics platform 'Matomo')
- Change of the system/logging logic, for example by entering several date points in the system
- Performing manual observations and/or time measurements

Some practical solutions for addressing duplicate activities/imprecise activity names:

- Application of activity aggregation
- Use of the discipline of ontology in conjunction with domain knowledge
- Omission of affected activities via filter function in the mining toolkit
- System adaptations for more meaningful loggings

Some practical solutions for addressing missing events:

- Accept the situation and communicate the problem to the users
- Use of source-independent, orchestrating workflow engine and processing of process steps via recordable channels
- Creation of mandatory rules and standards in process execution
- Adaptation of the archiving logic of the data warehouse
- Connection of different, also manually recorded protocols
- Implementation of employee training on data-creating process flows

Some practical solutions for addressing attribute designations in free-text format:

- Do not allow free-texts (implementation of corresponding system adaptations)
- Creation of system fields with clear entry specifications as well as a drop-down or drill-down menu with standardised selection options
- Application platforms for artificial intelligence and pattern recognition
- Use of robotic process automation in combination with artificial intelligence (to create attribute names and to define categories)
- Use of robots for text recognition and database input
- Creation of a plausibility check program for system input checking
- Application of search filter for clustering keywords in free-text

Source: Own illustration

6.4. Practice-based measures for enhanced data and event log quality

The conclusion from the guided expert interviews was that there is no 'one size fits all' solution for improving data and event log quality. One respondent, for example, suggested conducting a history-based error analysis at field level in the mining source to determine which fields (e.g. customer addresses) had frequently been entered incorrectly in the past. Based on the findings of such an analysis, the fields in question could subsequently have specific rules attached to them, such as making the 'street' field mandatory for every address. To enable the results of this error analysis to be compared over an extended period of time, the respondent proposed running overnight evaluations, which could then be presented in a data quality report and compared using sigma values.

In addition, it emerged that the establishment and definition of dedicated teams (such as master data and data processing units) or roles (data quality managers) can also contribute to improving data and event log quality.

Furthermore, it was clear from the expert interviews that improving data and event log quality is an ongoing process. One respondent noted that a proactive approach to data quality is crucial, as otherwise one runs the risk of losing valuable time for data preparation in the event of a mining use case. Another respondent stated that such a process could be facilitated, for example, by means of data quality workshops, where particularly the dos and don'ts of data management are conveyed.

It is further observed by respondents that the appropriate way to enhance data and event log quality is to improve the way in which process steps are registered and data are stored. This should enable more realistic logging. Following this line of thinking, it becomes evident that also dispensing as much as possible with free-text entry options and conducting data validation meetings with process and domain experts could serve as preventive measures in achieving more targeted data and event log quality.

One respondent concluded that a selection based on a specific and not-too-high number of process activities to be analysed is deemed helpful in ensuring data quality, since a dataset can hardly be evaluated beyond a certain variance. The respondent held the view that by reducing the number of activities considered, data homogeneity (similarity of data) can be increased and the variance of underlying data points decreased. The author finds this observation interesting, but questions it critically, given the fact that, by consciously selecting data, activities and contextual information which are relevant for analysis and interpretation may remain disregarded.

6.5. Findings with a focus on implementing a mining project

One key finding is that a mining project necessitates an interdisciplinary project team (to conduct data validation meetings, for instance), as the application of business process mining not only requires data knowledge, but also process skills. In this case study, without expert data validation, for example, it would scarcely have been possible to establish that the 'Order type' extraction field serves as a criterion for identifying payment orders that are issued via SWIFT or via physical payment orders. To illustrate the relevance of technical data validation, it is noted that without it, the distinction between the various workflow activities could not have been made. Only after an in-depth analysis by database administrators was it established that activities are given different add-ons depending on the processing stage of a payment order. This kind of insight would hardly have been possible without business process mining, since the individual activities would not have been accessible in such an aggregated, visually comprehensible form. Software engineers, database administrators and domain experts, among others, are thus relevant sources in identifying and resolving data quality problems.

In addition, it becomes evident how important it is to choose a process of manageable complexity at the start of a mining analysis, with a view to gaining initial experience in the area; a process where both the specialist and IT units share a common knowledge base.

The analysis also points to the enormous potential of combining data-centric and process-centric analysis techniques. Data analysis without process linking makes little sense, whereas process analysis without a data foundation is not very effective. This leads to the conclusion that processes and data should be analysed together and in line with each other.

In order to fill all the relevant project roles at the beginning of the case study analysis, it proved valuable to be able to demonstrate the discipline and its benefits using specific and readily understandable examples. If the

data extraction stage had been commenced too early, incorrectly selected data might have called for time-consuming adjustments that would have done little to inspire confidence in the analysis.

To conclude, the most significant challenges addressed in the case study analysis are discussed. The challenge during the data acquisition process was to identify the relevant data fields and gain an understanding of the elements that influence the process flow (e.g. timestamps). Moreover, defining a representative data extraction period that could only consider event data of completed cases was found to be complex. In addition, selecting the case study process to be used necessitated internal coordination between process owner, process employee and business process mining method expert. A further challenge to emerge was the handling of data with different degrees of granularity. These challenges were tackled by choosing a case study process that was understood by both the specialist and IT units and for which data could be technically extracted.

7. Future research

This study examines data and event log quality relevant to business process mining. The following observation by Rozinat (2011, n.p.), which hints at the complexity inherent in a mining project, aptly illustrates the need for further research:

> *"In fact, figuring out the semantics of existing IT logs can be anything between really easy and incredibly complicated. It largely depends on how distant the logs are from the actual business logic. For example, the performed business process steps may be recorded directly with their activity name, or you might need a mapping between some kind of cryptic action code and the actual business activity."*

While the existing mining methods indicate the importance of data quality, they provide no details about how the identification of data quality problems can be addressed in the planning, data extraction and logging stages of a process (a similar reference is also found in Andrews et al. (2018, p. 1)). Research in this area could enhance the acceptance of the discipline.

It is also evident that business process mining tends to be backward-looking (cf. van der Aalst, 2018, p. 1). In order for business process mining to prevail in the future, however, in-depth research is needed on potential design alternatives, such as real-time business process mining or process simulation using artificial intelligence.

Furthermore, there is a lack of research on how to determine when an organisation is 'fit' for the application of business process mining. In this respect, it would be beneficial from a practical point of view to draw up a checklist with specific questions that would allow a 'readiness assessment' to be conducted in advance of a mining project. Possible questions that could be included are: Does a mining project have the support of management? Do staff members have enough knowledge of the discipline to conduct a mining project? Is the quality of the underlying data sufficiently good or are the IT systems sufficiently advanced in terms of their data organisation to enable mining analyses? Do the specialist and IT units have enough resources to conduct a mining project?

It is also noted that the identification of data quality issues in event logs is not always straightforward and requires a methodical approach. In the context of this study, for instance, every issue identified in the academic literature was applied to the workflow log – in other words, the question was asked whether such an issue was visible in the event log. However, in future research on the subject, it would be more expedient to

develop methodologies, query languages or, for example, algorithms that support the systematic identification of data quality issues. This could substantially reduce the amount of time required for a mining analysis.

To conclude, further research is needed for an in-depth examination of the business objectives inherent in the discipline. Among other things, the relevance of the topic could also be further consolidated from a management theory perspective by examining the implications of using business process mining on decision-making ability or quality.

References

Aguirre, Santiago, Carlos Parra and Marcos Sepulveda (2017), methodological proposal for process mining projects, *International Journal of Business Process Integration and Management*, 8 (2), pp. 102–113.

Alves de Medeiros, Ana K., Wil M. P. van der Aalst and Ton Weijters (2003), *workflow mining: current status and future directions*, Research publication, Eindhoven University of Technology.

Andrews, Robert, Moe T. Wynn, Kirsten Vallmuur,, Arthur H. M. ter Hofstede, Emma Bosley, Mark Elcock and Stephen Rashford (2018), *pre-hospital retrieval and transport of road trauma patients in queensland. a process mining analysis,* Research publication, Queensland University of Technology.

Bose, Jagadeesh C., Ronny S. Mans and Wil M. P. van der Aalst, (2013), *wanna improve process mining results? it's high time we consider data quality issues seriously,* Research publication, Eindhoven University of Technology.

Brucker-Kley, Elke, Denisa Kykalová, Léonard Studer and Ines Leario (2018), fallstudie stadt lausanne: prozessanalyse in kombination mit process mining, in Elke Brucker-Kley, Denisa Kykalová and Thomas Keller (eds.), *prozessintelligenz – business-process-management-studie – status quo und erfolgsmuster*, Berlin, Heidelberg: Springer, pp. 45–59.

Carmona, Josep, Boudewijn van Dongen, Andreas Solti and Matthias Weidlich (2018), improving processes using conformance checking, in Josep Carmona, Boudewijn van Dongen, Andreas Solti and Matthias Weidlich (eds.), *conformance checking. relating processes and models*, Cham: Springer, pp. 229–240.

Cook, Jonathan E. and Alexander L. Wolf (1996), *discovering models of software processes from event-based data*, Research publication, University of Colorado.

Cook, Jonathan E. and Alexander L. Wolf (1998), *event-based detection of concurrency,* Research publication, New Mexico State University and University of Colorado.

Creswell, John W. (2007), *qualitative inquiry and research design: choosing among five approaches,* Thousand Oaks: Sage Publications, Inc.

Creswell, John W. (2013), *qualitative inquiry and research design: choosing among five approaches,* Thousand Oaks: Sage Publications, Inc.

Dunkl, Reinhold (2013), data improvement to enable process mining on integrated non-log data sources, in Roberto Moreno-Diaz, Franz Pichler and Alexis Quesada-Arencibia (eds.), *Computer Aided Systems Theory – EUROCAST 2013*, Berlin, Heidelberg: Springer, pp. 491–498.

Erdogan, Tugba G. and Ayça Tarhan (2018), *systematic mapping of process mining studies in healthcare,* Research publication, Hacettepe University.

Fluxicon (2018), hands-on tutorial, Online edition, www.fluxicon.com/book/read/tutorial/.

IDC (2017), data age 2025: the evolution of data to life-critical. don't focus on big data; focus on the data that's big, Online edition, www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf [requested on: 27.04.2018].

IEEE Task Force on Process Mining (2011), process mining manifesto, Online edition, www.win.tue.nl/ieeetfpm/downloads/Process%20Mining%20Manifesto.pdf.

Klevers, Markus (2018), *innovative optimierung ihrer geschäftsprozesse mit process mining,* Seminar documents*,* GEPRO mbH München.

Klevers, Markus (2018a), *process mining intensivseminar. einführung,* Seminar documents, GEPRO mbH München.

Mans, Ronny, Hajo Reijers, Hans Berends, Wasana Bandara and Rogier Prince (2013), *business process mining success,* Conference paper, Eindhoven University of Technology.

Mans, Ronny S., Wil M. P. van der Aalst, Rob J. B. Vanwersch and Arnold J. Moleman (2013a), process mining in healthcare: data challenges when answering frequently posed questions, in Richard Lenz, Silvia Miksch, Mor Peleg, Manfred Reichert, David Riaño and Annette Ten Teije (eds.), *process support and knowledge representation in health care*, Berlin, Heidelberg: Springer, pp. 140–153.

Mayring, Philipp (2009), qualitative inhaltsanalyse, in Uwe Flick, Ernst von Kardorff and Ines Steinke (eds.), *qualitative forschung. ein handbuch*, Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag, pp. 468–475.

Merkens, Hans (2012), auswahlverfahren, sampling, fallkonstruktion, in Uwe Flick, Ernst von Kardorff and Ines Steinke (eds.), *qualitative forschung. ein handbuch,* Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag, pp. 286–299.

Rozinat, Anne (n.d.), disco user guide, Online edition, www.fluxicon.com/disco/files/Disco-User-Guide.pdf.

Rozinat, Anne (2011), top 5 data quality problems for process mining, Online edition, www.fluxicon.com/blog/2011/06/data-quality-process-mining/.

Rozinat, Anne (2016), data quality problems in process mining and what to do about them – part 1: formatting errors, Online edition, www.fluxicon.com/blog/2016/01/data-quality-problems-in-process-mining-and-what-to-do-about-them-part-1-formatting-errors/.

Schmiedel, Danilo and Olaf Jessensky (2015), geschäftsprozesse identifizieren, prüfen und verbessern. process mining in der praxis, Online edition, www.opitz-consulting.com/fileadmin/user_upload/Collaterals/Artikel/business-technology-2015-02_process-mining_schmiedel-jessensky_sicher.pdf.

Tiwari, Ashutosh, Chris J. Turner, and Basim Majeed (2008), a review of business process mining: state-of-the-art and future trends, *Business Process Management Journal,* 14 (1), pp. 5–22.

Van der Aalst, Wil M. P. (2011), *process mining: discovery, conformance and enhancement of business processes,* Berlin, Heidelberg: Springer.

Van der Aalst, Wil M. P. (2013), business process management: a comprehensive survey, *ISRN Software Engineering*, 2013 (Article ID 507984), pp. 1–37.

Van der Aalst, Wil (2016), *process mining. data science in action,* Berlin, Heidelberg: Springer.

Van der Aalst, Wil M. P. (2018), *process mining and simulation: a match made in heaven,* Research publication, RWTH Aachen University.

Van Eck, Maikel L., Xixi Lu, Sander J. J. Leemans and Wil M. P. van der Aalst (2015), *pm^2: a process mining project methodology,* Research publication, Eindhoven University of Technology.

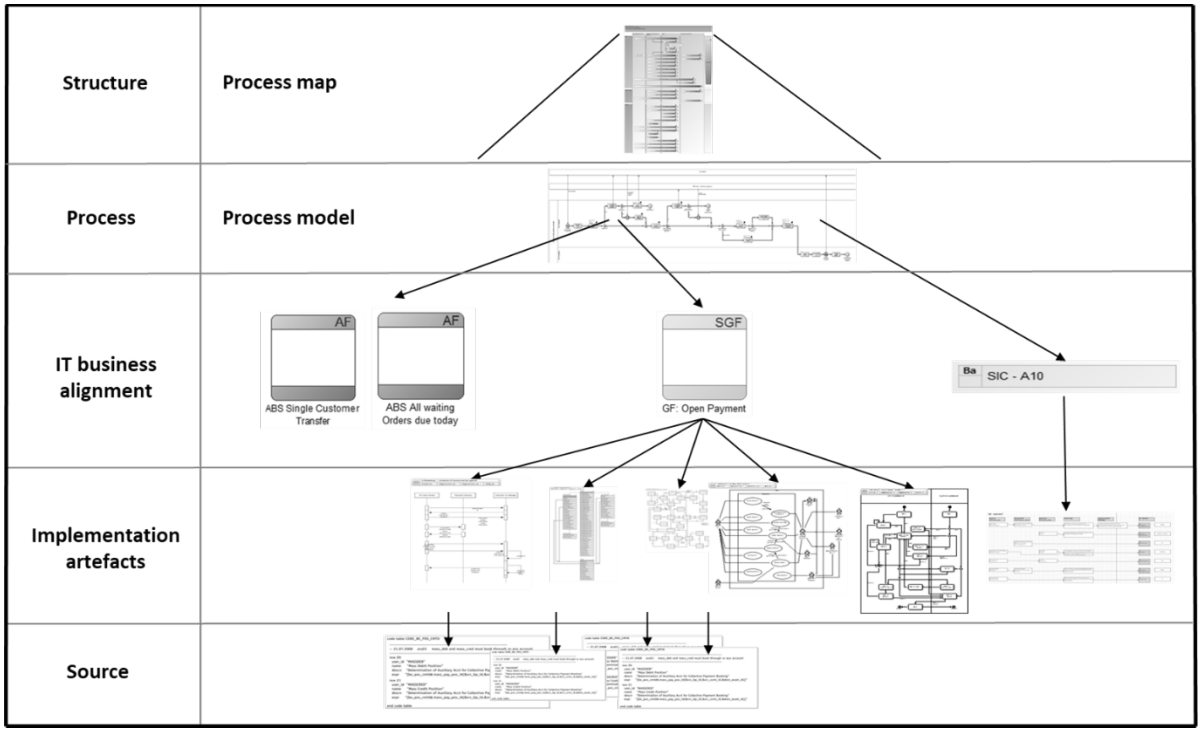
Appendix

I. Maturity levels for event logs

Level	Characterization	Examples
★★★★★	Highest level: the event log is of excellent quality (i.e., trustworthy and complete) and events are well-defined. Events are recorded in an automatic, systematic, reliable, and safe manner. Privacy and security considerations are addressed adequately. Moreover, the events recorded (and all of their attributes) have clear semantics. This implies the existence of one or more ontologies. Events and their attributes point to this ontology.	Semantically annotated logs of BPM systems.
★★★★	Events are recorded automatically and in a systematic and reliable manner, i.e., logs are trustworthy and complete. Unlike the systems operating at level ★★★, notions such as process instance (case) and activity are supported in an explicit manner.	Events logs of traditional BPM/ workflow systems.
★★★	Events are recorded automatically, but no systematic approach is followed to record events. However, unlike logs at level ★★, there is some level of guarantee that the events recorded match reality (i.e., the event log is trustworthy but not necessarily complete). Consider, for example, the events recorded by an ERP system. Although events need to be extracted from a variety of tables, the information can be assumed to be correct (e.g., it is safe to assume that a payment recorded by the ERP actually exists and vice versa).	Tables in ERP systems, event logs of CRM systems, transaction logs of messaging systems, event logs of high-tech systems, etc.
★★	Events are recorded automatically, i.e., as a by-product of some information system. Coverage varies, i.e., no systematic approach is followed to decide which events are recorded. Moreover, it is possible to bypass the information system. Hence, events may be missing or not recorded properly.	Event logs of document and product management systems, error logs of embedded systems, worksheets of service engineers, etc.
★	Lowest level: event logs are of poor quality. Recorded events may not correspond to reality and events may be missing. Event logs for which events are recorded by hand typically have such characteristics.	Trails left in paper documents routed through the organization ("yellow notes"), paper-based medical records, etc.

Source: IEEE Task Force on Process Mining, 2011, p. 7

II. Data architecture model



Source: Own illustration