A practitioner's guide for process mining on ERP systems the case of SAP order to cash

Roest, A.

Award date:
2012

Eindhoven, 22 November 2012

A PRACTITIONER'S GUIDE FOR PROCESS MINING ON ERP SYSTEMS – THE CASE OF SAP ORDER TO CASH

A. (Harmen) ROEST

BSc Industrial Engineering — TU/e 2011
Student identity number 0608628

in partial fulfilment of the requirements for the degree of

Master of Science
in Operations Management and Logistics

Supervisors:
dr.ir. R.M. (Remco) Dijkman, TU/e, IS
dr. M. (Marco) Comuzzi, TU/e, IS
H.J.R. (Dennis) van de Wiel MSc RE, KPMG IT Advisory

TUE. School of Industrial Engineering.
Series Master Theses Operations Management and Logistics

Subject headings: Process mining, Process analysis, SAP, Sales Process

ABSTRACT

This research presents a structured approach for practitioners to perform process mining on an ERP system. The generation of an event log is the first and very important step that practitioners have to take. Generating an event log poses technical and conceptual challenges. The approach has an emphasis on the design and generation of an event log. Furthermore it guides the practitioner in analyzing the event log and generating business value from it.

Through a literature study, an initial version of the approach is developed. It is further developed and specified through literature research and field research at KPMG IT Advisory. Combining literature and field research highlights the practitioner's perspective, but gives it a theoretical foundation. The approach is implemented and tested for SAP Order to Cash (OtC). The implementation results in an event log extraction script that is applicable on standard SAP implementations. The approach is tested in a case study with a Dutch based publisher of educational material.

The most important achievement of this study is the elaborate discussion on the data selection and event log design for process mining ERP systems. A step by step approach for practitioners is presented to design and extract an event log from an ERP system. Data availability has been underexposed in academic literature (Roest, 2012), and with this research an important step is taken to guide practitioners in Process Mining on ERP systems.

EXECUTIVE SUMMARY

Process mining is presented in academic literature as a powerful way to conduct business process analysis. However, process mining research does not support end to end process mining projects. The research focuses on the development of mining algorithms, which is only a small part of a process mining project. Therefore, this research aims to develop a structured approach for practitioners to support end-to-end process mining projects. The results of this research are twofold: a structured approach to conduct process mining projects on ERP processes is developed. This approach is process and platform independent and it covers the considerations and design decisions that every practitioner will face. Furthermore, the approach is implemented for the Order to Cash process supported by SAP. The implemented approach consists of a data model and event log design that enables process mining on standard implementations of the SAP OtC process. Process performance and process complexity questions can be answered with the event log. The implemented approach is tested on practicality and feasibility in a real life case study.

The approach is developed by using a combination of literature and field research. Firstly, Process Mining literature is used to describe the state of the art of process mining. The literature review highlights three elements: general overview of what process mining is, an evaluation of the practical applications of process mining and a direction for the development of a process mining approach.

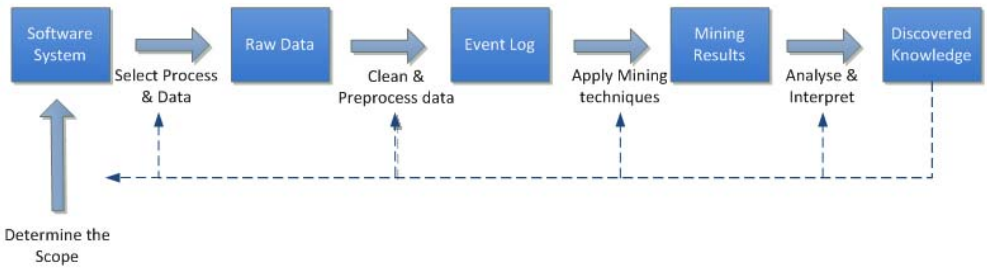


FIGURE 1: PROCESS MINING PROJECT APPROACH

The insights from the literature review are the starting point for the development of the Process Mining. Figure 1 depicts the process mining project approach. A project mining project consists of five stages.

Stage one: Determine the scope. Scoping a process mining project is captured in three elements: type of business process, project goals and analysis requirements. The type of business process influences the process mining project. Analyzing a structured administrative process requires different KPI's then the analysis of an unstructured health care process. Based on the type of process and the clients requirements, the project goals are defined. The GQM framework guides the determination of goals by breaking the high level goals down to focused analysis requirements. The definition of clear project goals gives the process mining project the right focus and guards the boundaries.

Stage two: Event log Design. The right data model has to be designed to enable process mining and to enable the accomplishment of the analysis requirements. The event log that is able to do both is designed in this stage. In event log design depends on four aspects: the case ID, the activity selection, the attribute selection and the time frame that the event log is covering. The

case ID is the common denominator the links all activities to a traceable case. The determination of the case ID is a challenge in ERP systems, due to the many to many relationships between process steps. ERP systems allow for the determination of activities on different levels of detail. It is important to select the activities that suit the analysis requirements. Attributes are data elements of the cases that are stored in the event log. They can be used for statistics analysis and for filtering the event log on certain characteristics. The event log is extracted for a finite period. It is important to choose the period such that all expected behavior (mean order time, seasonal effects, etc.) is captured in the event log.

Stage three: Preprocessing and cleaning the event log. In this stage, the event log design is translated into a data model that can be used to extract the event log from the ERP system. The data structures of an ERP system are not suitable for process mining. By mapping the event log design to the ERP systems data structure, the transactional data is structured as an event log. Since this stage is unique for each event log design, an implementation is shown for the OtC process that results in a data model that is generally applicable on all SAP implementations. The resulting event log has to be cleaned from noise such as incomplete cases or double counting of activities.

Stage four: Mining the Event log. Process Mining algorithms use the event log to analyze the process. An elaborate discussion on the technicalities of process mining algorithms is out of the scope of this research, rather a number of commercial process mining tools is described. The tool that is used to execute the case study and test the event log script is Disco. Disco is chosen as the preferred tool because of its availability in an academic license, the underlying algorithm that allow both a high level overview over the process and a detailed display of all possible flows. Furthermore, disco has extensive import and export capabilities and good performance.

Stage five: Analysis and interpretation of the mining results. The mining results have to be interpreted to give business meaning to the process characteristics and statistics. It is proposed that the tacit knowledge of process managers and process owner is important to give meaning to the mining results. Based on the type of project one or more iterations of discussing the results and refining the analysis is required to make an impact with the process mining project.

The approach is tested in a case study of a Dutch publisher of educational material. The five stages have been followed and showed to cover all challenges that were encountered during the case study. The sales process at the company has been analyzed with process mining and the analysis results are reported back to the management. The case study has shown that following the five stages enables process mining on ERP systems. A high quality event log is extracted from SAP and a Process Analysis with business value is presented to the case study company.

The results of the research are twofold. The five stages of the process mining approach enable practitioners to conduct end-to-end process mining projects. Practitioners lacked the support in the data extraction and data preparation stages of a process mining project. That gap in academic literature is bridged in this research. The implementation of the first three stages of the process mining approach for SAP OtC deliver a ready to use event log on standard implementations of SAP OtC. The event log automates the event log extraction from SAP, thereby enabling the analysis capabilities of process mining for practitioners. The practitioners are also guided in answering a standard set of analysis requirements that is representative for the analysis requirements that the market demands.

PREFACE

This thesis is the result of a graduation project in partial fulfillment of the Master of Science in Operations Management and Logistics at Eindhoven University of Technology. The research is conducted in cooperation with KPMG IT Advisory. Being part of a consultancy firm and an academic institution allowed me to combine the best of both worlds and challenge myself to find the balance between rigor and relevance.

First of all, I would like to thank my supervisors. Remco Dijkman has been my mentor and supervisor for the last two years. I would like to thank him for the interesting discussions, the fast and constructive feedback and the fact that he was always available for questions or discussions. I would also like to thank Marco Comuzzi for his support during the process. His high level overview over the project and useful feedback have been valuable.

Secondly, I would like to thank KPMG IT Advisory for allowing me to join the team. KPMG has given me the opportunity to work with experts in the field of data analytics and SAP and freely use the available resources in the organization. My colleagues have made my internship a really valuable experience where I got to know a lot about SAP, data analytics, the joys and troubles of working in a large organization and where we just had a lot of fun. Special gratitude goes to Dennis van de Wiel, my coach and supervisor. His analytic mindset and vast knowledge of data analytics and SAP have been a great help. Furthermore, I would like to thank Johan Steenstra and Gerben de Roest for helping me to bring my research to the real world in a case study. It was very valuable to experience the process of helping a client to improve his business.

Finally, I would like to thank everyone else who had a role in my studies and my graduation project. Friends, family and my girlfriend for the support and interest in what I am doing, for putting everything in the right perspective, and for distracting me at the right times.

The realization that this master thesis marks the end of two decades of schools, learning and studying has not come fully yet. However, I am looking forward to put all that I have learned into practice and keep on enjoying what I do, and doing what I enjoy.

Harmen Roest,
Eindhoven, 2012

CONTENTS

List of Figures

List of Tables

PART I: PRELIMINARIES

The first part of this master's thesis will introduce the research goal and its context. Furthermore, it is explained how the research is going to be conducted. The concept of process mining and a review of the practical applications of process mining are presented in the literature section.

The core elements of the preliminaries are summarized in Figure 2. Elements one and two are discussed in chapter one, and the third element is discussed in chapter three.

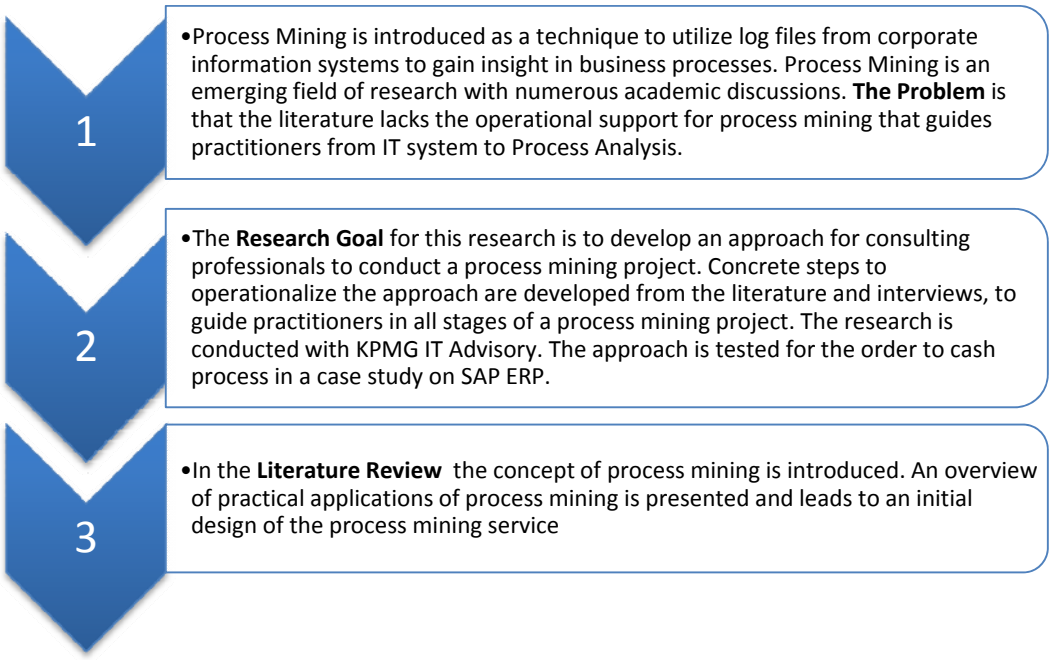


FIGURE 2: STRUCTURE OF PART I

INTRODUCTION

Companies are investing large amounts of money in the analysis and improvement of their business processes. Process analysis aims to understand what the company does, how and why they do it, which rules it follows and what types of results are desired. Traditional process analyses are time consuming sequences of reviewing procedures and interviews with people about the way the company works (van der Aalst, 2011).

Process mining is a set of tools and techniques that has the potential to utilize the data logs from the corporate information systems for business process analysis. Using process mining techniques to analyze business processes is claimed to be quicker, cheaper and often more reliable than the traditional process analysis approach (van der Aalst, 2011).

The field of process mining is lacking a practical approach to execute a process mining project. Especially the data preparation for process mining is underexposed in academic research. This research develops a practical approach to conduct process mining projects. The main contribution is found in the practical guidelines for the data preparation to enable process mining.

This chapter will introduce the problem statement and define the scope of the research. Furthermore, the research goal is presented and the roadmap to realizing the goal is presented in four sub questions. Finally the structure of the report is explained.

PROBLEM STATEMENT

Because process mining is a promising approach to business process analysis, it has been a hot topic in academic literature. The literature study by Roest (2012) shows that researchers are mainly focusing on the development of efficient and powerful analysis algorithms. However, Process mining literature does hardly cover the data collection and event log construction (Roest, 2012). Finding, merging and creating event logs that are suitable for process mining is not considered in academic research. Neither are the implications of the process mining results explained or discussed. There seem to be a number of challenges that are difficult to overcome by practitioners, which are not addressed by academic research.

A literature study on the practical applications of process mining by Roest (2012) concluded that the literature currently lacks guidelines for end-to-end business process analysis with process mining.

PROJECT GOAL

Therefore, this research aims to develop a practical approach to use process mining for business process analysis. The approach offers five steps that can be executed to perform process mining for platform and process independently. Furthermore, the approach is implemented and tested for the SAP Order to Cash process. The main contribution of this research is the practical guidelines for practitioners to design and extract an event log from an IT system. The project goal is summarized in the statement below:

Develop an approach for consultants to do business process analysis with Process Mining techniques.

In order to reach this goal, it is broken down in a number of questions and made specific for the scope. Answering those questions is the roadmap to realizing the research goal.

1) What are the analytic capabilities of Process Mining?
2) Which types of analyses are desired by consulting professionals and their clients?
3) How do the analytic capabilities of process mining match the analysis requirements desired by consulting professionals match?
4) Which steps are required to conduct a successful process mining project?

SUB QUESTION 1

What are the analytic capabilities of Process Mining?

The research of Roest (2012) is the basis for answering this question. The academic literature on practical applications of process mining is reviewed. It is used to give an overview of the analysis possibilities of process mining and a status quo of applied process mining research. Furthermore, literature is used to develop an initial approach to execute a process mining project.

SUB QUESTION 2

Which types of analyses are desired by the consulting professionals?

A process mining project will only be successful and lead to increased insight and knowledge if the analysis is focused. The construction of the event logs, the choice of algorithms and the analyses need to be guided by client questions.

The information needs are explored in interviews. A group of KPMG professionals is selected based on their experience with data analysis and their involvement in engagements with the case study companies. The KPMG professionals are encouraged to be specific and creative in defining their information needs. The information needs will result in a set of analysis requirements that need to be answered with process mining techniques in the remainder of the research.

From the KPMG clients where data analysis is conducted in the previous year, a list of possible cases is selected. Based on the availability of resources at KPMG and the case study company, a case study company is selected.

SUB QUESTION 3

How do the analytic capabilities of process mining match with the analysis requirements desired by consulting professionals match?

The analysis requirements from Sub Question 2 are mapped onto the process mining capabilities from Sub Question 1. It is highlighted with which mining techniques the analysis requirements can be answered. This question is answered throughout the steps developed in Sub Question 4.

SUB QUESTION 4

Which steps are required to conduct a successful process mining project?

The approach is designed by combining the initial design from the literature study with the requirements from the consulting professionals. The most important choices and considerations to operationalize each step are discussed. Academic literature and interviews with consulting professionals are used as input.

This results in a structured approach to conduct a process mining project that guides the choices and the steps that have to be made during a process mining project. All process mining stages will contain a generalized part that is applicable regardless of the process under consideration. Furthermore, each of the stages is implemented for the Order to Cash process as defined in the scope. The implementation for the Order to Cash project can be seen as a running example throughout the research.

The five steps in the approach are tested in a case study on a specific Order to Cash process. The data is provided by a KPMG client.

FINAL REMARKS

The results of this research are twofold:

1. A general approach to perform a platform and process independent, end to end process mining project on ERP systems.
   The practical and conceptual issues of process mining event data from ERP systems are addressed in the a five step process mining approach. By this general approach, the gap in academic literature is filled.
2. The approach is implemented for the SAP Order to Cash project. The implementation results in an event log extraction script covering the scope of an initial process scan on the OtC process. The event log extraction script is based on standard SAP an is applicable on any SAP system. It enables an initial process analysis on any Order to Cash process without consulting the client. In order to do a full process analysis, the analysis requirements can be further refined in cooperation with the client using the five step approach.

Every time the Order to Cash process is mentioned in the development of the approach, it is referring to the development of (2). Thus, referring to the *general Order to Cash process* without a specific company context. Only during the presentation of the case study in 0, the specific OtC process of the case study company is considered.

SCOPING THE RESEARCH

This section defines the scope and the context in which the research is performed and the kinds of questions that are going to be researched.

The industry context in which this research is conducted is provided by KPMG IT Advisory. KPMG has extensive experience with data analysis. Its audit department uses data analysis techniques to provide evidence for the annual audits. KPMG has developed a data analysis platform (Facts2Value) that supports over four hundred process performance indicators for audit and advisory purposes. The process performance indicators are computed based on data from the corporate information systems. Currently, over two hundred Facts2Value analyses engagements are performed annually. Facts2Value is still being expanded and improved. One of the current limitations is the extraction of process maps and process flows.

KPMG IT Advisory recognizes the potential of process mining as a set of techniques to perform business process analyses for their clients. They want to use process mining analyses for consulting purposes: quick business process analyses that pinpoint the potential wins for their clients. Operational questions related to throughput times (delays, bottlenecks, blockings, etc.) and quality levels of their services (% on time and in full), but also questions related to the quality of the administrative processes are asked (changes, cancellations, returns, etc.).

About 80% of their clients use SAP ERP as a corporate information system. There are no off-the-shelf solutions for the application of process mining techniques on SAP ERP systems. KPMG is interested in the opportunities that process mining gives for process analysis. It is expected that bridging the gap between SAP ERP and Process Mining techniques will offer valuable insight in the processes of their customers. The SAP Order to Cash (Sales) Process is chosen as the process of choice on which the research is tested in a case study. The sales process is chosen because it is a core process for most (all) customers and it is relatively difficult to analyze with regular data analysis techniques.

STRUCTURE OF THE REPORT

This report is structured in three parts. Part one introduces the research and its context. Furthermore a literature study is conducted where the concept of process mining is introduced and an initial version of the process mining approach is developed. Part two elaborates on the five stages of a process mining project. Each of the five stages will be discussed in a separate chapter. Part three is the practical evaluation of the process mining approach in a case study. Finally, the conclusions and directions for further research presented in chapter 9.

LITERATURE REVIEW

Process mining literature is used to describe the state of the art of process mining. The state of the art is researched in three elements: general overview of what process mining is, an evaluation of the practical applications of process mining and structural directions of a process mining approach.

PROCESS MINING OVERVIEW

The goal of process mining is to extract knowledge and value from transactional data stored in corporate information systems. By using the recorded data from the corporate information systems, Process Mining offers techniques to perform objective analyses based on factual event data. Traditional process analysis used interviews and workshops to 'extract' the process flows from the organization. Extracting 'PowerPoint realities' is a risk in traditional process analyses that is mitigated by the use of process mining techniques.

Figure 3 shows the relationship between the real world, the software systems that support the real world and the position and possibilities of process mining. The software systems record and store transactional data. The recorded data can be converted into event logs, which serve as the basis for a process mining analysis. By means of event logs and process models, process mining aims to analyze the real world and improve business processes and the software systems that support and control the real world.

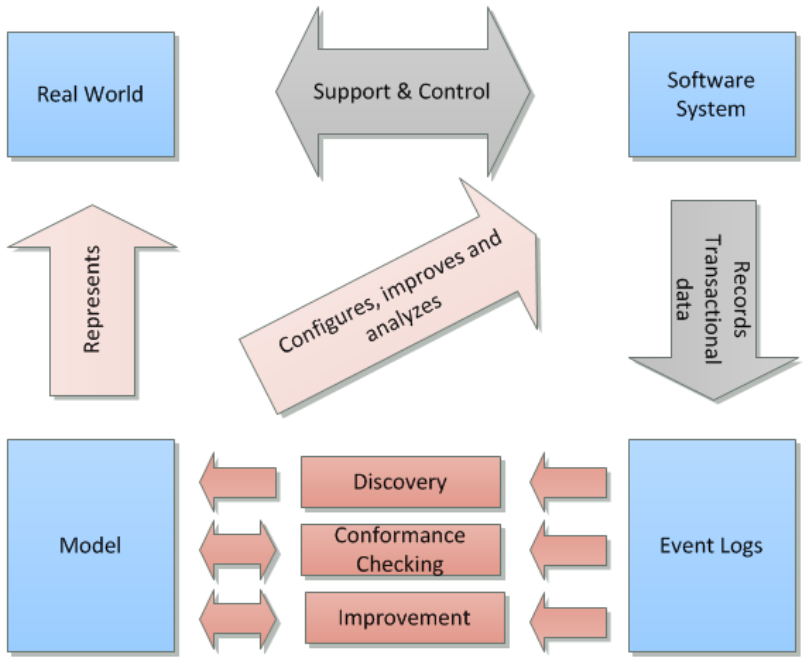


FIGURE 3: POSITIONING OF THE THREE MAIN TYPES OF PROCESS MINING BASED ON VAN DER AALST ET AL. (2012, P. 174, 175)

Three types of Process Mining

As depicted in Figure 3, there are three types of process mining: process discovery, conformance checking and process improvement. (van der Aalst et al., 2003, van der Aalst, 2011, van der Aalst et al., 2012, van der Aalst & Weijters, 2004). These three areas consist of techniques to link event logs with process models. Finally, the resulting models represent the real world, based on actual data, rather than assumed or tacit knowledge. Since the models are a valid representation of the real world they can be used to configure and improve the software systems that support the real world.

Process discovery is traditionally the most prominent area in process mining. It aims to extract information from an event log without a priori process models or process information. The event log is the only input for an algorithm to construct a (process) model (van der Aalst, 2011, van der Aalst & Weijters, 2004, Mans, Schonenberg, Song, van der Aalst, & Bakker, 2009).

Conformance checking can be used to compare the reality captured in the process model with the reality recorded in the event log. Conformance checking requires the input of an event log, but the input of a process model as well. This can be a handmade model from the past, or a result of another process mining effort. The model represents the process as it should be or is expected to be, whereas the event log shows what actually happens in the process (van der Aalst, 2011, van der Aalst & Weijters, 2004, Mans et al., 2009).

Process Improvement uses the process model and the event log to improve or extend the process model or the business process. Improvement action is required when the process model does not represent the reality well. Processes will change over time and process models are not always correct representations of the reality. Both require an adaption of the process model to increase the alignment between process model and reality. Besides ‘repairing’ the model, the process model can also be extended: new perspectives can be added to the model. New attributes in the event log or new mining algorithms may give opportunities to extract more information from the event log (van der Aalst, 2011, van der Aalst & Weijters, 2004, Mans et al., 2009).

Mining different perspectives

Process mining techniques offer the possibility to investigate the event logs from different perspectives. Just focusing on the control flow is a good way to get to know the business process. However, viewing the data from another perspective (Table 1) deepens the understanding of the business process and gives different insights for the analysis (van der Aalst, 2011, van der Aalst & Weijters, 2004). The applicability of the different perspectives is orthogonal to the types of process mining, but may sometimes depend on the information available in the event log. It does not make sense to do a temporal analysis of a process when there is no temporal information available in the event log.

Perspective	Explanation
Control flow perspective	The control flow perspective aims to find a good representation of the ordering of the activities. Using the event log, the algorithm will try to make a model that shows all the different paths that a particular case can follow through the process.
Organizational perspective	The organizational perspective focuses on the resources and their different roles and positions in the process. Who is involved in which places in the process, and how are they related to each other and to cases.
Case perspective	Instead of the activities, the case perspective focuses on the cases and their characteristics. For example: path in the process, resources working on the cases. Furthermore, the cases can be mined based on additional data elements such as costs, suppliers or quantities
Time perspective	The time perspective focuses on the temporal behavior of a process. It is possible to discover bottlenecks, monitor utilization, etc.

TABLE 1: PROCESS MINING PERSPECTIVES (VAN DER AALST, 2011, P. 11)

Following from the previous sections, process mining promises added value for process analysts. Extracting process maps based on real occurrence of activities, and the possibility to show the handovers of work in an organization can of great value both for consultants looking for improvement opportunities and for auditors looking for organizations that are in control. Despite the potential that is shown in the previous sections, process mining is not widely adopted amongst practitioners.

Process Mining Stages

Process mining follows a number of steps to reach the goal of knowledge discovery. Since process mining is a special type of data mining, the process mining project follow a structure that is closely related to the data mining process. An overview of the process mining steps is shown in Figure 4, based on the process of knowledge discovery in databases, described by Fayyad, Piatetsky-Shapiro, & Smyth (1996).

Figure 4 comprises the logical order of a process mining project. From a software system, a process is selected for analysis and the accompanying data is downloaded. This data is converted into the right format and where possible the data is cleaned. The preprocessing efforts result in an event log that will serve as the input for the selected process mining techniques. Thereafter, the mining results are interpreted and knowledge is discovered. The process steps are iterative: so it is possible (and often necessary) to go back and improve previous steps. The whole process takes place in a certain context, it is therefore important to realize the possible impact of the context for the results of the process mining process.

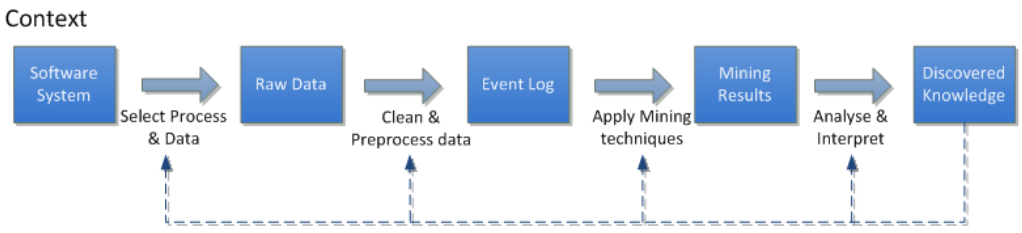


FIGURE 4: OVERVIEW OF PROCESS MINING STAGES, DEVELOPED FROM FAYYAD ET AL. (1996, P. 41)

CASE STUDY EVALUATION FRAMEWORK

The aim of this master thesis is to apply process mining in a business context. Academics have been conducting case studies to test their algorithms and show that process mining provides additional insights in business processes. Roest (2012) developed a framework for the evaluation and comparison of process mining case studies. All process mining case studies from the last decade were collected and evaluated and compared based on a set of criteria. This section will briefly introduce the evaluation criteria and present the results of the comparison framework.

EVALUATION CRITERIA

The developed framework is based on the process mining process described in the previous section. Following from Figure 4, three categories of characteristics are identified: Context, Preparation & event log, and Process Mining. Figure 5 shows the three categories:

- The software system(s) of the company and the selected process(es) are the context of the process mining process.
- The raw data that goes through cleaning and preprocessing to an event log can be described as preparation & event log.
- The application of process mining techniques and the analysis of the results are grouped under Process Mining.

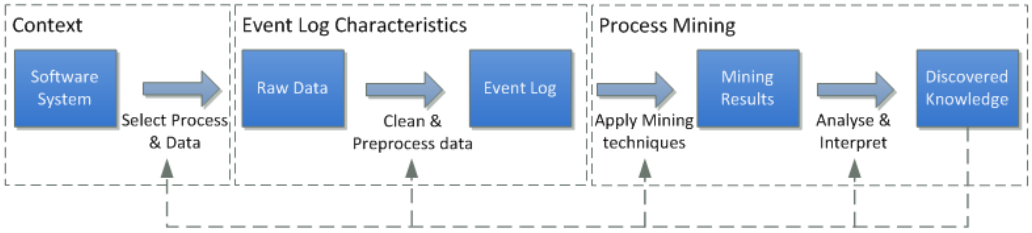


FIGURE 5: THREE CATEGORIES OF CHARACTERISTICS

For each of the three categories shown in Figure 5, a number of characteristics are given on which the practical applications of process mining will be assessed. Figure 6 shows the three categories and the characteristics for each category. The characteristics are based on and limited to the level of detail found in the case studies. Some of the case studies simply prove that a proposed algorithm is working and do not go into details about the process mining process. An explanation of the characteristics can be found in Roest (2012).

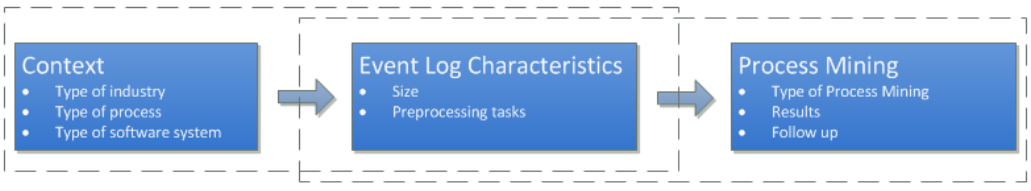


FIGURE 6: EVALUATION FRAMEWORK PRACTICAL APPLICATIONS OF PROCESS MINING

PRESENTATION OF THE RESULTS

Roest (2012) selected 42 case studies about Process Mining in a real life context. The case studies are compared using the framework described above. This section gives an overview of the results according to the three categories of characteristics. An elaborate discussion of the results can be found in Roest (2012).

CONTEXT

This section presents the most important observations from the context criteria.

Government and Healthcare industry are the two most frequently used industries to conduct a case study. They account for half (21 out of 42) the case studies in the framework. The other half of the case studies is taken from very different industries. They are ranging from agriculture to housing agencies and the gas industry.

Business processes range from very structured and predictable to unstructured. Structured or routine processes are processes that have a limited amount of variety and are characterized by a regular and predictable flow of tasks. Unstructured or non-routine processes are processes that adapt to information learned from the tasks as it unfolds (Lillrank, 2003). The processes that are analyzed in the case studies make up the whole spectrum of 'structuredness'. Structured processes like invoice handling and complaint handling are alternated by unstructured processes like testing or product development.

Workflow Management Systems (WfMSs) are the dominant source of event data. WfMSs suit requirements of process mining well because they provide a clear case ID and an intensive support of the process. ERP systems are well represented as well. A large number of organizations have implemented an ERP system, so ERP systems are a logical source of event data.

Table 2 provides an overview of the originating industry of the case studies. Table 3 gives an overview of the software systems that provided the event data .

Industry	
Government	14
Health Care	7
Gas Industry	3
University	3
Software Development	3
Banking	2
Manufacturing	2
Car Industry	2
Banking	2
Agriculture	2
Wafer Scanners	1
Spanish Authors association	1
House rental agency	1
Various	1

TABLE 2: OCCURRENCES OF INDUSTRIES

Computer System	
WfMS	10
ERP	5
Internal database system	5
Computer Logs	2
HTML website	2
Interviews	2
Financial administration	1
Machine logs	1
Observations	1
Tellstory	1
Administrative System	1

TABLE 3: COMPUTER SYSTEMS

PREPARATION & EVENT LOG

Preparation & event log is explained based on two criteria: size and preprocessing tasks (Table 4). Information on the event logs was less detailed in the case studies. Most case studies gave some information on the size of the event log (37 out of 44), but only half of them described the preprocessing tasks that were needed before the event log was suitable for mining (23 out of 44).

The event log sizes are different in terms of the number of process instances taken into account in the analysis. The number of instances ranges from 24 to 130136, the same holds for the number of recorded events: 3023 to 279333 events. However, the event logs show similar patterns. The number of activities is rather low (5-130) compared to the number of recorded events and the number of process instances (6,000-130,000). This indicates a large number of relatively small log traces in the event log. Despite the low number of activities in the event log, there are still very many possible traces between those activities. The large number of traces will show a large portion of frequent behavior, pointing in the direction of the 'right' model. However, there is still a high change on infrequent behavior that makes it harder to extract an unambiguous process model.

There is one exception to the ratio between activities and process instances: Rozinat, de Jong, Günther, & van der Aalst (2009) investigated the test process of wafer scanners at ASML. In that case study, only 24 cases were found, but 154966 events were recorded in the event log. That means a small number of very long log traces in contrast to a large number of short traces. Mining these kinds of unstructured processes comes with additional challenges to extract knowledge from the event data. Rozinat et al. (2009) handled these challenges by applying filtering techniques to the event log. Rather than focusing on frequent traces, they started looking for frequent test activities that were carried out in all traces. Hereby, their process model does reflect on the common tests in the process.

Pre-processing tasks are hardly reported in full detail in the case studies. Only half the case studies report on pre-processing at all and the most frequently seen pre-processing task is a conversion of the event log into an appropriate format for process mining or the application of a filter for the most frequent behavior (Table 4). Furthermore, some case studies aggregate the data to the same level of detail and others redefine time stamps to make sure that it is possible to mine the right level of detail and the right perspectives.

Pre-processing activities	
Conversion to suitable format	9
Filter	8
Consistent aggregation level	3
Redefine time stamps	2
Cleaning	2
Remove start/stop events	1
Merging of event logs	1
Change Application to extract log	1

TABLE 4: PRE-PROCESSING ACTIVITIES

PROCESS MINING

The last category of criteria is explaining the actual process mining tasks and results. This section will discuss the process mining criteria: type of process mining, results and follow up.

van der Aalst et al. (2012) state that process discovery is still the most dominant type of process mining. The case studies in the framework support that claim, 35 out of 44 case studies are focusing on discovery. The case studies show discovery in all the four mining perspectives described in section 0 (control flow, organizational, case and time perspective). Control flow seems to be the dominant perspective, followed by the organizational perspective. Mining the control flow can be the goal of a process mining effort, but can also be part of a larger package where the analyses zoom in on particular parts of the process. Extracting the real control flow gives insight in the process, also for conformance checking and process improvement tasks.

For most case studies, the results are straightforward. Processes are discovered, and insight and knowledge is gained from the event logs. Parts of the case studies are written as evidence for newly developed algorithms. Testing them on a real event log adds to their value. However, the result of that approach is that it is merely shown that the algorithms are working rather than an in-depth analysis of the process that is researched. Only a handful of case studies give a detailed analysis of the process mining results. It is promising to see from those analyses that valuable knowledge is extracted from the event logs. Dominant loops are discovered, as well as bottlenecks and other flaws in the control flow. Furthermore, possible causes of delays in the process can be demonstrated with process mining techniques. For example, van der Aalst et al. (2007) investigate an invoice handling process. They found out that there is a relation between the amount of money involved and the performance of the case in the process. These types of detailed analysis can make process mining valuable for applications in the real world, not only to support a new algorithm, but to do real analyses and contribute to organizational improvements.

Following from the claim that process mining contributes to business process analysis and improvement, the follow up criteria was added to the framework. Only one case study reported on concrete process improvements based on the process mining analysis (van der Aalst et al., 2007). Some others reported on directions for further research, but the majority of the case studies did not report on any follow up on their process mining efforts. This observation can be explained by the nature of academic research and the types of papers that are published. The real analyses and results of process mining are a lot less relevant in academic terms than supporting a new algorithm with a case study. The detailed analyses with high practical value can still be present with practitioners of process mining.

Process mining Project Approach

The evaluation framework shows that the academic literature on process mining has a focus on a small section of the process mining phases (Section 0). Researchers are interested in the robustness and applicability of algorithms, and often take data quality and data availability as a given. They do not support end-to-end process mining projects. Practitioners face a different reality where the application of algorithms is not a goal in itself, but just a means to achieve the goal of knowledge discovery. With the current state of applied process mining research, practitioners are insufficiently supported to conduct a process mining project.

The goal of the research is therefore to deliver a general approach to execute a process mining project on any process supported by any IT system. The approach is starting from the data collection and construction of high quality event logs, and ranging to a guided interpretation of the results.

The process mining stages (Figure 4) derived from Fayyad et al. (1996) are used as an initial version of the approach. These process mining stages by however mostly consist of practical steps. The conceptual step of scoping the research (determining project goals and analysis requirements) is an important step that is taken into account in the development of the process mining approach, and will therefore be added to the approach (Figure 7).

In the main part of the research, the five process mining phases are discussed elaborately.

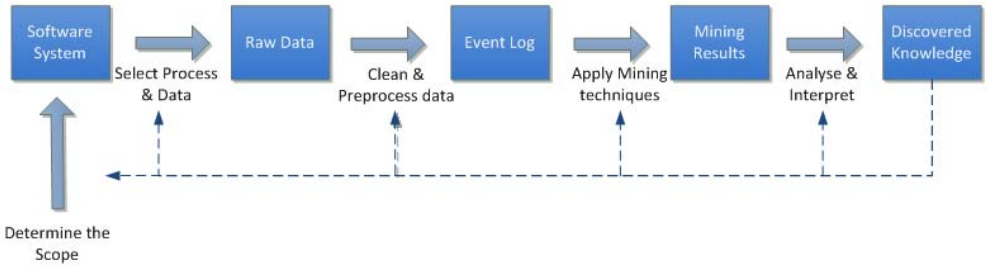


FIGURE 7: PROCESS MINING APPROACH

Determine the Scope

As a prelude to every process mining project, the scope of the research has to be defined before the project can actually start. Determining the scope was not part of the evaluation framework discussed in the previous section. However, when a process mining project is going to be conducted in practice, determining the scope is an essential step for process mining success.

For a process mining project the scope is determined in three elements: type of process, project goals and analysis requirements. These three elements are discussed in 0.

Event Log Design

Selection of the process to be researched and the data that is used to analyze the process is a crucial step in every process mining project.

The event log is designed in three steps: selection of activities, determination of the case ID, attribute selection and time frame determination. 0 elaborates on these four steps of the event log design.

Cleaning and Preprocessing

When the goal and the scope of the process mining project have been determined and the necessary data is downloaded, the next stage is to critically asses the data.

Cleaning and preprocessing ranges from converting the data into an event log format to dealing with incomplete cases and noise in the data, to selecting the appropriate timeframe. Similar to the previous stage, the quality of the data correlates strongly with the quality of the analysis results. van der Aalst et al. (2012) urge practitioners to treat their data as 'first class citizens': complete and trustworthy event logs will lead to reliable analysis results.

0 discusses the steps related to this stage.

Apply Mining techniques

With an event log consisting of the right events and attributes – cleaned and in the right format – the actual mining algorithms can be applied. The questions that have been driving the process mining project so far are going to be answered by selecting the right mining techniques and applying the right algorithms.

Algorithms are contained in process mining tools. A selection of process mining tools is described in 0.

Analyze and Interpret Results

After the mining algorithms have been applied to the event log, the results have to be interpreted. Process analysis is not an exact science: even when the analyses are based on real event data the interpretation of the results strongly depends on the context. Valid conclusions can only be drawn in close cooperation with the process owners. They have the knowledge to clarify odd results and point the researcher into the right direction for further research and investigation. The stage of analyzing the results will typically lead to several iterations of previous steps. The scope can be adjusted as a result of intermediate results, which will lead to going through the all the stages again. Similarly, the analysis and interpretation of results might lead to revisiting any of the previous stages. When the analysis requirements in the scope are satisfactory answered, the project is finished.

0 provides a discussion on the analysis and interpretation of mining results.

PART II: PROCESS MINING APPROACH

The literature study in the previous chapter shows the stages in a process mining project, based on the stages in a knowledge discovery in databases (KDD) project (Fayyad et al., 1996). These stages are recognized in the process mining case studies in the case study evaluation framework (Roest, 2012). As explained in the end of chapter 1, a scoping stage is needed to guide the four practical stages of a process mining project.

This second part of the thesis elaborates on the five stages of a process mining project. Each of the five stages will be discussed in a separate chapter (Figure 8).

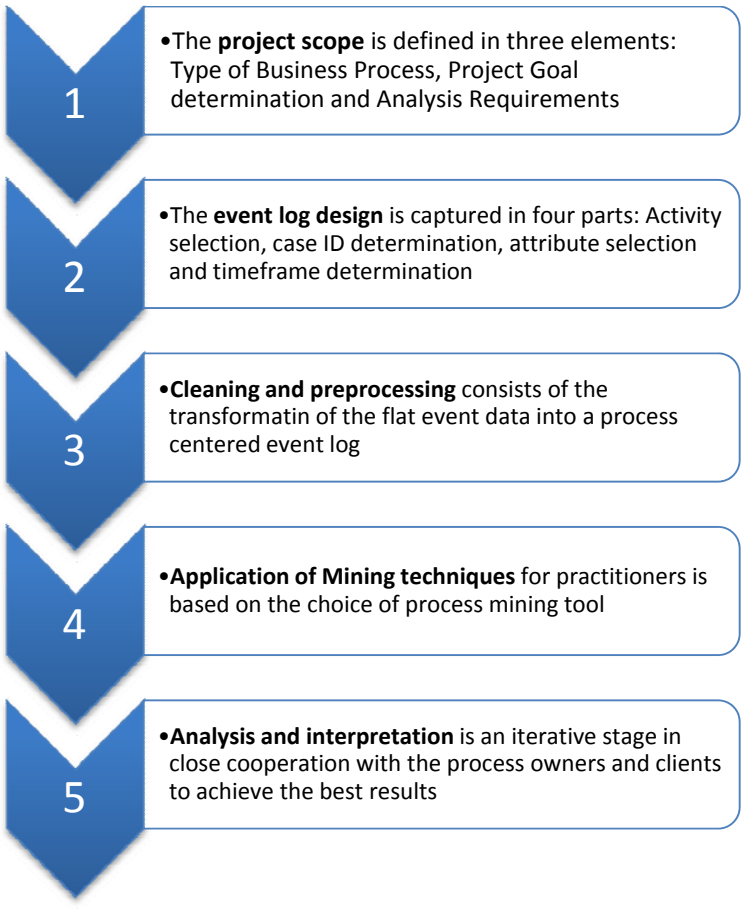


FIGURE 8: STRUCTURE OF PART II

DEFINING THE SCOPE

The scope of the research has to be determined as a prelude to every process mining project, before the project can actually start. Determining the scope was not part of the evaluation framework discussed in the previous section. However, when a process mining project is going to be conducted in practice, determining the scope is an essential step for process mining success.

For a process mining project the scope is determined in three elements: type of process, project goals and analysis requirements. The first paragraph gives an overview of the different types of processes and supportive software systems that can be encountered in a process mining project. The second paragraph elaborates on the determination of project goals, related to different types of process mining projects that are identified in the literature, and the types of process mining that can be used to execute the projects. A good understanding of the type of process, the type of software system and the project goals is essential for the specification of the analysis requirements. The analysis requirements that are identified in interviews with consulting professionals are discussed in paragraph three.

TYPE OF BUSINESS PROCESS

The business process under consideration needs to be analyzed as part of the project scope. For process owners, this might by a trivial step, but consultants need to take the type of process into consideration when the project goals and analysis requirements are determined. For example, a hospital process (unstructured, end-to-end process is not known a priori) requires different KPI's and project goals than an invoicing project (structured, administrative, all steps are known a priori).

A business process is defined as a series of activities that will lead to a clearly defined business goal (Davenport & Short, 1990). Processes are found everywhere: ranging from a simple process of doing grocery shopping to very formal processes to grant permits. Processes are found everywhere, but they are very different in characteristics.

Business processes range from very structured and predictable to unstructured. Structured or routine processes are processes that have a limited amount of variety and are characterized by a regular and predictable flow of tasks. Unstructured or non-routine processes are processes that adapt to information learned from the tasks as it unfolds (Lillrank, 2003). Different types of processes pose different opportunities and challenges for process mining projects.

A treatment process in a hospital is very unstructured, the diagnosis of an individual patient determines the next step in the treatment process. It does not make real sense to investigate the number of variants through the treatment process and asses process quality based on that metric. It makes a lot more sense to go one step deeper and investigate the relation between diagnoses and the route through the treatment process. Another possibility is to compare similar treatment processes across different hospitals (Mans et al., 2008, 2009). A highly standardized administrative process like handling invoices or collecting fines is more likely to be investigated by determining control flow and process metrics like ''number of invoices' or 'number of days to collect a fine'

A business process can only be investigated with process mining if event data is available. Nowadays most processes are supported by information systems. Event data is available for most processes, and can be either 'object centered' or 'process centered'. In a 'process centered' information system all logging information is related to a process instance that can be monitored

through the entire process. An example of a process centered information system is a workflow system that supports a customer service process. Every service request is a case and all actions that are performed to answer the request are logged as activities related to the case. In an, object centered information systems all log activities and characteristics are related to *objects* in the process. For example, ERP systems information is contained in the document rather than in an event log. The system's change log will contain 'events' but they are also uniquely linked to the document, not to the process. On document can be involved in many processes and on process can involve many documents. In a sales process: when a sales order is posted, a sales document is created that contains all relevant information about the sales order. The sales order will continue to follow the process, but for every step in the sales process a new document is posted. When the end to end sales process is monitored, the different documents have to be merged into a traceable case.

This research develops a process mining approach for consultants. A business process selection from a consultants perspective adds the dimension of repeatability. A consultant wants to utilize the development efforts (designing and constructing the event log, choice of mining algorithms, etc.) in more than one assignment. Standardizing the process mining approach requires a business process that is easy to analyze and that is supported by a widely used information system.

The business process that is going to be investigated is the sales process (Order to Cash), supported by SAP ERP systems. The sales process is a structured process, and because SAP is a widely used ERP system, the SAP implementation of the sales process will be operational with a lot of potential clients.

PROJECT GOAL DETERMINATION

This section describes the determination of project goals for process mining projects. van der Aalst et al. (2012) identifies the determination of clear project goals as an important success factor of process mining projects. The different types of process mining projects are explained. Furthermore, the Goal, Question, Metric framework (Caldiera & Rombach, 1994) is proposed as a way of defining the goals for a process mining project. Finally, it is shown that the determination of project goals leads to the selection of process mining types and perspectives.

van der Aalst (2011) identifies three main types of process mining projects:

- A *Question Driven* project. Specific questions about the process or the performance are guiding the project. For example: "How many different ways of handling an invoice are found in my purchasing process?" or "How often is an invoice payment overdue?"
- A *Goal Driven* project aims to change or improve a process on certain KPI's or characteristics. For example: "I want to simplify my process of granting building permits." or "I want to decrease my average lead time with 10%"
- A *Curiosity Driven* project. An explorative project without the guidance of specific questions or goals. The available data is put to the test with the expectation that it will provide valuable insight or will lead to more specific questions that require further analysis

Question driven projects are usually the easiest to perform. The goals and objectives are very specific and clear, making it easy to define the data requirements and perform the mining techniques. A goal driven project usually requires several iterations of defining problem areas and investigating root causes. It is not clear from the start which mining techniques are required

and what the exact data requirements are. Goal driven projects require more time and resources to accomplish than question driven projects. A curiosity driven project is the hardest type of project to perform, especially when the event data is scattered around different systems and extracting an high quality event log is far from trivial. Current process mining algorithms can analyze event logs from various perspectives, and with virtually unlimited filtering options, it is easy to get lost in the data. However, if time and resources can be made available, a cooperation between a process mining expert and a domain expert can lead to surprising insights in a business process.

Regardless of the type of project that is conducted, the analysis direction should be defined in the scoping stage. A good way of translating goals or general analysis directions into focused and concrete analyses is the Goal, Question, Metric (GQM) framework. GQM is a top down framework from the determination of project goals towards concrete analyses. It can be particularly helpful in question driven and goal driven projects to use the GQM framework to ask the right questions and define the right data requirements.

GQM starts with defining a conceptual project goal. The project goal is specified in:

- A purpose (improve, shorten, increase, remove, etc.)
- An issue (utilization, throughput, quality, etc.)
- An object (process, resource, product, etc.)
- A viewpoint (cost perspective, managers perspective, time perspective, etc.)

The resulting goal is of a conceptual nature, therefore hard to reach directly. The goal is refined into a set of questions. The questions represent the different aspects of the goal and allow the goal to be measured quantitatively. Answering the questions requires one or more metrics that can answer the question on an operational level. The metrics have specific data and analysis requirements and combine the data to an answer on the questions. Together, the questions answer the high level project goal.

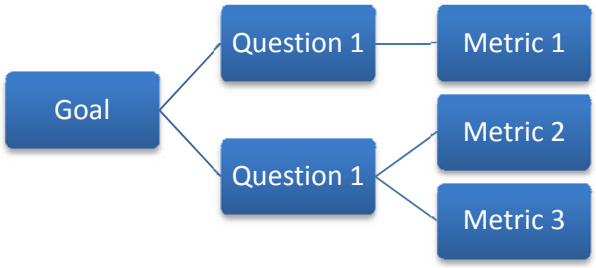


FIGURE 9: GQM METHOD

For process mining projects, the definition of questions and metrics is guided by the different types and perspectives of process mining. As stated above, the definition of metrics is a set of data that answers the question in a quantitative way. Each metric should be defined in a combination of the type of process mining and the process mining perspective. For example the number of different paths through the process is found in (1), the throughput time of priority orders vs. the throughput time of normal order is found in (2), and the employees that violate the segregation of duties policy are found in (3) (see Table 5). Each type of process mining project (with associated goals) can be executed with a combination of the type of process mining and the mining perspective.

	Control flow	Organizational	Case	Time
Process Discovery	# unique process flows (1)			Throughput Time (2)
Conformance Checking		SOD policy (3)		
Process Improvement				

TABLE 5: POSITIONING METRICS

This approach is developed as a goal driven process mining project. Companies often involve consulting professionals when they need external expertise for achieving some predefined goal. Consulting professionals have the experience with similar clients having similar questions. It is therefore very well possible to develop a standardized approach to perform process mining on certain types of engagements.

Analysis Requirements

This paragraph shows the selection of process goals, questions and metrics for analyzing the Order to Cash process that is chosen in paragraph 0. The GQM approach from paragraph 0 is used to specify the analysis requirements from consulting professionals. This paragraph is specific for the OtC Process. However, the GQM method is generally application, regardless of the business process under consideration.

The requirements analysis is focused on the Order to Cash (sales) process in SAP ERP. SAP ERP is an object centered ERP system. It is a widely adopted ERP system with implementations in tens of thousands of companies worldwide. The selected sales process is structured administrative process, which is a core value adding process in all companies. These factors makes it suitable and attractive for the design of a multi-functional process mining approach.

In the scoping stage of the process mining approach, interviews have been conducted with eight consulting professionals. Two aspects are explored in the unstructured interviews: process mining analysis requirements for consulting professionals, and data availability in SAP. This section will elaborate on the analysis requirements resulting from the interviews. The general approach of the GQM framework is used to classify the project goals, questions and metrics for a process mining project.

The general response to mining the sales process was similar throughout the interviews:

- Very powerful that the analyses are dynamic: zooming in and out from different perspectives is really an advantage.
- Allows for “quick-and-dirty” process scans and identification of problem areas for root cause analysis projects
- My clients do not have a clear understanding of what their processes look like
- Is it possible to measure the time aspect of the process?

The interviews show that a good business process is a process with *optimized effectiveness* that is realized against *reasonable costs* and under *adequate control*. This statement is translated in three categories of analysis requirements:

- Process Performance: How well does the process perform in terms of reliability, validity, efficiency, completeness, etc.
- Process complexity and consistency: structure, how many different paths and variations through the process are found, etc.

- Process Control: is the sales process in control from a risk perspective (fraud risk, financial risk, operational risk)?

Please note that the analysis requirements are derived from interviews with practitioners. The set of required analysis will not be complete, and is subject to the background and experience of the practitioners that were interviewed. A complete and verified set of possible analysis on the sales process requires a larger number of interviews and a theoretical validation. However, the requirements defined in the interviews are representing relevant questions for any company. They will enable an initial process analysis that is applicable on any sales process, without consulting the client for company specific research goals or requirements.

Table 6 shows the application of the GQM metric on the analysis requirements derived from the interviews. The overall goal is to optimize the OtC process. The goal is broken down into three questions representing the three categories of analysis. For each category, a number of metrics are defined that answer the question.

Goal	Question	Metric
Optimize the SAP OtC Process	How well does my process perform?	What Percentage of my orders is a no touch order?
		How many changes occur?
		What are my bottlenecks?
		What are my top 10 changes?
		What is the average Throughput time?
		What is the average Payment Term?
	How complex is my process?	What are the top 10 interface users?
		How many different order flows can be identified?
		How is the 80/20 distribution of unique flows divided over the cases?
		What are the differences between BU1 and BU2?
		What are the differences between Order Type 1 and Order Type 2?
		What are the differences between Product Group 1 and Product Group 2?
	Is my Process In Control?	Are there any violation of the segregation of duties
		Is there any unauthorized behavior found
		Are there any indicators of fraudulent behavior found in the process?
		Does the process map confirm the procedural process maps defined by the company?
		Does the configuration of the (SAP) system support the policies and procedures in place (ITACS)

TABLE 6: GQM EXAMPLE FOR OTC

OUTLOOK TO EVENT LOG DESIGN

In this first stage, the scope of the project is defined. With a clear notion of the process, the information system that provides the event data, and the type of project, the next step is to design an event log that will enable answering the analysis requirements. The first two categories are included in the event log design. Analyzing if the process is in control requires an event log that includes system settings and other information from the company. The scope of the initial event log design is to be able to do an initial analysis without consulting the client.

EVENT LOG DESIGN

Process Mining is a data driven analysis technique. Without data there is no analysis. Furthermore, process mining success strongly correlates with the quality of the event log. Selecting the right data and designing an event log that is sufficient for answering the research questions is a key step in every process mining project. This chapter describes the data selection stage in a process mining project.

An event log has three minimal requirements for process mining (van der Aalst & Weijters, 2004, van Dongen & van der Aalst, 2005):

- Each event corresponds to an activity that was executed in the process (Paragraph 0)
- Multiple events are linked together in a process instance or case (Paragraph 0)
- Each case forms a sequence of events (activities) that are ordered by their timestamp (Paragraph 0)

Besides the minimal requirements, event logs can be enriched with additional attributes (such as resources) that enable the use of the different mining perspectives (organizational mining) and enables the use of filters to make cross sections of the process. The selection of attributes is discussed in paragraph 0. The download scope of the event log has to be determined. Paragraph 0 elaborates on a suitable time frame for downloading event data. Paragraph 0 shows the necessary activities and attributes for answering the requirements analysis from paragraph 0. Finally, paragraph 0 gives an outlook to the next stage in the process mining approach.

ACTIVITY SELECTION

Process mining is a process centered approach, and in the previous section it is stated that every activity should be related to a process instance or case. Firstly, the activities are selected and secondly the case ID that relates to the activities is determined. The scoping phase has determined the analysis perspective and the analysis requirements. The next step is to determine the activities that are related to the analysis requirements and perspective. Activity selection depends on analysis requirements in four aspects: process boundaries, analysis perspective, level of aggregation and level of IT support. All four aspects are discussed below.

PROCESS BOUNDARIES

Defining process boundaries consists of determining the beginning and the end of the process as well as defining which activities are in scope or out of scope of the process. Every process consists of one or more activities. Defining the process with a list of activities and designing an event log is a manual step in a process mining project. Processes supported by a workflow system have a straightforward set of activities that can be related to a case. When a process is considered that is supported by an ERP system, the process boundaries have to be chosen and the activities that are part of the process have to be selected. Processes supported by ERP systems do not have a clearly defined end-to-end structure. Decisions made in one part of the process, trigger the start of another processes that is needed to finish the first process. Consider for example the sales process in SAP. In case the goods that are ordered are not available in stock, the sales process triggers the procurement process and/or the production process. When the goods are procured or produced, the materials become available in stock and the sales process can continue (

Figure 10).

It is not possible to mine to interrelated system of business processes as a whole. The object centered structure of ERP systems makes it impossible to define a case that is traceable through the whole system. A clear begin and end of a process must be defined in order to be able to mine the process. When interrelated processes need to be mined, that will results in more than one event log.

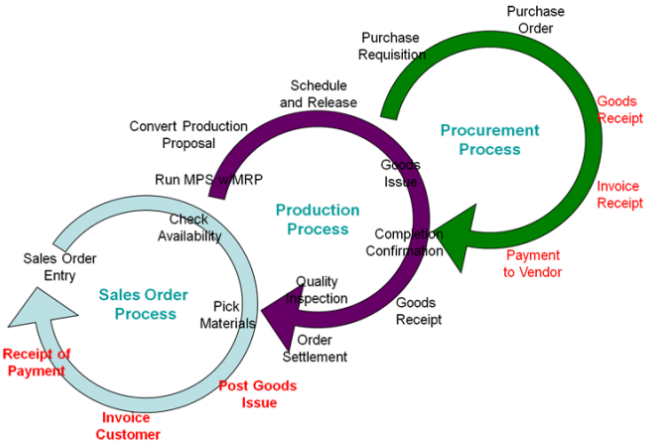


FIGURE 10: INTERRELATION OF BUSINESS PROCESSES (PIESSENS, 2011)

ANALYSIS PERSPECTIVE

The analysis perspectives defines which actions are chosen as activities in an event log design. The definition of process goals and metrics determine the analysis perspective that is analyzed. The analysis perspective determines the perspective from which activities are selected. This is again caused by the object centered architecture of ERP systems. A sales process can be viewed from the perspective of a sales order: what is the order flow through the process, what types of sales orders are processed, etc. In that case, every sales order will have a unique case ID and all activities are related to a sales order. It is also possible to analyze a sales process from a customer perspective: which products are sold, in what quantities, etc. Every customer will have a unique case ID and all the sales activities related to the customer are logged in the event log.

The project scope guides the definition of project boundaries. Paragraph 0 refers to the GQM method as a good way of refining high level project goals into concrete goals and metrics. The required activities per metric should be selected. In principle, every metric with its associated activities can result in a separate event log that is tailored to answer just that specific metric. In practice it will often be preferred to combine as many metrics as possible in one event log. Metrics can be grouped into one event log when the analysis perspective and the level of aggregation (explained in paragraph 0) are similar. It is then possible to determine a Case ID (explained in paragraph 0) and mine the defined process.

Aggregation Level

The aggregation level of a process mining analysis is the level of detail of which the activities are chosen. Activities can be selected on different levels of aggregation. On a strategic level a process is usually summarized in a small number of high level activities. However, on an operational level, these high level activity consist of many steps. Process mining can be performed on different levels of aggregation. Figure 11 shows an example of different aggregation levels for the sales process supported by SAP ERP. The main flow of the SAP sales process is Sales Document, Delivery, Invoice and Payment. Each of these documents can have a 'creation', a 'change' and a 'deletion' event. Creation of the sales document can be broken down in to create standard order, credit nota, quotation, return, etc. Lowering the level of detail increases the number of activities, and consequently the complexity of the mining results. However, it also increases the insight in the level of detail in the analysis, which might be required according to the project goals and analysis requirements.

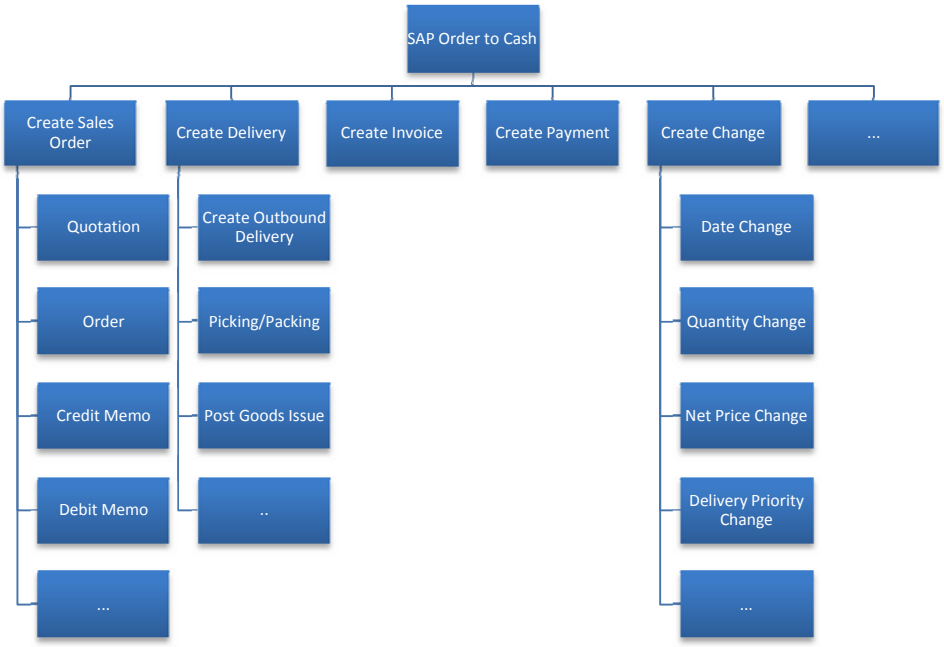


FIGURE 11: ACTIVITIES ON DIFFERENT AGGREGATION LEVELS

Current process mining tool support is not able to aggregate continuously (e.g. it is not possible to generate a low level event log and aggregate activities during the mining analysis). Therefore, the right level of aggregation has to be determined beforehand.

The different project types (explained in paragraph 0) choose their aggregation level in different fashions. It is shown that that process mining projects may require more than one event log on different levels of aggregation.

Curiosity driven: these projects usually go through several iterations of targeting problem areas and conducting root cause analysis. For the definition of problem areas, an event log with high level activities will point in the direction of problem areas, without the information overload of a large number of low level activities. However, when a problem area is defined and root cause analyses are performed the required level of detail increases. It is likely that curiosity driven projects need more than one event log on different levels of aggregation before the results are satisfactory

Goal Driven: For goal driven projects the level of aggregation is easier to determine. The direction of the problem area is known, but the exact root cause is still unknown. The logical choice is a level of detail that provides an overview of the problem area with enough details to point out the root causes.

Question Driven: Concrete questions answer the question of the right aggregation level in itself. For example “What are the top-ten changes on standard sales order?” requires a low level of aggregation on changes. For every metric that needs to be answered, a list of required activities should be made. The metrics that require similar levels of detail and have the same analysis perspective can be combined into an event log.

Summarizing can be said that the first activity selection should be on a high level. Select just the activities that show the process on a high level. The analysis will give insight and raise questions on certain parts of the process. For those parts, the aggregation level can be lowered to increase the level of detail in the activities and thereby in the analysis. Iterate over these two steps until the analyses deliver satisfactory results.

The example above shows that the aggregation level of activities depends on the type of process mining project. Similarly, the level of aggregation can depend on the type of process mining or the process mining perspective that is chosen. In practice, a combination of the three will be used to determine a set of activities that suits the project scope. The set of activities will lead to one or more event logs.

Level of Process Integration with IT

The final consideration in the activity selection is the requirement that the activity has to be supported by an IT system. Process Mining is a data driven analysis approach: without data, no analysis. In a real life business process a lot of hidden activities can occur, even if the main flow of a process is supported by IT system, Hidden activities are activities in the process that are not visible in IT systems. Hidden activities can have very different forms: for example quick consultations between colleagues, walking from one workstation to another, classifying a case ‘complex’ or ‘simple’, etc. The hidden activities are impacting the process, but they are not logged and therefore not part of the process mining analysis De Medeiros & Weijters (2005) . As a result, the mining analysis can be biased: It might well be possible that an analysis from a time perspective gives a very reassuring result, but only because a couple of very time consuming activities are performed outside the IT system.

IT systems do not only support activities, usually some activities in the process are automated steps where no human step is required. The automated steps can be added as activities to the event log, but if the analysis is focused on manual steps, the automated steps are of less importance.

As a conclusion: selecting the activities in a process mining project might seem a trivial step at first. However, a lot of important design decisions for the event log are made in this stage. It is therefore important that the list of activities that are included in the event are confirmed with the process owners and put in the right context (automated or manual, hidden activities, level of aggregation, analysis perspective).

CASE ID DETERMINATION

The next challenge is to find a case ID that is the common denominator of the selected activities. The selection of the case ID strongly relates to the aggregation level of the activities discussed in the previous paragraph. Low level activities might require a different case ID than high level activities.

Where process centered software systems are organized around cases (paragraph 0), ERP systems do not have a predefined case. SAP is object centered in nature, it stores documents that are related to activities in the sales process. For example, the creation of a sales order results in the creation of a sales document, shipping the goods will result in a delivery document, preparing the invoice gives a payment document, etc. Furthermore, one sales document can relate to one or more deliveries, and every delivery can relate to one or more payment documents, etc. (Figure 12). In order to analyze a SAP process from a process perspective, the transformation from object centered to process centered is a step that needs to be performed by the process miner in the data selection phase. The most important step in the transformation from object centered to process centered is the definition of an appropriate case ID. This paragraph will show how a suitable case ID can be derived from an ERP system by using the OtC process as an example.

An appropriate case ID is a case ID that generates an event log with the right level of detail and the right characteristics to realize the project goals. Consider the following situation in the sales process of a production firm (Figure 12). The figure shows that a case ID for the highest level of detail for the sales process is chosen:

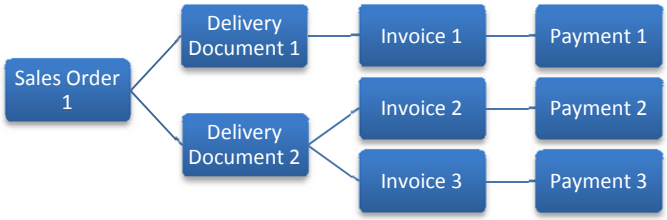


FIGURE 12: VISUALIZATION OF THE SALES EXAMPLE

1) A sales order is created: 60 items of product 1, and 50 items of product 2
2) The demand for product 1 can be delivered out of stock and is delivered immediately (delivery Document 1)
3) An invoice covering the first shipment is sent (invoice 1)
4) A payment is received for the first invoice (payment 1)
5) Product 2 is shipped (delivery document 1)
6) Two invoices are sent, one for product 1 and one for product 2
7) A payment is received for both invoices
8) Figure 12 represents the example described below.

Even though Figure 12 shows that Sales Order 1 results in two deliveries, three invoices and three payments, it still follows the expected flow of a sales order. Consequently, processing this sales order could be represented in the process mining map as follows in Figure 13.

FIGURE 13: EXPECTED FLOW

However, when the sales order number is used as a case ID, process mining algorithms will represent Sales Order 1 in a different way. Process mining algorithms relate all activities to a particular process instance, and order the activities based on their time stamps. Process mining algorithms will present the divergence in the data as a sequence of activities, rather than activities that occur in parallel (see Figure 14). Imagine representing 10000+ sales orders with different levels of convergence based on their sales order number. It will result in a big 'spaghetti process' (van der Aalst, 2011) that is impossible to analyze.

FIGURE 14: POLUTED FLOW DUE TO CONVERGENCE

In order to overcome this representation issue, the case ID can be chosen on a lower level of detail. Sales documents in SAP consist of different levels of detail:

- Header: Data in the header of a sales document are applicable to the entire document. This includes, e.g., customer-related data.
- Items: Each item of a sales document contains its own data. This includes, for example, material data and order quantities. Each sales document can contain multiple items, while each individual item can be processed differently. Examples are material items, service items, free-of-charge items or text items.

A lower level case ID is a concatenation of the sales order number (Header level) with the item number (item level). Looking back to example from Figure 12, Sales order 1 is split into two cases, item 1 and item 2 (Figure 15).

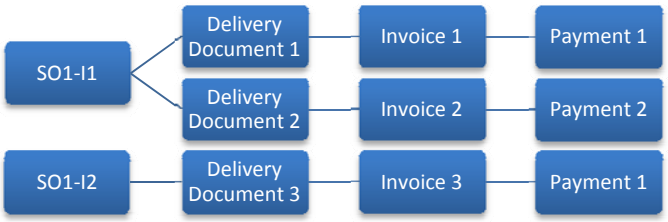


FIGURE 15: CASE ID ON ITEM LEVEL

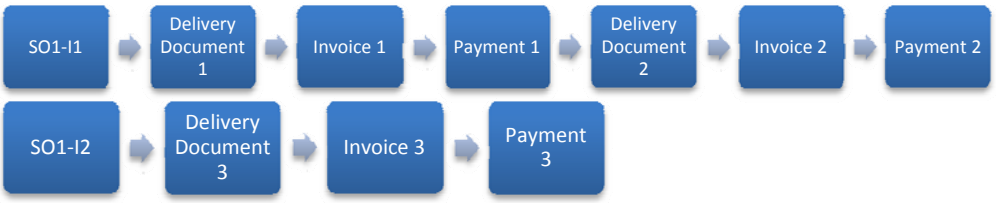


FIGURE 16: PROCESS MINING ON ITEM LEVEL

Figure 15 and Figure 16 show that part of the representation problem is solved. Sales Order 1 is split into two cases that show a separate sequence of activities in the process map. Since SO1-I2 is related to only one delivery document, invoice and payment, the expected flow is shown in Figure 16. However, SO1-I1 is related to two delivery documents, invoices and payments. The representation of SO1-I1 still differs from the expectations. However, the downside of lowering the level of detail for the case ID is the loss of intuition in the event log statistics. Figure 15 and

Figure 16 show that one *sales order* represents two *cases* in the event log. The data is contaminated with duplicate counting of sales order that are related to more than one delivery.

To remove the difference between expectation and representation, the case ID is determined at an additional level of detail. The gap between representation and expectation occurs when a document refers to more than one documents in the remainder of the process. Defining the unique combination of Sales Order Item, Delivery Document and invoice as a case closes the gap between expectation and representation.

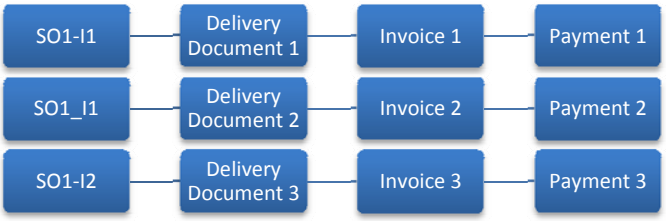


FIGURE 17: CASE ID IS UNIQUE COMBINATION OF SALES ORDER, DELIVERY AND INVOICE

Looking at Figure 17, the gap between expectation and representation is solved using this case ID. However, using a unique combination of three documents as a case ID has a downside. The number of cases and the number of sales order is the process are different. The long case ID leads to duplicate counting of various of activities. The event log statistics do not represent the process on sales order level, the level that is commonly used by companies to discuss the sales process. Gaining intuition in the visualizations of the process, means losing intuition in the event log statistics and the counting of cases and situations. Figure 17 shows three cases, but only one Sales order. Imagine a business process with 10000+ sales orders with multiple order lines, sometimes multiple deliveries and invoices per order line. The event log statistics need to be analyzed with care and a clear definition of what is represented as a case.

In the above section three levels of detail in defining the case ID are discussed. It is important to note that there is no single correct answer to give in the case ID determination. Depending on the project goals and analysis requirements, a level of detail can be chosen and the emphasis can be put on clear statistics or clear process maps. Current process mining software is not able to work with these considerations dynamically, so the level of detail in the case ID should be made explicit in the data selection phase.

Attribute Selection

In the introduction to this chapter, the minimal event log requirements are given (van der Aalst & Weijters, 2004, van Dongen & van der Aalst, 2005):

- Events correspond with activities in the business process
- All events are associated with a particular case
- The temporal order of activities must be ensured (by for example time stamps).

The activity selection and the case ID determination have been discussed in paragraph 0 and 0. The temporal order of activities is discussed in this paragraph. Furthermore, the value of additional attribute is discussed in this paragraph .

In a business process, the temporal order of activities is important. It is practically impossible to deliver goods that are not available in stock or collect a fine that is not registered in the system.

Process Mining algorithms strongly depend on the temporal order of activities: control flow discovery without knowing the order of activities is impossible. To ensure the temporal order of activities, a time stamp is required for every activity. Chronological order of all the timestamps ensure the right temporal order.

The three minimal requirements are sufficient for creating a minimal process map, but they do not give an in-depth insight in the process. The events can be enriched with different kinds of data (Figure 18):

- Resources: who performed the activity
- Data elements: customer name, size of order, etc.
- Additional Timestamps: beginning and end time of an event to measure processing times of activities

It is important to relate the choice of attributes to the goal of the process mining project. It is not possible to do an in depth analysis on throughput times of activities when there are no beginning and end time stamps available in the event log. Furthermore, when comparative analyses are required on for example business units, product categories or order values, it is necessary that they are added as attributes to the activities.

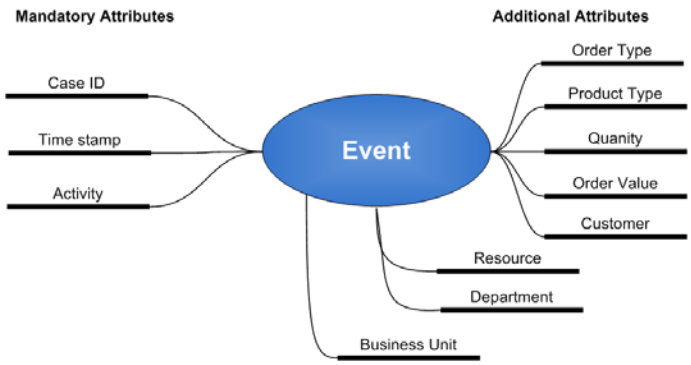


FIGURE 18: ATTRIBUTE SUGGESTIONS

Some mining algorithms require additional attributes. One of the mining perspectives described in the literature review is organizational mining. The algorithms that make for example make organizational maps or show handovers of work need resource information. At least the resource name, but preferably also the department that the resources belong to.

Many process mining tools allow for filtering the event log. Filtering can be used to make cross sections of the event log to be able to compare the control flow between different product groups or compare the temporal performance between business units. The analysis requirements in the scope phase will give pointers to the attributes that need to be added for filtering and comparison.

TIME FRAME DETERMINATION

The three sections above describe the considerations for the event log design. When the event log design is finished, the data download has to be done. An appropriate timeframe for a mining project is determined based on the mean run time of a process instance, seasonal effects and other deviations in the intensity of cases and capacity of the process mining tool.

The event log should cover a representative period of time to give a good overview of the process. The period of time representative when it captures the 'normal' behavior of the process in the real world. The 'normal' behavior is summarized in four points:

- The event log should contain more than one complete case that started and ended after each other. In some processes it takes months to complete a typical case (building permits, legal procedures). Process mining is only meaningful when multiple process instances can be followed from end-to-end.
- Seasonal effects, promotional effects or other deviations in the intensity of cases should be captured in the process. Industries that are sensitive to promotions or seasonal effect may encounter periods with 'peak' activities. These peak periods may influence throughput times, mistakes or manual steps. It can be important to the analysis that these 'peak' periods are part of the event log.
- Implementations of new systems: new information systems encounter a warm up period before the systems is running on full operational speed. In the warm up period, waiting times and quantities might not be representative for reality. Ideally, only downloads from the steady state are made.
- For a larger time frame, the number of events increases rapidly. Process mining tools do not have an infinite capacity. Mining and analysis activities can encounter performance issues, while a much smaller log might result in the same conclusions.

SALES PROCESS EXAMPLE

A set of analysis questions and metric is proposed in paragraph 0. The analysis requirements are specified for an analysis on de OtC process supported by SAP. This paragraph shows the activities and attributes that need to be included in the event log design in order to meet the analysis requirements. This paragraph is specific for the set of analysis requirements defined in paragraph 0. A similar mapping can be made for every possible set of analysis requirements to make sure that the event log design answers the analysis requirements.

Table 7 presents the analysis requirements from Table 6, and includes two extra columns showing the necessary activities and attributes to complete the analysis. The first two questions on process performance and process complexity are considered in the implementation of the process mining approach. The analysis question about process control is out of scope for this project.

Table 7 shows several types of metrics. The first type of metric does not require a specific combination of activities and attributes. An example is the number of possible process flows. The number of process flows does not depend on activities or attributes. The second type of metric compares different situations in the process. A comparison requires an attribute that enables the comparison, but is does not require a specific set of activities. The third type of metric focuses on a specific activity. For example analyzing changes requires the change activity in the event log. There is no specific requirement for attributes in this type of metrics. The last type of metrics needs a specific combination of activities and attributes. For example the throughput time between two specific activities. Date and time information is required to analyze throughput times, and the specified activities are required to measure the throughput time between them.

Goal	Question	Metric	Activities	Attributes
Optimize the SAP OtC Process	How well does my process perform?	What Percentage of my orders is a no touch order?	All	All
		How many changes occur?	All Change Activities	All
		What are my bottlenecks?	All	All
		What are my top 10 changes?	All Change Activities	All
		What is the average Throughput time?	All	Date
		What is the average Payment Term?	Invoice, Payment	Date
	How complex is my process?	What are the top 10 interface users?	All	Resource
		How many different order flows can be identified?	All	All
		How is the 80/20 distribution of unique flows divided over the cases?	All	All
		What are the differences between BU1 and BU2?	All	Profit Centre
		What are the differences between Order Type 1 and Order Type 2?	All	Order Type
		What are the differences between Product Group 1 and Product Group 2?	All	Product Group
	Is my Process In Control?	Are there any violation of the segregation of duties	Out of Scope	Out of Scope
		Is there any unauthorized behavior found	Out of Scope	Out of Scope
		Are there any indicators of fraudulent behavior found in the process?	Out of Scope	Out of Scope
		Does the process map confirm the procedural process maps defined by the company?	Out of Scope	Out of Scope
		Does the configuration of the (SAP) system support the policies and procedures in place (ITACS)	Out of Scope	Out of Scope

TABLE 7: RELATION BETWEEN ANALYSIS REQUIREMENTS AND EVENT LOG DESIGN

OUTLOOK TO CLEANING AND PREPROCESSING

When the event log is designed, the next stage is to extract the event log from the IT system. The Case ID, the list of activities and attributes are mapped to transactional data in the IT system and for a given timeframe, an event log is extracted.

CLEANING AND PREPROCESSING

Cleaning and Preprocessing the event log is used to make the raw event data suitable for process mining tools, and to increase the quality of the analysis results. Preprocessing consists of the steps that need to be taken to transform the 'flat' event data logged by the IT system to a 2 dimensional event log. As stated before, process mining tools assume that the resulting event log is correct, and will treat it likewise. Data errors, incomplete cases and infrequent behavior is considered by the algorithms as a reliable representation of the reality. In order to minimize these effects the event logs needs to be cleaned.

This section explains how the flattened event data can be transformed to an event log. In other words, how the event log designed in 0 can be extracted from the SAP System. At first, a short overview of different approaches from the literature are presented. Secondly, the approach used in this research is explained. The four elements designed in 0 are mapped on the SAP system. This chapter describes the mapping for the OtC process step by step. Furthermore, this chapter describes how the event log can be cleaned to increase to value of the process mining analysis.

Every event log design requires a design specific mapping to the information system. Therefore, this step cannot be generalized process and platform independently. The event log extraction presented in the section is however applicable on all SAP OtC processes, independent of the company context.

APPROACHES PRESENTED IN LITERATURE

Piessens (2011) presents an overview of the work that has been done on extracting event logs in SAP. The most notable approaches are summarized in this section. A complete overview is found in Piessens (2011).

The four steps shown in Figure 19 are recognized in all the approaches that are found in Piessens (2011). The level of automation of step three is evolving from mostly manual towards mostly automated. The three most notable implementations of the event log creation are summarized below.

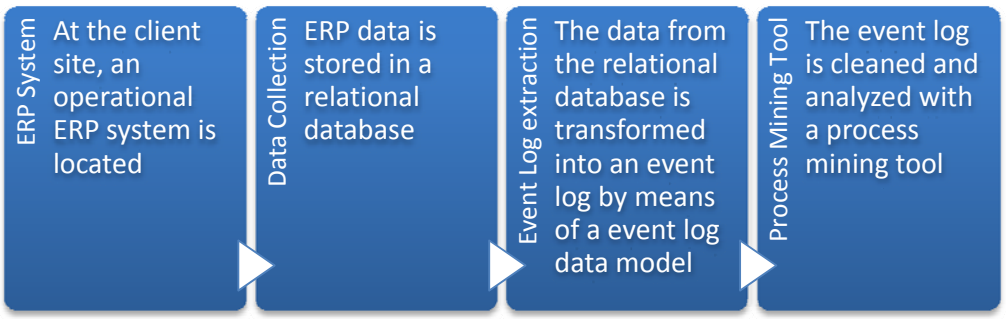


FIGURE 19: FOUR STEPS IN PREPROCESSING ERP DATA TO EVENT LOG

TABLE FINDER

Giessel (2004) uses the SAP reference models to find the relevant SAP tables from which data needs to be extracted. Furthermore, he develops a TableFinder that automates the discovery of tables related to SAP business objects that are found in the SAP reference model. He manually derives a document flow in excel and manually creates an XML event log.

XES MAPPER / XESAME

Buijs (2010) develops a generic way to construct event logs from data sources. The XES mapper is developed, a tool that converts data stored in a relational data base to the XES event log format. The relations between the tables, and the mapping of the right fields for an event log have to be made manually. The conversion execution is automated and an event log in the XES format is constructed.

SAP LOG EXTRACTOR

Piessens (2011) develops the SAP Log extractor in his master thesis. The SAP LogExtractor is a java based tool that converts SAP tables (stored in a PostgreSQL database) into an event log. It uses a predefined process repository that consists of a list of activities with a mapping of those activities to the SAP tables. The required activities can be selected in the LogExtractor, a corresponding Case ID is determined by the tool via so called 'Case-Table Mappings' and the log is extracted into a .csv file.

Both Buijs (2010) and Piessens (2011) address the problem of convergence and divergence in the data (paragraph 0) that complicates the definition of a case ID.

EXTRACTING AN EVENT LOG FROM SAP

The three master theses conducted in the past 8 year show a development towards a generic and automated approach for extraction event logs from SAP. Concepts and ideas from those theses have been used in this research. However, the event log extraction in this thesis is done on the existing data analysis platform at KPMG IT Advisory. KPMG IT Advisory uses SQL servers for the analysis of the SAP systems of their clients. A SQL database with data downloads of SAP tables are the starting point for the extraction of the event log.

A SQL script is developed that combines the relevant SAP tables and fields into an event log. The script is usable on all standard SAP implementations of the sales process and is tailored to the analysis requirements as described in section 0. Figure 20 show the flow from the SAP system at the client site, to an event log cleaned and ready for process mining.

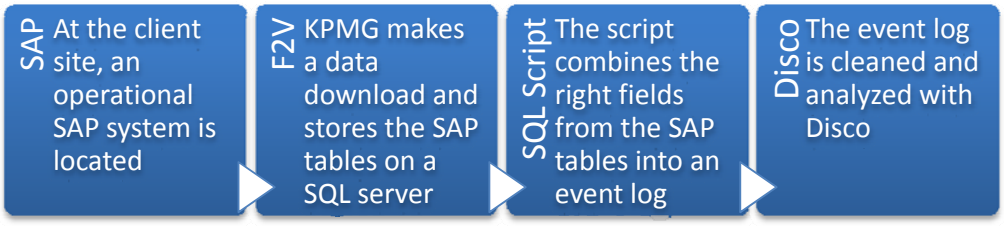


FIGURE 20: FROM SAP TO DISCO

Activity Selection

The activity selection is based on the four criteria presented in paragraph 0: process boundaries, analysis perspectives, aggregation level and level of IT support.

SAP Sales and Distribution (SD) is a SAP module that consists of all data and transactions that are needed to perform the sales process. Despite the fact that SAP SD integrates with several other SAP modules (Logistics Execution for the availability checks of sold materials, etc.), the process is defined as all activities that are performed within the SD module. However, an exception is made for payments. The payment that is related to the sales order is recorded in the SAP Financial accounting module (FI). It is the intuitive end of the sales process and will be implemented as such.

In the sales process in SAP, 5 standard documents are recognized:

- Sales Document (SD)
- Delivery Document (SD)
- Billing Document (SD)
- Payment Document (FI)
- Change Document (ECC)

These five documents are the document flow in a SAP sales process. Therefore, they are the backbone of the SQL script.

The sales process will be analyzed from a sales order perspective. From the five documents actions related to the sales order are identified as activities. With the creation of each document, one or more activities are recognized. The posting of a sales document can be the result of the creation of a sales inquiry, but also of the creation of a sales order or a contract. The level of aggregation is such that it is possible to distinguish between those two. In order to do so, the document type is used. Every document posted in SAP Sales and Distribution (the module where the sales process is executed) has a certain document type (‘A’ for Inquiry, ‘B’ for Quotation, etc.). Every document type relates to a creation activity, a full list of all the document types is found in Appendix A.

Customer payments are the final step in the sales process. The booking of a payment generates a payment document. The generation of the payment document is chosen as the ‘create payment’ activity.

Changes to each of the documents are possible and result in the creation of a change document. The creation of a change document is chosen as an activities. Changes occur on various field in the documents, such as date changes, quantity changes, price changes, etc. It is possible to specify every type of change as a separate activity. However, to increase the readability and understandability of the process maps, the changes are added as activities on an aggregate level. A change on the four documents results in four change activities. The characteristics of each change are stored in additional attributes which allow for detailed analysis of the changes in the sales process.

The mapping of these activities to the SAP tables is explained in Appendix B

CASE ID

In order to relate the activities to each other in process instances, a Case ID is determined. Following the discussion in paragraph 0, in determining a Case ID SAP a trade-off has to be made between intuitive event log statistics and intuitive process maps. For this event log, the intuitive process maps are the most important. The event logs are designed for consultants: in the communication with their clients they considered intuitive process maps as more valuable because the strong visual capabilities of process mining attracts the clients attention and directly opens the discussion.

Following the specifications from the consultants, the Case ID is chosen on the lowest level of detail discussed in paragraph 0. The unique combination of a sales order, delivery and invoice is considered a case.

Case	Sales Order	Delivery	Invoice
1	1000	2000	3000
2	1000	2001	3001
3	1000	2001	3002
4	1001	2002	3003
5	1002	2003	3004
6	1002	2003	3005
7	0000	2004	3006

TABLE 8: EVERY UNIQUE COMBINATION IS A CASE

Not all sales orders are related to a delivery or an invoice. Table 9 shows the possible document flows in an SAP system. The 'regular' flow of a sales order has a sales document, a delivery and an invoice. However, some sales orders do not have all three documents available in the process. This can be due to canceled orders, or other flows in the sales process such as direct invoices or deliveries without reference.

	Sales Document	Delivery	Invoice
'normal' flow	X	X	X
No invoice (i.e. canceled order)	X	X	
No Delivery (i.e. service)	X		X
No delivery, No Invoice (i.e. quantity contract)	X		
No Sales Document (i.e. Delivery w/o reference)		X	X
Only Delivery		X	
Only Invoice (i.e. Direct invoice)			X

TABLE 9: POSSIBLE DOCUMENT FLOWS

In cases that do not have the three related documents, the case ID is complemented with zero's for the missing document(s), indicating that no reference documents are available. Case 7 in Table 8 shows a delivery without reference, with a case ID '000020043006'. It is ensured that documents with reference 0000 are not found in the dataset (i.e. not contained in the timeframe, or not existing).

ATTRIBUTES

Besides the Case ID and the activity, the last minimal requirement for an event log is a timestamp to ensure the chronological order of events in a process instance. All the documents in SAP create the time and date of their creation. The timestamp is thus available and added as an attribute to the event log.

However, data downloads from a Clients SAP systems usually consists of a selection of SAP tables, and per SAP tables a selection of fields. In case one of the tables only the date was included in the download scope and not the time. It is than only possible to distinguish the activities per day, and not the order on that particular day. Since the order of events is the foundation of Process Mining this is a problem that will create noise in the event log. However, the SAP Sales process is structured such that the involved documents are created in a particular order. Thus, the sequence of events on a particular day can be hardcoded into the event log Table 10. The result of this measure is that every creation of a sales document is recorded at midnight, on the date provided by the SAP table. Logically, the temporal analysis is flawed by this measure. The time between the creation of a sales document and a delivery document on the same day is always 6 hours.

This problem however only occurs when analysis are performed on datasets that are acquired in the past. The download scope is adapted to include the invoice creation time, thereby the temporal analysis are enabled.

Activity	Date	Time
Create Sales Document	From SAP Table	00:00
Change Sales Document	From SAP Table	03:00
Create Delivery Document	From SAP Table	06:00
Create Billing Document	From SAP Table	12:00
Change Billing Document	From SAP Table	15:00
Create Payment Document	From SAP Table	18:00

TABLE 10: HARDCODING EVENT TIMES

Besides the three minimum requirements of an event log, additional attributes are added to the event log. According to the analysis requirements in paragraph 0 attributes are added to provide detailed information on the case. The attributes will be used to enable filters on different aspects of the cases. The attributes on changes allow for further research on the changes in the sales process. By adding change information as attributes, the process map maintains its clean and high level overview, but the event log contains enough information for an analysis on the specifics of each change.

A list of attributes is given in Table 11. It is important to note that this list of attributes is a dynamic list. The number of fields in the SAP tables is practically endless. Different choices in the scoping an analysis requirements will demand different or additional attributes. A mapping of the activities to the relevant fields in SAP is given in Appendix C.

Attributes			
Case ID	Material Name	Payment Status	Change Table
Activity	Company Code	Customer Name	Change Field
Timestamp	Profit centre	Order Value	Old and New values
Sales Order Number	Resource	Change Transaction	

TABLE 11: LIST OF INITIAL ATTRIBUTES

TIME FRAME

The considerations for the choice of the time frame are discussed in paragraph 0. The extraction method described in this section is not tailored to a specific industry. Therefore it is not possible to relate the timeframe of the data extraction to the mean order time or the type of industry the company is working in.

The data extraction by KPMG is usually one year of business data. Since most companies will have a mean order time that is less than one year, the data extraction scope for the event log is initially taken from all the data that is extracted by KPMG. However, if the resulting event log contains too many activities for the capacity of the process mining tool support, the script allows for the specification of a timeframe within the available data. Take into consideration that seasonal effects might be influencing the outcomes when a shorter time interval is extracted.

STRUCTURE OF THE SQL SCRIPT

The SQL script is structured is presented in Figure 21.

First all case IDs are determined and stored in a 'case library'. In the case library a number of attributes that hold for the whole case (such as company code, profit centre, customer) are stored. The case ID library is used to link all activities in the sales process to a particular case (event log requirement).

The next step is to relate the activities to the case ID and record them in the event log. The activities are taken from the relevant SAP tables, and matched with the case ID from the Case ID Library. For every document type, the activities and the specific attributes for that activity are added to the event log.

The change activities use a change information library together with the Case ID library to generate the events for the event log. For performance reasons the detailed information on the table that is changed and the field that is changed are stored in a change library. This library together with the Case ID library is used to record the change activities.

The relevant information for the Payment activities is stored in three separate tables in SAP: Open Payments, Closed Payments and Payment Headers. The open en closed payments are gathered into a payments library. This library together with the Case ID library is used to record the payment activities. The complete script is found in Appendix D.

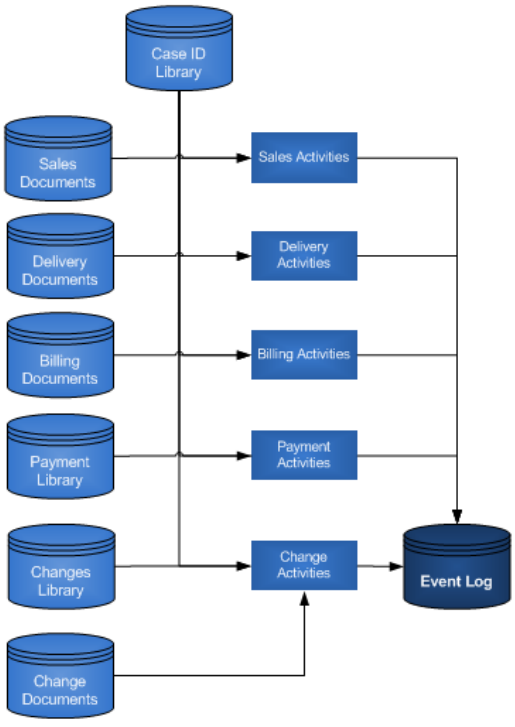


FIGURE 21: STRUCTURE OF THE SQL SCRIPT

CLEANING THE EVENT LOG

The approach described above leads to two main sources of noise in the event log. Double counting of activities and incomplete cases. In this paragraph both types of noise are briefly explained.

DOUBLE COUNTING OF ACTIVITIES

An occurrence of an activity in the event log does not have a one on one relationship with an occurrence of an activity in the real life sales process. This is caused by the way the case ID is determined. A more elaborate discussion on this problem is found in paragraph 0.

This impact of this issue is mainly seen in the analysis of the business process. Case statistics lose their intuitiveness. There is currently no solution to resolve the issue of the double counting of activities. Process analysts should be aware of the fact that the statistics are not intuitively representative for the real life situation in the process.

Possible approaches to resolve the issues of double counting can be explored in different areas. Firstly, a possible solution is to design different event logs tailored to answer a specific analysis question. Fewer activities are easier to capture in a case ID that has fewer double counting of activities. However, a large disadvantage is that one process analysis will require more event logs. The power of process mining is found in being able to analyze the process from individual cases to high level process flows. Designing different event logs makes it more difficult to fully utilize the process centered capabilities of process mining. Secondly, another solution can be found in further developments in process mining tool support. For example, include two types of case ID attributes in the event log. One 'traditional' case ID that is used to link all activities to a case and add 'statistics' attributes that are used to build the statistics. Thereby, it is recognized that it is possible that several process instances can belong to one sales order). Both options are not further explored in this research and require further research.

In this research, an attribute is added to the event log that holds the sales order number. It is then possible to relate the number of cases to the number of sales orders. The sales order is a recognizable element of discussion for the process owner. The disadvantage of this approach is that interpretation of the event log statistics requires extra attribute filters before the number of involved sales order can be extracted. This is best illustrated by an example. Consider an analysis on the top 10 types of changes on sales order for Business Unit 1. Table 12 shows the different steps that have to be taken before the answer can be given. Despite the additional analysis steps, adding statistics attributes that are analyzed by extra filters seem the best option for this research.

	Intuitive Statistics	Inclusion of 'statistics' attribute
1	Attribute filter on BU	Attribute filter on BU
2	View statistics of change activity	View statistics of change activity
3		For every change activity: attribute filter on that specific activity
4		View statistics of Sales Order Numbers

TABLE 12: ILLUSTRATION STATISTICS ANALYSIS

INCOMPLETE CASES

The event log is always extracted for a predefined period of time. Generally speaking, sales orders do not 'perfectly fit' in the time frame for which the event log is extracted. Orders will start before the event log extraction starts, or orders are not yet finished at the time of extraction. The difficulty for the notion of incomplete cases in ERP systems is that a case does not have a 'hard' start or end event. The number of possibilities to start and a process are almost limitless. By just deleting all cases that do not follow the 'standard' pattern, also 'false negatives' are deleted.

It is recommended to be conservative in defining start and end events. Especially because the complexity of ERP systems allows more process flows and shortcuts than are officially recognized. It is particularly interesting to be able to extract these 'shortcuts' with a process mining analysis (Jans, Lybaert, Vanhoof, & van der Werf, 2011).

When the start and end events are defined, incomplete cases are relatively easy to filter. Most process mining tools provide a filtering functionality that allows the definition of possible start and end events. All cases with different beginning and end events are than removed from the event log.

OUTLOOK TO MINING THE EVENT LOG

By finishing this stage in a process mining project, an event log that enables a process mining analysis is ready. The event log contains all the attributes and activities that are required for the analysis are included, and the event log is extracted from the information system. The next step is importing the log in a process mining tool that uses the event log as input for the actual analysis. The next chapter gives an overview of commercial mining tools that enable the process mining analysis of the event log.

MINING THE EVENT LOG

The event log in itself is just a way of representing transactional data. It does not give insight in the operational processes. In order to translate the information (currently structured in an event log) process mining algorithms are used.

Over the last 15 years algorithms have been developed for the different types of process mining (discovery, conformance checking and improvement) and for different mining perspectives (control flow, time, case, etc.) as presented in paragraph 0. The algorithms however, are academic achievements that do not always translate to the requirements of end-users in a business context. In the end, to a business user the algorithms are just a means to an end. He is mainly interested in the insights in the process that can be obtained by the algorithms. And maybe even more importantly in a smooth way towards these results.

Since this research is developing an approach that brings process mining to business consultants, an elaborate discussion on the technicalities of process mining algorithms is out of the scope of this research. Rather, this chapter will briefly describe a number of commercial process mining tools and make a choice for the preferred process mining used in the analysis of the case study event log.

Ailenei (2011) performed a comparative analysis of the main commercial process mining tools and tested them based on several characteristics and use cases. The comparison by Ailenei (2011) is focused on the technical capabilities of the process mining tools. For example, to what extend are they able to capture different types of patterns in a process model (Sequential activities, parallel activities, xor- and or-splits, etc.), is it possible to analyze the handovers of work, is a throughput time analysis possible, etc. However, these are only technical criteria. Elements like ease of use, capacity, speed and reliability, possible input and output formats are not considered in Ailenei (2011).

The remainder of this chapter gives a brief overview of the commercial process mining tools that are currently on the market. It is not in scope of this research to give a full analysis of the process mining and business capabilities of the different process mining tools. For a technical review of the process mining tools is referred to Ailenei (2011). The business capabilities of the process mining tools can only be evaluated for the tools that have been used in this research. Firstly five tools are described that have process mining capabilities.

PROCESS MINING TOOLS

There is number of Process Mining tools available on the market that use an event log of some sorts to do a process mining analysis of the process. The following five tools are briefly described based on the information provided by Ailenei (2011) and (in case of Disco) testing the tool in context of this research.

- Software AG – ARIS Process Performance Manager (PPM)[1]
  ARIS PPM is part of the ARIS controlling platform and it aims to analyze processes based on historical data (event log). ARIS PPM supports event logs in many different file formats, and the discovered model and statistics can be exported for reporting or further analysis.

[1] http://www.softwareag.com/nl/products/aris_platform/aris_controlling/aris_process_performance/overview/ default.asp

Besides the discovery of process models, ARIS PPM allow for the definition of Key Performance Indicators (KPI's). The defined KPI's are automatically calculated based on the event log and the results can be shown in performance dashboards. There are process benchmarking capabilities implemented that allow comparisons of models based on their structures, organizational units involved and KPI's.

Ailenei (2011) concludes that ARIS PPM covers the entire process mining spectrum and identifies that the KPI based analysis is its strongest point.

- Fourspark – Flow[2]

  Flow is a process mining tool developed by the Norwegian company Fourspark. Flow allows for the discovery and analysis of processes based on historical data. The user interface is built as a dashboard that allows showing different elements such as process maps, geographical locations, statistics, etc. Event logs can be filtered and searched to analyze parts of the process that have certain criteria.

- Future Process Intelligence – Futura Reflect/ Perceptive Software – Perceptive Reflect[3]

  Futura Process Intelligence was part of Pallas Athena, a Dutch IT company. Pallas Athena has been acquired by Perceptive Software, a Lexmark company. Futura Reflect is incorporated in Perceptive Reflect, providing similar functionality. Claims by Ailenei (2011) about the process mining capabilities of Futura Reflect will therefore be true for Perceptive Reflect as well.

  Perceptive Reflect can be used as a standalone process mining tool, or as part of the Perceptive Process suite. It has four main areas of functionality: Mine, Explore, Animate and Charting. Mine and Explore are two process discovery functions with different nuances. Mine develops a model that is representative for a user defined percentage of the cases. Explore assumes sequential behavior and will therefore not extract parallel activities. Explore allows for filtering the log and zooming in on the organizational perspective. Animate replays the case on the event log, thus allowing the identification of bottlenecks. Charting give the opportunity to create different kinds of charts for analysis.

- QPR – Process Analyzer[4]

  Process Analyzer is an excel based process mining tool developed by the Finnish company QPR. When the event log is imported in excel, all standard excel functionality is still enabled. Different process mining capabilities such as mining the control flow, analyzing process variants and filtering the log on different attributes is possible. In case the event log exceeds 1 mln events (excel limit), the process analyzer can connect with an sql server to load its event logs.

- Fluxicon – Disco[5]

  Disco is a process mining tool developed by Fluxicon. Disco allows for the analysis of processes with process mining techniques. After importing an event log, Disco provides process maps, detailed event statistics and the possibility to drill down to case level. To analyze parts of the process, Disco has various filter options (variation filters, attribute filters, time frame filters, etc.). Disco does not allow for creating organizational models.

  Hands on experience with Disco shows a process mining tool that quickly and dynamically analyzes large and complex log files. Ailenei (2011) did not review the process mining capabilities of Disco, since it was released after the research was finished.

[2] http://fourspark.no/?page_id=11
[3] http://www.perceptivesoftware.com/products/perceptive-process/process-mining
[4] http://www.qpr.com/products/qpr-processanalyzer.htm
[5] http://www.fluxicon.com/disco/

AUTOMATED PROCESS DISCOVERY TOOLS

There are several systems on the market that do not use an event log with historical transactional data from one or more corporate IT systems to extract and mine business process models. They collect user information at workstation level and combine to collect data into business events. The sequence of business events than allow for the construction of a process map and the analysis of the gathered data.

An advantage of this approach is that it is platform independent. Any application that is used to execute the process is recorded by the systems and will be part of the process model. A disadvantage can be the data model behind the business activities and the possibility to analyze the process on case level. However, this type of system has not been tested in this research and therefore strong claims about both the process mining capabilities or the business capabilities of the tools cannot be made.

A short description based on the factsheets and websites of the tools is given below.

- Iontas – Process Discovery Focus[6]
  After installing the tool on all relevant workstations, a start and end event of the process have to be designed. Process Discovery Focus automatically records what the activities in the process are and by whom they are performed. Based on the collected user information, a process map is constructed and an analysis can be performed.
- Open Connect – Comprehend[7]
  By installing the tool on all workstations, it collects user activity. The user activity is automatically clustered into business events. The business events are analyzed and can be presented in dashboards .
- Fujitsu – Interstage Automated Process Discovery[8]
  Fujitsu's interstage Automated Process Discovery collects user information by installing sensors on the workstations. The user information collected is then used to discover the process flow. The tool enables different types of analysis and dashboard features. It is however, not a process mining tool that uses an event log from existing systems. It rather collects user information that is translated to business activities.
- StereoLOGIC – Discovery Analyst[9]
  The Discovery analyst records user activity, the recorded information is automatically translated into a process map. Frequency statistics are collected and can be analyzed and the Discovery Analyst provides conformance checking functionality.

[6] http://www.iontas.com/pages/products/pdf.php
[7] http://www.oc.com/technology/
[8] http://www.fujitsu.com/global/services/software/interstage/solutions/bpmgt/bpma/
[9] http://www.stereologic.com/stereologic_software.htm

TOOL USED IN THIS RESEARCH

The choice of process mining tool support depends on the process mining capabilities of the tool. However, in a business context there are more factors that play a role in choosing the right tool for your process mining project. Integration with currently available tooling or information systems is important. Ease of use, capacity and speed might also be important factors in the choice for the right process mining tool.

In this research Disco is used as the process mining tool. The choice for disco is based on several criteria.

- Availability – Disco is made available to students and researchers via an academic initiative.
- Underlying Algorithm – Disco is based on the genetic algorithm, which allows to dynamically select the most frequent paths or activities as well as viewing all available activities and paths.
- Import/export capabilities – Disco is able to import and export all major event log formats such as txt, mxml, xes, csv and xslx.
- Ease of use – Easy to understand user interface, allows for quick and dynamic analyses
- Performance – Disco has a very good performance with event logs up to 5 million events.

SALES PROCESS EXAMPLE

In paragraph 0 a set of analysis requirements is proposed. In paragraph 0 the required attributes and activities are added to the analysis requirements. This paragraph will explain the process mining techniques that can be used to answer the analysis requirements. This paragraph is specific for the OtC process, but the process mining techniques are generally applicable on similar analysis metrics.

Table 13 shows the mining techniques that are used to answer the analysis requirements. For all the metrics a combination of statistics and filters is used. Through filters excerpts of the event log can be made based on specified characteristics. Possible filters are attribute filters, activity filters, variation filters, time frame filters, etc. Including or excluding specific parts of the event log shows the process map and accompanying statistics that give insight in the business process.

Event log statistics give general information about the event log such as variants, number of resources, process instances, etc. Furthermore, statistics are presented on all the attributes that are included in the event log. Analyzing the statistics in combination with the filters give valuable insight in the business process.

Goal	Question	Metric	Activities	Attributes	Mining Technique
Optimize the SAP OtC Process	How well does my process perform?	What Percentage of my orders is a no touch order?	All	All	Activity Filter; exclude disrupting events
		How many changes occur?	All Change Activities	All	Activity Statistics
		What are my bottlenecks?	All	All	Activity Statistics, Throughput Times between activities

		What are my top 10 changes?	All Change Activities	All	Attribute Statistics
		What is the average Throughput time?	All	Date	Time Perspective, Throughput time between the first and last activity
		What is the average Payment Term?	Invoice, Payment	Date	Time Perspective, Throughput time between the two activities
	How complex is my process?	What are the top 10 interface users?	All	Resource	Resource Statistics
		How many different order flows can be identified?	All	All	Case Statistics
		How is the 80/20 distribution of unique flows divided over the cases?	All	All	Variation Filter
		What are the differences between BU1 and BU2?	All	Profit Centre	Attribute Filter on Profit Centre
		What are the differences between Order Type 1 and Order Type 2?	All	Order Type	Attribute Filter on Order Type
		What are the differences between Product Group 1 and Product Group 2?	All	Product Group	Attribute Filter on Product Group
	Is my Process In Control?	Are there any violation of the segregation of duties	Not in Scope	Not in Scope	Not in Scope
		Is there any unauthorized behavior found	Not in Scope	Not in Scope	Not in Scope
		Are there any indicators of fraudulent behavior found in the process?	Not in Scope	Not in Scope	Not in Scope
		Does the process map confirm the procedural process maps defined by the company?	Not in Scope	Not in Scope	Not in Scope
		Does the configuration of the (SAP) system support the policies and procedures in place (ITACS)	Not in Scope	Not in Scope	Not in Scope

TABLE 13: RELATION BETWEEN ANALYSIS REQUIREMENTS AND MINING TECHNIQUES

OUTLOOK TO ANALYSIS AND INTERPRETATION

This chapter gives a brief overview of process mining tools. Just the tool does not deliver a process analysis. An additional steps is needed to interpret the results and give business meaning to the process characteristics and statistics that are made visible by the process mining tool.

ANALYZE AND INTERPRET MINING RESULTS

The final step in a process mining project is the stage of analyzing and interpreting the results from the mining tools. Process Mining tools give a lot of process characteristics and statistics that needs to be translated into valuable business information.

The literature does not present a general and structured approach for this stage. The literature study by Roest (2012) states that practical applications of process mining in academic literature hardly report on the analysis and interpretation of the mining results.

This chapter builds on the experience of introducing Process Mining in the KPMG environment, and conducting a case study with a KPMG client. In order to generalize the findings presented in this chapter, further research is needed.

Analysis and interpretation of the results links back to the project scope determined in 0. The choice of the project type, and the analysis requirements agreed upon with the client are leading in the way the mining results are analyzed and interpreted.

In the remainder of this chapter, a number of general focus points are discussed and the specifics for the three different project types are highlighted.

ANALYSIS IN ITS CONTEXT

A business process does not exist without its context, and should be analyzed likewise. Process mining is a good way to show the process owners what is going on in their system, and observations can always be made. However, finding root causes and redesigning (parts of) a process can never be done without cooperation with the process owners (van der Aalst, 2011). Process owners have tacit knowledge that is not contained in the corporate it systems (and thus not in the process mining results), but is vital for the interpretation of the mining results.

In the process of analyzing the mining results it is important to plan one or more discussion sessions where the initial observations are discussed and put into context. The practical experience from the process owners will help to give value to the mining results, and will most likely lead to a refinement of the analysis and the identification of new analysis directions. The root causes of the observations in the process are extracted in a cooperation of fact based process mining results and practical and tacit process knowledge of the practitioners.

INTERPRETATION RELATED TO THE PROJECT TYPES

In section 0 the three types of process mining projects are described. The type of process mining project determines for a large part the way the mining results are realized and require a slightly different approach of giving them value. This section will discuss the analysis and interpretation specifics for the three types of process mining projects.

CURIOSITY DRIVEN PROJECT

In a curiosity driven project, the analyst would be about the research without a predefined set of questions or research objective. Analyzing without a predefined set of requirements makes the interpretation of results difficult. Process mining allows for many different analysis options (mining perspectives, filter options, aggregation levels, etc.).

To give some structure to the analysis in a curiosity driven project, a 'layered' analysis approach can be helpful. In a layered approach the process is analyzed at a high level first. Focus points can be frequencies of activities or resources, control flow patterns, etc. Based on this high level analysis, one or more focus areas can be defined in cooperation with the client. Possible focus areas can for example be: analysis of the returns flow, focus on sales order changes, where do the differences in control flow come from, etc. The results of the second layer can be the basis for a third layer, etc. By systematically drilling down in the process in cooperation with the customer structures the analysis and interpretation of results for curiosity driven projects and maximizes the business value of the process mining project.

GOAL DRIVEN PROJECT

In goal driven projects, the first layer of focus areas is known beforehand by the customer and is defined in the scoping phase. For goal driven project, this allows the analysis to drill down the process more quickly and focus on the problem areas from the beginning of the project.

Similar to curiosity driven projects, drilling down the process in an analysis requires the tacit knowledge and the business context that can only be provided by the process owners and the customers.

QUESTION DRIVEN PROJECT

For question driven projects, the scoping phase has been defining specific questions that need to be answered by means of a process mining project. The context of the business process is already known at the time of the scoping stage and implemented in the analysis requirements.

However, even when the questions are specific and are asked in their specific business context. Answering the questions and drawing conclusions from those answers still requires cooperation with the customer to actually drill down to the root cause of the answers as they appear in the process mining analysis.

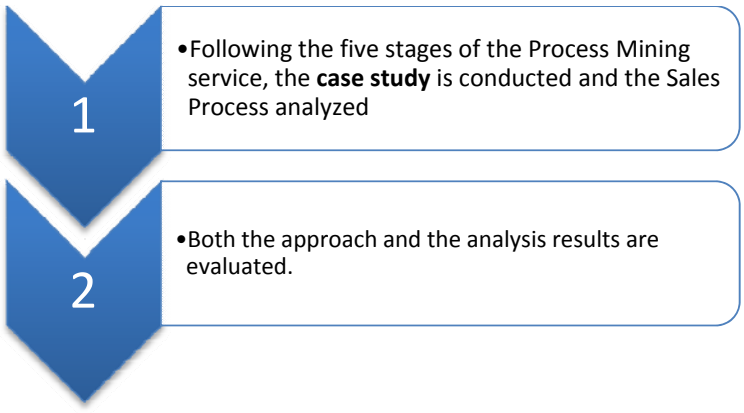
CONCLUDING REMARKS

In the sections above, it is repeatedly emphasized that cooperation with the customer and feedback on the mining results is essential for obtaining valuable analysis results. It can be a pitfall to rely completely on the feedback of the customer and process owners. The analysis possibilities of process mining are covering many aspects (see 0) and allow for new inventive ways to analyze business processes. The creativity and curiosity of the analyst can still result in remarkable insights in the business process.

Without weakening the call for cooperation and putting results into context, the keywords for a valuable process mining analysis are iteration and creativity in the analysis process.

PART III: CASE STUDY

The process mining approach developed in Part II is tested in a real life situation. The five stages of the process mining service are conducted at a Dutch publisher of educational material. After the execution of the case study, both the approach and the results are evaluated.



CASE STUDY

This case study is testing the 5 stage process mining approach developed in this research. The case study is conducted at an European Dutch based education company. They have a SAP single instance system supporting three operating companies in three different countries. For this case study, one company code is considered due to the license limitations of Disco. The data is made available through a data download conducted by KPMG IT advisory and is located on a secured server.

The case study is structured by following the 5 stages of the process mining approach developed in the research. Through the case study several aspects of the approach are tested. Practicality and feasibility of the approach are tested in the case study. Furthermore, it is verified that the approach leads to a useful process analysis with business value for the client. The complete process analysis is found in 0.

SCOPING

Besides a generally applicable process mining approach, Part II also develops an event log extraction script that enables an initial process analysis without consulting the client. The event log that resulted in executing the script is used for the first step in the case study.

In an interactive workshop with the process owner of the case study company (in the remainder referred to as company), insight is given in the sales process of the company. Being able to introduce process mining to the company with an interactive workshop on their own process data is considered a powerful approach.

The workshop resulted in eight analysis questions (Table 14). The first three questions are related to a comparison. The fourth questions asks for the analysis of a specific flow. Question zooms in on throughput times for certain orders. The last three questions focus on change actions

	Question
1	Compare the control flow of the different business units to be able to identify differences and improvement possibilities
2	Compare the control flow for different sales order types to be able to identify differences and improvement actions
3	Compare the control flow for different product groups to be able to identify differences and improvement actions
4	Identify the returns process: activities, throughput times between order and returns delivery per business unit and per customer category
5	Average time between order intake and delivery for sales orders with material status 02 or 05
6	Identify the first time right (orders without manual order changes) sales order, divided per business unit and per order type
7	Distinguish the different change activities explicitly, manual changes vs. automated changes, date change vs. price change, etc.
8	Identify the top 10 manual change actions for different sales order types

TABLE 14: ANALYSIS QUESTIONS CASE STUDY

EVENT LOG DESIGN

The event log design is refined where necessary to be able to answer the questions presented in Table 14.

The required activities to answer question 1-7 do not differ from the standard list of activities in the standard event log. Question 8 requires a lower level of detail in the activity change sales order. In the standard event log, this detail is added in specifying which fields and which tables are changed in the change sales order activity. Since a specification to separate change activities is implicitly captured in the standard event log, it is decided to keep the specification to separate change activities the same as in the standard event log.

The main point of interest for the client is to cluster the order flows due to several characteristics such as business unit, material group and order type, customer group, etc. The possibilities for clustering have a close relationship with the attribute selection. When for example the order flows are analyzed per business unit, an attribute determining the business unit should be added to the event log.

Table 15 shows the attributes that will allow answering the analysis questions. Question three and five require attributes that are not contained in the standard event log. The attributes that are needed to answer those questions are not contained in the download scope of the KPMG data download. It is therefore in this case study not possible to answer these questions.

Question	Attributes
1	Profit Center
2	Sales Order Type
3	
4	Activity Type
5	
6	Resource, Profit Center, Order Type
7	Resource, Table & Field Change
8	Resource, Sales Order Type

TABLE 15: RELATION BETWEEN ATTRIBUTES AND ANALYSIS QUESTIONS

PREPROCESSING AND CLEANING

The preprocessing and cleaning stage transforms the event data logged in SAP to an event log that is usable in process mining tools. Furthermore, the event log is cleaned to increase the value of the results

The incomplete cases are filtered from the event log. Eight start events are found in the event log. The event ‘Change Sales Order’ is the only event that indicates an incomplete cases. Cases starting with ‘ Change Sales Order’ are therefore excluded from the event log. For the end events, it is a bit more complicated. For cases that end in a ‘strange’ event (i.e. Order w/o Charge) it is not sure if it is an incomplete case that is finished out of the time frame, or that it is a case that is ‘stuck’ in the process and will never be finished. Only instances with end event ‘order’ are excluded from the event log.

Application of Mining Techniques

The event log that is prepared in the previous stage is analyzed with Disco. Disco has three areas of functionality: the process map, event log statistics and information on case level. Furthermore, it allows the analyst to apply different types of filters on all attributes of the event log.

Answering the questions determined in the scoping stage (Table 14) require filtering the event log on different attributes. When a comparison is asked between for example business units or sales order types. The event log is filtered on the specific attribute: and process maps, statistics and cases can be compared.

Analysis and Interpretation of the results

The process of analysis and interpretation of the results is described in this section. Furthermore the management summary that is reported to the client is found in this section. The full report is found in 0.

Analyzing Mining Results

This stage contained iterations of refining the analysis and adding business value to it.

1. After the scoping workshop, an initial report covering the eight analysis questions was made. This report was found to be mostly a summary of event log statistics without business value.
2. In a workshop with a data analysis expert, the analyses were refined giving value to the bare statistics. The resulting report was able to translate the statistics into observations that contained business value.
3. In a workshop with a business consultant OtC, the observations were refined and next steps or recommendations were added to the observations.

Following these three iterations shows that analysis without its context is delivering interesting statistics about the process. However, interactive session(s) with process owners or process consultants are needed before the statistics can be translated into observations and recommendations.

Management Summary

The management summary contains of two parts. Firstly, the two main observations of the process are given. Secondly, a summary of the answer to the eight questions from Table 14 is given (Table 16).

Main observations (Process Profiling):

- We have observed that 80% of the Order to Cash (OTC) transactions (68,000 sales orders) in SAP (cases) are handled via 10 unique process flows. Given the level of complexity of the OTC business process (and level of variants) this seems adequate (throughput times are subject to further investigation).
- However, we have also observed that 20% of the OTC transactions (17,000 sales orders) in SAP are handled via more than 1,900 unique process flows. Handling all these exceptions impacts process efficiency (is time-consuming). We have also observed that the throughput time for certain sales orders is rather (too?) long (0, Slide 10, page XXVI). Standardization of these sales order processes will lead to significant process improvements resulting in material cost savings and an improved uninterrupted order flow.

Abbr. question	Summarized Answer (details in 0)
Q1. BU comparison	• BU1 (47% - 39,988) and BU2 (41% - 38,569) have most sales orders in SAP. (BU3, 7% - 6,685). • BU2 has 67% promotional orders, where BU1 has 31% and BU3 32% (rest is mainly standard orders) • BU1 has an emphasis on contracts (64%), BU2 (<1%) and BU3 (5%) mostly have Standard Orders.
Q2. Order Type comparison	• There are several order types in use. The order types standard orders, promotional orders and contracts are used in most cases. See detailed observations
Q3. Product Group comparison	*Not possible to answer with current data model. Further Research Required.*
Q4. Return Process	Approximately 2% of the sales orders are returned (1495 sales orders). Most returns are observed for BU2 (1,135).
Q5. Throughput times for materials with status 02 and 05	*Not possible to answer with current data model. Further Research Required.*
Q6. First Time Right Orders	Sales orders are changed often (25%) after initial registration. There are different reasons for changing sales orders. See top 10 changes for reasoning.

Q7. Change Activities	• Mostly schedule changes in Promotional Orders by the System User
	  • 95% of the changes are performed by the System User
	  • 95% of the System User changes are schedule changes (Process inefficiency)
	• Dialog Users change: PO Number (37%), Billing date (13%), Reason for Rejection (7%), Schedule (5%)

Q8. Top 10 changes per Order Type	• Standard: PO Number (45%), Billing Date (17%), Terms of Payment (6%), Schedule (5%) • CNTRCT1, CNTRCT2, CNTRCT3, CNTRCT4: Reason for Rejection (30%), Pricing Date (17%), Reason for Cancellation (14%) • Promo: Schedule (22%), Confirmed quantity (9%)

TABLE 16: SUMMARIZED ANSWERS TO ANALYSIS QUESTIONS

CASE STUDY EVALUATION

The case study is used to test practicality and feasibility of the process mining approach. Furthermore, the value of the process analysis is evaluated.

The approach is showed to be highly practical. It is possible to extract an event log and deliver a process mining analysis using the five stages. The event log extraction script covers the main interests of the case study company and also of the interviewed consultants. The feasibility of the process mining approach is showed in the case study as well. Without knowledge of the SAP implementation at the case study company, it was possible to extract an event log and answer most of the analysis questions. It is not possible to generalize the practicality and feasibility o the approach since it is tested in a single case study. The extraction script is based on the standard SAP tables that have been downloaded over 200 times by KPMG as part of their data analysis platform. It is expected that it is possible to extract an event log on all KPMG data extractions.

The reference to the KPMG data download is also a weakness in terms of practicality and feasibility. The approach assumes that the relevant SAP tables are available in a relational SQL database. Data availability is a critical success factor, and the approach does not cover the access to SAP tables.

The analysis report is presented to the case study company. The audience existed of line managers, process managers, an enterprise architect and a member of the management team. The analysis results were received positively. It was noted by the audience that the analysis results were both reassuring and challenging. The dominant flows, customers and order types were successfully identified by the analysis. Furthermore, highlighting the large number of process variants and the large number of changes on promotional orders challenged them to follow up on those issues.

CONCLUSIONS

This final chapter will highlight the main conclusions following from the research. In the first section the research is summarized. In the second section the practical implications of this research for KPMG professionals and practitioners in general are highlighted. Finally, directions for further research are identified, both for the five step process mining approach and the implementation of the approach for SAP Order to Cash (OtC).

SUMMARY

The aim of this research was to develop an approach for consultants to do a business process analysis with process mining techniques. A literature review on the current state of the art of process mining and its practical applications is done. It was found that process mining research has a focus on the development of mining algorithms. However, for practitioners there is more to a process mining project than just applying mining techniques. Process mining practitioners deal with challenges like constructing the right event log from complex IT application logs and having to translate mining results into business value.

These challenges are addressed in the approach that is developed in this research. The structured approach for process mining projects is developed based on the literature and interviews. The execution of a process mining project is captured in five stages.

1. Define scope: The scope sets the boundaries for the process mining projects and it gives the necessary focus by defining a clear set of project goals and analysis requirements.
2. Design event log: The analysis requirements and project goals are translated into an event log design. An event log contains four elements:
3. Preprocess and clean event data: The event log design is mapped on the IT landscape enabling the extraction of the right information to construct an event log. The translation of the event log design to a data model is implemented for the SAP OtC process. Furthermore, event logs contain noise (i.e. double counting of activities or incomplete cases) that needs to be filtered from the event log to get the best results.
4. Mine event log: Process mining tools are used to analyze the event log. A list of commercial process mining tools is presented in this stage. Disco is chosen as the process mining tool in this research.
5. Analyze and Interpret mining results: Several iterations of analysis and interpretation in cooperation with process managers are needed to translate the factual process characteristics and statistics. This final stage of knowledge discovery works best in a cooperation of process mining experts and process owners to combine mining expertise with tacit knowledge of the process and its context.

The approach and the implementation of the event log for SAP OtC are tested in a case study with a Dutch publisher of educational material. The approach is found to cover all major challenges and issues that are encountered during the execution of the case study. The way the case ID is chosen in the event log design leads to a double counting of activities in the event log. This double counting makes the analysis of the event log statistics less intuitive. This issue is sufficiently covered by adding ‘statistics’ attributes that make it possible to relate the statistics back to an easy to grasp unit of measure. The results of the analysis are presented to the management team and were well received by the case study company. The results were reassuring in showing what their dominant product flows and customers are. But also renewing by pointing towards improvement possibilities and flaws in the sales process.

PRACTICAL IMPLICATIONS

The results of this research are twofold: a process and platform independent approach to guide process mining on ERP systems (1). Furthermore, an implementation of this approach for the SAP Order to Cash process is developed that enables the analysis of all Order to Cash process independent of the company context (2).

The practical implications of (1) are

- The gap between process mining research and practitioners has been bridged by guiding an end-to-end process mining project with all conceptual and practical considerations
- The main considerations of process mining projects on ERP systems are addressed. It enables the design of event logs on ERP systems (process and platform independently) for practitioners.

The practical implications of (2) are:

- The implementation of the event log extraction script enables process mining on the implementations of the OtC process in SAP (regardless of the specific company).
- The event log script offers an answer to a standard set of analysis questions. Hereby, it is possible to give the client insight in their processes during initial scoping discussions. This is found to be a powerful approach.

LIMITATIONS AND FURTHER RESEARCH

The research has several limitations and directions for further research. Both limitations and directions for further research are summarized below:

- Due to technical complexity of mining ERP systems, the first 3 stages of the process mining approach have received the most attention. Stage 4 and 5 are explained on a higher level. Further research is needed to directly link analysis requirements with activities, attributes, mining tools and filters. A solution direction is to develop a catalog that contains a list of analysis questions. For each analysis question, a step-by-step guide from activity and attribute selection, towards the choice and application of process mining tools.
- Due to the object centered structure of ERP systems it is difficult to find a suitable case ID. Due to convergence and divergence observed between different process steps the case ID is not easily determined.
  This case ID problem is recognized, but no full solution is found. Two solutions directions are proposed: a dedicated event log for every analysis question will solve most of the problem. However, that approach loses the powerful dynamic of process mining to analyze both high level and case specific in one event log. Another solution direction is the introduction of an extended case ID in process mining tools, one case ID is used to relate all activities to one case (traditional case ID), the other part is used to calculate appropriate statistics (extended 'statistics case ID').
- SAP lacks Start and End times for its activities. It is impossible to trace how long it took to complete an activity. It is therefore hard to elaborate on throughput times and processing times.
- This research focused on the analysis of the process flow and case statistics. Process mining offers more types and perspectives such as organizational mining and conformance checking. Further research is needed to include those types and perspectives of process mining into the different stages of the process mining approach.

BIBLIOGRAPHY

van der Aalst, W. M. P. (2011). Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer.

van der Aalst, W. M. P., Andriansyah, A., Alves de Medeiros, A. K., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., et al. (2012). Process mining manifesto. *BPM 2011 Workshops Proceedings*, 169–194.

van der Aalst, W. M. P., van Dongen, B. F., Herbst, J., Măruşter, L., Schimm, G. & Weijters, A. (2003). Workflow mining: A survey of issues and approaches. *Data \& Knowledge Engineering*, *47*(2), 237–267. Elsevier.

van der Aalst, W. M. P., Reijers, H. A., Weijters, A. J. M. M., van Dongen, B. F., Alves de Medeiros, A., Song, M. & Verbeek, H. (2007). Business process mining: An industrial application. *Information Systems*, *32*(5), 713–732. Elsevier.

van der Aalst, W. M. P. & Weijters, A. (2004). Process mining: a research agenda. *Computers in Industry*, *53*(3), 231–244. Elsevier.

Ailenei, I. M. (2011). Process Mining Tools: A Comparative Analysis. *Master Thesis, Eindhoven University of Technology*.

Buijs, J. (2010). Mapping Data Sources to XES in a generic way. *Master Thesis, Eindhoven University of Technology*.

Caldiera, V. R. B. G. & Rombach, H. D. (1994). The goal question metric approach. *Encyclopedia of software engineering*, *2*, 528–532.

Davenport, T. H. & Short, J. E. (1990). The New Industrial Engineering: Information Technology and Business Process Redesign. *Sloan Management Review*, *31*(4), 11–28.

van Dongen, B. F. & van der Aalst, W. M. P. (2005). A meta model for process mining data. *Proceedings of the CAiSE*, *5*, 309–320.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Giessel, M. van. (2004). Process Mining in SAP R/3. *Master Thesis, Eindhoven University of Technology*.

Jans, M., Lybaert, N., Vanhoof, K. & van der Werf, J. M. (2011). A business process mining application for internal transaction fraud mitigation. *Expert Systems with Applications*. Elsevier.

Lillrank, P. (2003). The quality of standard, routine and nonroutine processes. *Organization Studies*, *24*(2), 215–233. Sage Publications.

Mans, R. S., Schonenberg, H., Leonardi, G., Panzarasa, S., Cavallini, A., Quaglini, S. & van der Aalst, W. (2008). Process mining techniques: an application to stroke care. *Studies in health technology and informatics*, *136*, 573. IOS Press; 1999.

Mans, R. S., Schonenberg, M., Song, M., van der Aalst, W. M. P. & Bakker, P. (2009). Application of process mining in healthcare-a case study in a dutch hospital. *Biomedical Engineering Systems and Technologies*, 425–438. Springer.

De Medeiros, A. K. A. & Weijters, A. (2005). Genetic process mining. *Applications and Theory of Petri Nets 2005, volume 3536 of Lecture Notes in Computer Science*.

Piessens, D. A. M. (2011). Event Log Extraction from SAP ECC6.0. *Master Thesis, Eindhoven University of Technology*.

Roest, A. (2012). Practical Applications of Process Mining. *Literature Review, Eindhoven University of Technology*.

Rozinat, A., de Jong, I. S. M., Günther, C. & van der Aalst, W. M. P. (2009). Process mining applied to the test process of wafer scanners in ASML. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *39*(4), 474–479. IEEE.

APPENDIX A SD DOCUMENT TYPES

SD Document Type	Description
A	Inquiry
B	Quotation
C	Order
D	Item proposal
E	Scheduling agreement
F	Scheduling agreement with external service agent
G	Contract
H	Returns
I	Order w/o charge
J	Delivery
K	Credit memo request
L	Debit memo request
M	Invoice
N	Invoice cancellation
O	Credit memo
P	Debit memo
Q	WMS transfer order
R	Goods movement
S	Credit memo cancellation
T	Returns delivery for order
U	Pro forma invoice
V	Purchase order
W	Independent reqts plan
X	Handling unit
0	Master contract
1	Sales activities (CAS)
2	External transaction
3	Invoice list
4	Credit memo list
5	Intercompany invoice
6	Intercompany credit memo
7	Delivery/shipping notification
8	Shipment
a	Shipment costs
b	CRM Opportunity
c	Unverified delivery
d	Trading Contract
e	Allocation table
g	Rough Goods Receipt (only IS-Retail)
h	Cancel goods issue
i	Goods receipt
j	JIT call

p	Goods Movement (Documentation)
r	TD Transport (only IS-Oil)
s	Load confirmation, reposting (only IS-Oil)
t	Gain / loss (only IS-Oil)
u	Reentry into storage (only IS-Oil)
v	2-step Goods receipt (only IS-Oil)
w	Reservation (only IS-Oil)
x	Load confirmation, goods receipt (only IS-Oil)
$	(AFS)

TABLE 17: SD DOCUMENT TYPES

APPENDIX B MAPPING ACTIVITIES TO SAP

Activity	
Create Sales Document	Posting of a sales document in VBAK. The document type VBTYP is the activity name
Change Sales Document	Posting of a change document in CDHDR, related to the case
Create Delivery Document	Posting of a delivery document in LIKP. The document type VBTYP is the activity name
Create Billing Document	Posting of a billing document in VBRK. The document type VBTYP is the activity Name
Change Billing Document	Posting of a change document in CDHDR, related to the case
Create Payment Document	Posting of a financial document in BKPF.

TABLE 18: MAPPING OF ACTIVITIES TO SAP

APPENDIX C MAPPING ATTRIBUTES TO SAP

Attribute	Sap field
Resource	CDHDR.USERNAME, VBAP.ERNAM, LIKP.ERNAM, VBRP.ERNAM, BSID.USNAM, BSAD.USNAM
Time Stamp	CDHDR.UDATE, CDHDR.UTIME, LIKP.ERDAT, LIKP.ERZET, VBRP.ERDAT, VBRP.ERZET, BSID.CPUDT, BSAD.CPUDT
Sales Order Number	CDPOS.TABKEY, LIPS.VGBEL, VBRP.AUBEL, VBAP.VBELN
Company Code	TVKO.BUKRS
Company Code Descr	T001.BUTXT
Customer Name	KNA1.NAME1
Order Type	TVAKT.BEZEI
Material	MAKT.MAKTX, VBAP.ARKTX
Order Value	VBRP.NETWR, VBAK.NETWR
Profit Center	CEPCT.LTEXT
Payment Status	BSAD.* (closed), BSID.* (open)
Transaction	CDPOS.TCODE
Table	CDPOS.TABNAME
Table Descr	DD02T.DDTEXT
Field	CDPOS.FNAME
Field Descr	DDFTX.SCRTEXT_L
Old Value	CDPOS.VALUE_OLD
New Value	CDPOS.VALUE_NEW

TABLE 19: MAPPING ATTRIBUTES TO SAP

APPENDIX D EVENT LOG EXTRACTION SCRIPT

Vertrouwelijk

APPENDIX E CASE STUDY ANALYSIS OF THE MINING RESULTS

Vertrouwelijk