

# Data quality within process mining in the auditing context

SIMONE VAN SCHEEPSTAL  
ANR: 892315



Master Thesis Information Management

Author	S.J.C.M. (Simone) van Scheepstal ANR: 892315
Supervisors	
University supervisor	Dr. H (hans) Weigand Tilburg University
Co-reader university	Prof. dr. ir. H.A.M. (Hennie) Daniels Tilburg University
Company supervisor	W.M.J. (Wesley) Wiertz Msc RA EMITA BDO
Company	BDO Eindhoven Dr. Holtropaan 15 5652 XR Eindhoven
University	Tilburg University Warandelaan 2 5037 AB Tilburg

## **Preface**

I wrote this thesis in order to graduate from the master Information Management at Tilburg University. This research is written in assignment of BDO. The issue of data quality within process mining is a well-known hurdle, however it has not have great attention of the academic world in the financial auditing context. Therefore my interest was raised to develop more knowledge in this topic.

I would like to thank my university supervisor Hans Weigand and my company supervisor Wesley Wiertz for guiding me through this research by providing feedback and giving access to relevant sources needed to perform this research. Besides this, I would like to thank my colleagues at BDO for every helping hand.

Simone van Scheepstal

## **Abstract**

The power of process mining is dependent on the quality of the process mining data, however this is not sufficient yet. Handling data quality issues within process mining is not yet researched in the field of financial auditing, for this reason this research is performed.

The problem statement of this research is to find an answer of how to handle the data quality issues within process mining in the financial auditing context. In order to answer this problem statement three research questions need to be posed, the first is which quality issues are known to arise in the application of process mining. This question is answered using a literature review. Afterwards, the question is posed if these identified data quality issues influence the application of process mining in the financial auditing context. This question is answered using a holistic multi-case study with participants who perform process mining analysis in the context of financial auditing. As the processes in scope for the audit are subject to a certain extent of control in order to comply with regulatory and law requirements, are these not comparable with any other processes. The last question which is answered is how these identified issues which influence the application of process mining in the financial auditing context can be handled. This question is answered by creating a guideline with the use of literature, and this is verified by means of a personally administered questionnaire with participants who perform process mining analysis in the context of financial auditing.

The outcome of this research is that all identified data quality issues are influencing the application of process mining in the auditing context. These data quality issues can be categorized within two categories, these are; data quality issues as a result of contemporary company process deviations, and data quality issues as a result of event log deficiencies. The conclusion of this is that there is not yet enough academic research undertaken to adequately handle the all identified issues that influence the application of process mining. Therefore, more progress in academic research has to be undertaken. The proposed guideline for handling the identified issues: voluminous, irrelevant data, concept drift and incorrect data are found as relevant for handling the issue at hand and feasible to execute in the process mining procedure. This entails for the issue voluminous data conducting a process analysis in collaboration with the customer and scoping of the process mining analysis based on key research questions and the process analysis. For the issue concept drift this entails a statistical test in order to uncover any changes within the company process. As last, for incorrect data this means validating the process mining model by the customer.

# TABLE OF CONTENT

1. Introduction .....	1
1.1 Context.....	2
1.2 Relevance of this study .....	2
1.3 Research questions .....	2
1.4 General overview .....	2
2 Context .....	4
2.1 Introduction.....	4
2.2 Financial auditing.....	4
2.2.1 Definition .....	4
2.2.2 Processes in scope.....	5
2.3 Process mining .....	6
2.3.1 Definition .....	6
2.3.2 Process mining perspectives and methodologies .....	7
2.3.3 Process mining applied in the audit .....	9
2.3.4 Added value of process mining in the financial audit.....	10
2.3.5 Added value in the financial auditing process .....	12
2.4 Conclusion .....	14
3 Data quality within process mining .....	16
3.1 introduction.....	16
3.2 definition.....	16
3.3 Quality dimensions .....	16
3.4 Data quality within process mining.....	21
3.4.1 Identified data quality issues within process mining .....	22
3.5 Guideline.....	1
3.5.1 Voluminous data .....	1
3.5.2 Case heterogeneity .....	2
3.5.3 Granularity .....	3
3.5.4 Concept drift .....	3
3.5.5 Object centric data.....	4
3.5.6 Missing data.....	5
3.5.7 Incorrect data .....	5
3.5.8 Imprecise data .....	6
3.5.9 Irrelevant data .....	6
3.5.10 Guideline.....	7
3.6 Conclusion .....	9
4 Methodology.....	11

4.1	Introduction.....	11
4.2	Research design .....	11
4.2.1	Research questions .....	11
4.2.2	Unit of analysis .....	11
4.2.3	Process .....	11
4.3	Propositions .....	12
4.4	Reliability and validity .....	13
4.5	Data collection .....	14
4.5.1	Literature review .....	14
4.5.2	Case study Interviews .....	14
4.5.3	Validation.....	15
4.6	Data analysis .....	16
4.7	Conclusion .....	16
5	Results and analysis.....	17
5.1	Introduction.....	17
5.2	Multi-case study interviews .....	17
5.2.1	External IT auditors.....	17
5.2.2	External financial auditors.....	20
5.2.3	Internal auditors .....	23
5.3	Cross case analysis.....	27
5.3.1	Introduction.....	27
5.4	Propositions .....	30
5.5	Guideline validation .....	31
5.5.1	Results validation .....	31
5.5.2	Analysis validation.....	32
5.6	Conclusion .....	32
6	Discussion .....	33
6.1	Introduction.....	33
6.2	Discussion .....	33
6.3	Limitations .....	36
6.4	Recommandations .....	36
7	Conclusion.....	38
8	References .....	40
9	Appendixes.....	44
	Appendix A - Interview protocol - invitations.....	44
	Appendix B – Basic structure multi-case interviews .....	45
	Appendix C – Summurized results multi-case interviews .....	50
	Appendix D – Milti-case interview transcripts.....	57
	Appendix E - Questionnaire: Guideline validation (21 questions) .....	72

Appendix F – Results validation ..... 85



## List of tables

Table 1 Quality dimensions (Cai & Zhu, 2015). .....	18
Table 2 Overview of quality dimensions. ....	20
Table 3 Identified issues as a result of process characteristics. ....	24
Table 4 Overview of all quality issues as a result of event log deficiencies. ....	1
Table 5 Proposed guideline for the identified issues out of literature. ....	7
Table 6 Distribution of interviewees (multi-case interviews). ....	15
Table 7 Distribution of participants (personally administered questionnaires). ....	16
Table 8 Results multi-case study IT auditors. ....	17
Table 9 Results multi-case study financial auditors. ....	21
Table 10 Results multi-case study internal auditors. ....	24
Table 11 Results proposition 1. ....	31
Table 12 Results proposition 2. ....	31
Table 13 Item CVI of relevancy and feasibility per item. ....	32
Table 14 Results: Relevancy. ....	85
Table 15 Results: easy to administer. ....	85
Table 16 Results: easy to administer. ....	85

## List of figures

Figure 1 Process mining procedure (Bozkaya, Gabriels, & van der Werf, 2009).....	9
Figure 2 Replication approach of a multi-case study (Yin, 2013). .....	12
Figure 3 Measure of occurrence: Quality issues as a result of contemporary process characteristics: IT Auditors. ....	18
Figure 4 Measure of occurrence: Quality issues as a result of event log deficiencies: IT Auditors. ....	19
Figure 5 Measure of occurrence: Quality issues as a result of contemporary process characteristics: Financial auditors. ....	22
Figure 6 Measure of occurrence: Quality issues as a result of event log deficiencies: Financial auditors. ....	22
Figure 7 Measure of occurrence: Quality issues as a result of contemporary process characteristics: Internal auditors. ....	25
Figure 8 Measure of occurrence: Quality issues as a result of event log deficiencies: Internal auditors. ....	25
Figure 9 Relevancy of issue. ....	27
Figure 10 Measure of occurrence in case the issue is relevant according functional groups. ....	27

# 1. INTRODUCTION

The upcoming section will contain the introduction of this research. The introduction will commence with outlining the topic of this research and the context in which this research is positioned. The practical and academic relevance are discussed afterwards. Then, the research questions and which will lead this research are stated along with the general overview of this research.

Data science evolved into an important field of study for both the business and academic world as the future of this field promises a magnitude of sustainable and relevant opportunities which could have a high impact on both the scientific and societal world (Chen, Chiang & Storey, 2012). This gives the opportunity to make sense of the overwhelming plentitude of data in this era of digitalization which surpasses the capability of traditional approaches. One field of emerging research is process mining which is on one hand a closely related part of data science, which actually means the extraction of knowledge from data, and on the other hand to process modeling and analysis (Provost, Fawcett, 2013; Chen, Chiang & Storey, 2012; Van Der Aalst et al., 2011).

The Process mining manifesto (2011, p172) describes process mining as “extracting knowledge from event logs readily available in today’s information systems in order to discover, monitor and improve real processes (i.e., how processes are executed in reality)”. Due to more availability of recorded event logs within information systems, business are now able to extract and analyze knowledge of their enterprise resource planning (ERP) systems, in order to see how processes are undertaken in reality versus how they are designed to be undertaken (Jans, Alles, & Vasarhelyi, 2013) (Van Der Aalst, 2012). However the power of process mining is dependent on the quality of the process mining data, and this is not sufficient yet (Jans, Alles, & Vasarhelyi, Process Mining of Event Logs in Auditing: Opportunities and Challenges, 2010).

A study by Cai and Zhu (2015) shows that within big data analytics, poor quality of data and the know-how of solving this issue has become a major challenge for both the business and academic world, as this results in low utilization of the data, lost efficiency, and jeopardizing correct decision making. One thing which is clear upon studying several studies is that the impact of data quality issues is immense as the importance of data quality increases as a result of the digital transformation to both system makers and users (Fox, Levitin, & Redman, 1994) (Eckerson, 2002) (Cai & Zhu, 2015) however unfortunately this subject has received little attention from academia.

According to a survey which was administered among companies, nearly half of these companies has experienced damages as a result of insufficient data quality. These damages they experienced was sourced back to extra time necessary to resolve insufficient data quality and a loss of integrity in the system or application (Eckerson, 2002). Other literature also acknowledges these problems and

additionally states that these problems may lead to a decrease in data utilization and therefore also negatively influence the decision making processes within companies (Cai & Zhu, 2015)

## **1.1 CONTEXT**

The context of this research is the application of process mining within the field of financial auditing. The processes which are in scope for financial auditing are subject to a certain extent of control in order to comply with law and regulatory requirements, for this reason are these processes not comparable with any other processes.

## **1.2 RELEVANCE OF THIS STUDY**

The academic relevance of this study is to fill the gap of knowledge about data quality within process mining in the financial auditing context. Current literature about data quality improvement methods in this context are lacking probably because process mining concerns an emerging field of research (Cai, Zhu, 2015) (Hazen, Ezell, & Jones-Farmer 2014).

Studies do already exist of this subject in the field of for example of web analytics or in the context of for example healthcare, biology or for the big data field as a whole (Cai, Zhu, 2015). However not yet about data quality of process mining in the in the financial auditing context.

## **1.3 RESEARCH QUESTIONS**

For this reason, the question which is sought to be answered in this study is:

The following research questions are defined in order to answer the problem of this research:

1. Which data quality issues are known to arise in the application of process mining?

This question is answered using a literature review.

2. Do the arisen data quality issues influence the application of process mining in financial auditing?

This question is answered using a holistic multi-case study with participants who perform process mining analysis in the context of financial auditing.

3. How do you handle quality issues which have a relevant influence on the application of process mining in the auditing context?

This question is answered by creating a guideline with the use of literature, and this is verified by means of a personally administered questionnaire with participants who perform process mining analysis in the context of financial auditing.

## **1.4 GENERAL OVERVIEW**

The following chapter concerns the literature review where a detailed analysis can be found upon the articles and books analyzed in order to obtain a thorough understanding of the topics of interest.

Afterwards data quality issues within process mining are is scrutinized by means of a literature review. This chapter is followed by the methodology of this research. The methodology chapter addresses the motives for the sampling methods used and the data collection process. Afterwards, the results are discussed and analyzed. The results of the data collection are stated in this chapter, these are discussed later on in the discussion chapter. In addition limitations of this research and

recommendations for further research are described, which will be followed by the conclusion of this research (Sekaran & Bougie, 2010).

## 2 CONTEXT

### 2.1 INTRODUCTION

The following chapter will contain basic information about the context of this research. First financial auditing explained, along with the processes which are in scope in a financial audit. Secondly, process mining is discussed, and this chapter ends with explaining how process mining adds value in the financial auditing context.

### 2.2 FINANCIAL AUDITING

The upcoming section explains the context of this research namely, the financial audit. First of all, the financial audit is defined with its relevant stakeholders, afterwards it discusses which processes are in scope for the external financial audit.

#### 2.2.1 Definition

The following section will describe the main actors, namely the external and internal financial auditor for the audit on the financial statement of a company.

Company nowadays are obliged to formulate financial statements for their stakeholders, in order to inform them about their financial situation. These financial statements are subject to independent audits by financial auditors in order to mitigate the risk of misinformation (Werner, Gehrke, & Nüttgens, 2012). These financial audits can be executed by internal and external auditors.

The external financial auditor who provides an independent opinion about the financial statement of a company in accordance with predefined standards. For instance the Generally accepted accounting principles (GAAP) (focused on U.S. companies), or the International financial reporting standards (IFRS) (focused on international companies) or to another locally accepted regulatory form of accounting standards, principles and procedures during the execution of compiling their financial statement. These principles or standards set a common language for reporting (Investopedia, 2016). For example, the GAAP has to be used by U.S. companies who distributes its financial statement outside the company. The final goal of an audit is a report which indicates that the auditees' financial statements are free of material misstatement (Auditing and Assurance Services: An Applied Approach, 2012). Material misstatements are defined as, if present in the financial statement, misstatements which may reasonably be expected to influence the economic decisions of the users taken on the basis of the financial statement, misstatements can arise from an error or fraud.

Judgements about materiality are made in light of surrounding circumstances, and are affected by the size or nature of a misstatement, or a combination of both (Auditing and Assurance Services: An Applied Approach, 2012). In other words, is the goal of an external financial audit to answer the following question: Is the financial statement fairly stated? In order to come to this conclusion the auditor has to perform a process in order to collect sufficient evidence to come towards a conclusion about the financial statement. First, the auditor has to understand the circumstances that the company is operating in, including how they govern their operation. Afterwards, the auditor needs to know which risks the company is facing, and how they are managing these risks, especially the risks which

may influence the financial statement. This knowledge is needed to form an opinion about the financial statement of a company. This is done by obtaining an understanding of the information systems which support these processes, including related company processes (Gehrke & Mueller-Wickop, 2010).

Internal auditors are also obliged to national, as well as international regulatory and law standards. An important international institution is the Institute of internal auditors (IIA) (Gehrke & Mueller-Wickop, 2010). Internal auditing is also obliged to deliver independent and objective assurance about the company's operations, in order to improve risk management, control, and governance processes. The focus of this research is on the financial audit on the financial statements of a company, in this scope, are the requirements set for external financial auditors for the financial statements of a company in general the same for internal auditors (Gehrke & Mueller-Wickop, 2010).

The difference between an internal and an external financial auditor is that internal auditors execute an audit in assignment of the management of the company, and external auditors execute an audit in assignment of external stakeholders of a company (Deckers & Van Kollenburg, 2016).

### **2.2.2 Processes in scope**

The focus of this study of is on financial auditing. This entails obtaining a thorough understanding internal control of all company processes and procedures which lead up to financial reporting, and besides this, also includes obtaining an understanding of the relevant information systems of the company (Akkerman, Admiraal, Brekelmans, & Oost, 2006). The connection between the financial statement and the company's processes are that the financial statement is a (re)production of the processes which the company has in place. This means that in order for the auditor to get assurance in the financial statement of a company, the processes which lead up to financial reporting need to be trustworthy too. For this reason it is important to depict the processes which are in scope for the financial audit.

The processes which are in scope differs per company, however the most common processes are purchasing, human resources, sales, production and financial processes. These processes need to be in controlled by the company in order to produce trustworthy output which lead to financial reporting. The purchasing processes need to be in control such that the purchased goods and/or services are completely accountable for the audit, this also means that the purchase process is audited on points for attention such as segregation of duty, the distribution of authorization of persons who participate in the process, the manner of control in the process of changes in the purchase prices, and acceptance of goods and invoices (Deckers & Van Kollenburg, 2016).

Another example is the production process, in here the process is audited if the company invested all their sacrificed input value in their production process in order to retrieve an adequate output value, as with the purchase process, this also means that the production process is audited on points for attention such as segregation of duty, the existence of production norms and effort of personnel and systems, the manner of control in the process of production planning, registration, progress control

and quality controls (Deckers & Van Kollenburg, 2016). These are only two examples of the most common processes which are scoped within an audit. All these processes have in common that much attention is devoted to the actual execution of the processes, which must possess a certain level of control, and the control of employee who execute the process such as the segregation of duty and authorization. The control of these processes is of interest of both internal and external stakeholders, especially in the audit with the goal to assure that these processes are in compliance with regulatory and law requirements and do not possess misinformation.

## 2.3 PROCESS MINING

In the upcoming section the analytical technique; process mining will be discussed. This commences with explaining the definition of process mining along with the process mining procedure and afterwards the added value of process mining in the financial audit will be discussed.

### 2.3.1 Definition

The analytical technique process mining has been evolving since the 1990s, where it was first researched in software engineering context and afterwards in the workflow management context. The first researchers who introduced this called Agrawal, Gunopulos, and Leymann introduced this concept in 1998. Through the time, tools were designed to apply process mining, and retrieve value out of process mining in practice (Werner, Gehrke, & Nüttgens, 2012). Research provided methodologies, applications and algorithms which allowed a clear visual representation of the output data in the area of workflow management (Werner, Gehrke, & Nüttgens, 2012).

Van der Aalst (2011) defined process mining as a term subsuming all methods of distilling structured process descriptions from a set of real executions. This could be placed in any context, and for this reason this definition may perhaps not be specific enough for the context of this thesis. One other key researcher in this field is Jans, she performed substantial research of process mining in the auditing context. She defined process mining as “The basic idea of process mining is to extract knowledge from event logs recorded by an information system” where an event log is “a chronological record of computer system activities which are saved to a file on the system” (Jans, Alles, & Vasarhelyi, A field study on the use of process mining of event logs as an analytical procedure in auditing, 2014, p. 1752). Thus, an event log is a chronological record of an automated process activity which are saved to a file in the system or application. Event logs are also known as an audit trail, however this may be confusing as audit trail may indicate that it has something to do with auditing itself. As this is not the case, it may be confusing to use audit trail as definition in this thesis as this research is conducted in the auditing context. For this reason, event log as definition will be used instead. What however is missing from the latter definition, and is mentioned in the first, is the ‘process’ facet of process mining. Why is this important? Well, because the focus of process mining are the relevant processes within the organization, processes are in this thesis defined as a “defined set of business activities that represent the steps required to achieve a business objective” (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013, p. 2)



### 2.3.2 Process mining perspectives and methodologies

The application of process mining in the auditing context are broad and versatile. Process mining is not only used for detecting anomalies and exceptions, also for research on processes in general (Prima Kurniati, Kusuma, Agung, & Wisudiawan, 2015). There are three fundamental process mining perspectives: First of all the process or data perspective, in this perspective the question is answered how the process was executed. This perspective is used to compare the designed and observed behavior. Secondly, the organizational perspective which answers the question who executed the process. This is used in order to research for example the 'four eyes principle'. And as last, the case perspective which answers the question what happened in a specific transaction. This perspective is used to scrutinize a specific part of a business process. (Jans, Alles, & Vasarhelyi, 2013).

Within these perspective, several analysis are possible. In the process perspective; process discovery, conformance checking and performance analysis is possible. The idea of process discovery is the starting point for every process mining methodology. This methodology aims to model a process based on the logs retrieved from information systems (Van der Aalst W. , 2011). Thus, by applying process discovery, the actual execution of the process is 'discovered' and visualized using data from information systems.

The output of process a graphical output of a summary of the followed activities in a process. The granularity of the output can be changed if needed in order to scrutinize any discrepancies.

With the aid of process discovery the auditor is able to find out in an unbiased manner how the processes are executed (van der Aalst, van Hee, van der Werf, & Verdonk, 2016). This could mean that process mining could reveal processes which are not supposed to take place (Jans, Alles, & Vasarhelyi, Process Mining of Event Logs in Auditing: Opportunities and Challenges, 2010)

Conformance checking is used in order to compare "the behavior of a process model and the behavior recorded in an event log... to find commonalities and discrepancies" (Van der Aalst W. , 2011, p. 2).

The output of such analysis are conformance measures which communicate the overall conformance of the model and the log, and local diagnostics which highlights a certain part of the logs which disagrees with the model (Van der Aalst W. , 2011).

So a conformance check tries to find the answer if there are any deviations between the model of the process, and the execution of the process (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013). The auditors framework in this perspective are de jure models which describe the desired process, thus as designed by the client including the boundaries of the process and on the other hand, de facto models describe the processes as they are executed, thus the reality where potential violations occurred which are described in the de jure models (van der Aalst, van Hee, van der Werf, & Verdonk, 2016). The de facto models can be retrieved using the process discovery perspective. The metrics which are used to describe the conformance or in other words alignment between these two models are the fitness and the appropriateness measure, these are calculated when analyzing the conformance check (Jans, Alles, &

Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013).

Conformance check is done in order to see how the process is executed in the auditee's organization when they encounter for example an (unexpected) constraint (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013). Besides this, since there an a-priori model used, you are able to check controls, for example the four-eyes principle.

Another advantage is that you are able to check whether the process is executed as designed, thus checking for deviations, and in case there are deviations, how severe these are (e.g. the probability of a certain outcome). This manner of applying the process mining technique is highly relevant for auditing as it enables the auditor to detect violations and segregation of duties (van der Aalst, van Hee, van der Werf, & Verdonk, 2016).

Performance analysis focuses on the performance of business processes and is extensively used in other context than auditing. However this kind of analysis can be of value for the auditor, for example to track outliers in the process, with regard to performance (Jans, Alles, & Vasarhelyi, Process Mining of Event Logs in Auditing: Opportunities and Challenges, 2010). Perhaps these outliers are a result of process deviations, or employee in the client's organization which deliberately caused this to happen. According to literature allows the organizational perspective for social network analysis and role analysis.

A social network analysis gives insight to all interactions between employees involved in a certain process, this may serve as a basis for understanding the meaning of all transactions in a process, and how this is influenced by employees. This is made possible by the originator entry in an event log. Research states that this kind of analysis gives most insight when it is applied to a small subset or group of interest (Jans, Alles, & Vasarhelyi, 2014).

Social network analysis could possibly detect collusive fraud, this perspective analyzes the information contained in an event log of every user for each transaction. This makes it possible to see for example an unexpected pattern of invoicing, and authorization between the same set of employees (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013).

Another organizational perspective is role analysis, this types of analysis looks into the meta-data in combination with the originator which is present in the event logs. The goal of role analysis is to gain insight which employee had responsibility for a certain process step

(Jans, Alles, & Vasarhelyi, A field study on the use of process mining of event logs as an analytical procedure in auditing, 2014). With the aid of this analysis the financial auditor is able to know if for instance 'four eyes principle' is in place within the organization.

The last perspective discussed is the case perspective, according to literature decision mining and verification analysis is possible here. In here, the subject of analysis are the cases and thereby the stored attributes in both the process instances and audit trails. This type of analysis focuses on the

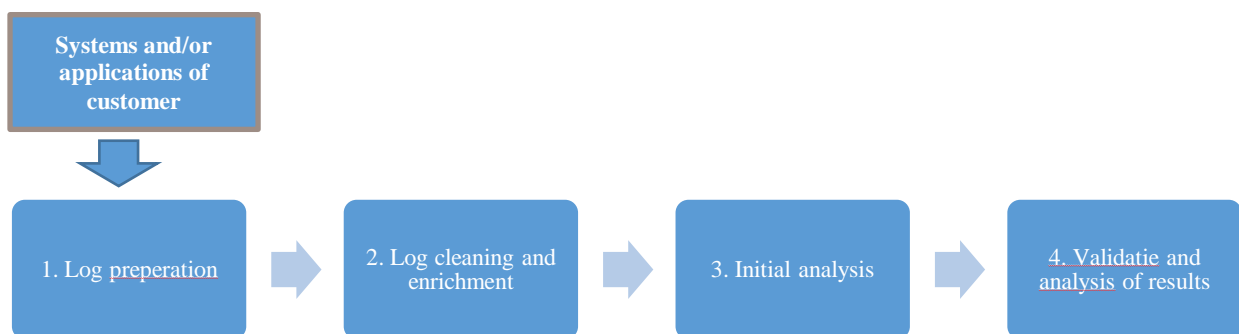
decision and verification tasks in a process. This analysis can be used in combination with other data analytics techniques. The output of this methodology could be an algorithm of how a process is executed. An auditor can within this perspective check if all internal controls were executed as designed by the client's organization.

Decision mining and verification analyses are not fully described yet in current literature in the auditing context (Jans, Alles, & Vasarhelyi, *Process Mining of Event Logs in Auditing: Opportunities and Challenges*, 2010).

### 2.3.3 Process mining applied in the audit

Most business store structure their data in systems and/or applications which support the company's processes. This data consist out of Meta data and input data. Input data are entries made by users of that system and metadata consist out of information recorded automatically by the system's database, such as the timestamp of when the activity of the data entry has taken place (Jans, Alles, & Vasarhelyi, *A field study on the use of process mining of event logs as an analytical procedure in auditing*, 2014). Metadata is thus contextual data which gives information about the entries made by users. The input and Meta data that the systems and/or applications that support a company's processes produces are formed into structured event logs.

Bozkaya, Gabriels, and van der Werf (2009) proposes a procedure for using process mining as a tool to research processes. This information will give insight into the procedure of the application of process mining within auditing.



**Figure 1 Process mining procedure (Bozkaya, Gabriels, & van der Werf, 2009)**

#### 1. Log preparation

In the first phase, named log preparation, data is extracted from the systems and/or application which support the processes which are in scope for the audit.

The extracted data can differ, this is dependent on the systems and/or application which support these processes, for example: Some data entails plain text and other data might entail complete databases

(Bozkaya, Gabriels, & van der Werf, 2009). For this reason the data should be (pre)processed in order to form proper event logs.

#### 1. Log cleaning and enrichment

Important here for the translation into standard formats is the selection of suitable data, as you have an overwhelming amount of data, only the event logs which are applicable for the relevant processes need to be selected (Van der Aalst W. , 2011) (Bozkaya, Gabriels, & van der Werf, 2009).

By the aid of filtering and querying the data minimum information is acquired which is needed within a standard format of an event log ( (Van der Aalst W. , 2011)

- The activity; which contains process steps that took place such as, pay invoice, receive order,
- The case ID; and
- The information stated in chronological order according to the events,

A case ID is a unique identifier for a process instance, which for example could be an invoice or an order. Also additional information is possible, these are called standard attributes such as timestamp (the time that the activity took place), resources (who performed the activity), transaction type and costs associated with the event.

Thus in this phase get the first overview of the process, statistics are gathered in this phase to research the event logs on for example completion.

#### 2. Initial analysis

In this phase the analysis which are intended to be performed, are executed these concern the process mining methodologies as mentioned previously, such as performance analysis, conformance checking and role analysis. Multiple methodologies and software tools are available for implementing process mining data are available today, the decision depends for instance on the kind of information system; process or object centric. These consist out of generic software tools such as open-source software, and specific tools based on one particular ERP system (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013).

#### 3. Validation and analysis of results

After the initial analysis, the you have information about how the process is executed in reality, however the auditor is not able to analyze these all by themselves as they are not able to decide whether behavior is wanted or unwanted, therefore validation of the customer is needed in order to form a valid advise with the results from the process mining analysis (Bozkaya, Gabriels, & van der Werf, 2009).

### **2.3.4 Added value of process mining in the financial audit**

In the following section the added value of process mining in the financial audit is discussed. First, the shortcomings of the traditional methods in the changed business environment. Secondly, how process mining adds value in financial auditing as it fulfills the shortcomings of the traditional methods of financial audit in the changed business environment of today.

**Shortcomings traditional method**

Through digitalization, most business processes are executed with the aid of information systems (Werner, Gehrke, & Nüttgens, 2012) (ICAEW, 2016). The data generated through these processed transactions, provides the basis for internal and external reporting. Contemporary audit methods do not take large voluminous data into account, for this reason, a discrepancy is formed between the highly digitalized business processes of companies on one side, and lacking audit methods to cope with this data on the other side.

Laws, regulations and standards differ between countries, however recent research found a confluence trend between the most significant international accounting frameworks. For instance, the International Federation of accountants (IFAC) is an international organization for the accountancy profession, they set international requirements such as the international standards on auditing (ISAs). They require the use of a risk based audit methodology which includes the business processes and information systems in place. Because of the digitalization, a lot of these business processes automated within ERP systems. This means that a majority of the internal controls are in fact ERP system application controls. What are exactly ERP system application controls? Within ERP systems there are control mechanisms in order to manage the business processes of a company. These control mechanisms are called ERP system application controls. These control mechanisms can be helpful within an audit. The external financial auditor can access the settings within an ERP system, which govern these control mechanisms instead of testing the transactions itself (Werner, Gehrke, & Nüttgens, 2012). Traditional audit methodologies are however not system based. This leads to a discrepancy between automated processes on the business side, and manually risk based evaluation methodologies on the other side. Clearly, the auditor is lacking modern audit methodologies, and for this reason lacking in efficient and effective audits.

Thus, the traditional approaches seems to lack the ability to incorporate the changes in the auditing environment, the consequence of this are several shortcomings which are present nowadays within traditional financial auditing (ICAEW, 2016). First of all, the traditional approach is cumbersome with the increase of business transactions in important business processes (ICAEW, 2016). Moreover, they do not take into account the changed business environment, which means encountering the digitalization. Furthermore, research within traditional auditing approaches is performed with a sample and not the whole population (Werner, Gehrke, & Nüttgens, 2012) (ICAEW, 2016). Thus, they do not give a full insight into the business processes, as they are mostly automated, and only the exceptions are executed by persons, moreover the persons they are interviewing in order to retrieve knowledge about the processes within traditional approaches for information are for this reason not reliable. Therefore, also not enough evidence to rely on. Research also states that financial external auditors nowadays lack knowledge to test these ERP system application controls. Therefore do not have sufficient knowledge at hand to make a correct judgement. This calls for a better financial

auditing approach, in the following section is discussed how process mining can overcome the shortcomings of traditional approaches.

#### **Added value process mining**

Process mining now, has not yet been fully understood in the auditing context, this is because it has been seen as a pure statistical concept, and not as an approach which can complement the methodology in the audit practice (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013); (de Kok, 2016).

The first problem to tackle is that the traditional approaches are cumbersome as business transactions increase. A field study has been performed in 2014 by Jans, Alles, & Vasarhelyi on the use of process mining of event logs as an analytical procedure in auditing and they found out that during the application of process mining they found a significant amount of relevant anomalies for auditing which were not found with traditional audit methods. Anomalies are transactions that may indicate financial accounting irregularities, breakdowns in internal controls, and/or fraud, which means that process mining gives auditors new information, which was not discoverable with traditional auditing methodologies. For this reason it is safe to say that process mining can complement traditional audit methodologies and overcome the large increase of business transactions.

Process mining does take into account the changed business environment, as it uses the entries and contextual data from information systems as a data source. Contextual data gives information about the time on which a certain activity happened, and for example by whom this activity was executed (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013) (Jans, Alles, & Vasarhelyi, 2014). This data will enable the auditor to consider the whole population instead of a sample in traditional audit approaches. However a large constraint here is the quality of data (Jans, Alles, & Vasarhelyi, 2013).

Process mining within auditing has a different focus than traditional audit methods. Process mining focuses on the business processes, rather than the value of transactions and its accumulations (Jans, Alles, & Vasarhelyi, 2014). So it gives the auditor full insight into the business processes of the client, including the processes which are fully automated. For this reason, this methodology is suited for the audit practice as it allows analytical procedures on the process level. Another advantage is that unlike with traditional approaches, where most of the data is entered by an employee of the auditee, process mining, specifically the contextual data, gives the auditor the opportunity to study data which is automatically and independently entered, which increases the trustworthiness of the data as such. In conclusion, the added value of process mining complements the audit methodology by filling up the shortcomings of the traditional approach.

#### **2.3.5 Added value in the financial auditing process**

As discussed in the previous section, the added value of process mining in the previous section is explained as complementing the audit methodology by filling up the shortcomings of the traditional

method, however there are more places in the audit process where process mining adds value according to literature.

Coney (2014) who performed a case study in the audit context on the integration of process mining in the audit states that the difference between the traditional audit approach and what process mining can offer is not only to only communicate the deviations but also advise the client in improving its internal control. Other objectives which can be accomplished by integrating process mining in the auditing practice is that it will increase the scope that auditors can handle, and provide instant business process insights. Moreover, it is possible to test the controls framework that the clients have in place with the use of process mining.

van der Aalst, van Hee, van der Werf, & Verdonk (2016) also researched the application of process mining in the auditing (both internal and external) context. They found that the auditor of tomorrow could use process mining even continuously in an organization as it does not interfere with the daily operation. For process mining analysis you could use, historic but also present-day information out of the systems without affecting the execution of these processes. In the audit process this could improve the phase where the auditor attains audit evidence.

Koopman and De Kok (2014) looked into the question if process mining is truly essential for the audit process. They found that process mining offers chances to this specialization to increase their added value towards the customer during the execution of the audit process. The application of process mining is supportive in every phase of external auditing. In the planning phase with understanding all processes and possible risks in scope, and in the execution phase with the test procedures. For example: With the aid of the process mining technique, the auditor is able to receive feedback information about not only acknowledged risks, but it also assist to recognize unknown risks in the clients processes as a result of the amount of data that process mining can handle, by comparing the design of the process with the actual execution of the process (Koopman & de Kok, 2014) (van der Heijden & Bajnath, 2015).

For example, an internal control can be the 'four-eyes' principle, this means that a certain activity, i.e. a decision, transaction must be approved by at least two people. This internal control principle can be tested with the aid of the event logs, these logs gives the auditor insight in the process steps, who executed which activity on which instant.

Another example could be tested the existence of a change management process, as designed by the company. With the use of process mining, you are able to see if all process steps are being executed in the same sequel as in the designed process, whether there are any exceptions, and the functionality of the internal controls.

A publication of ICAEW on data analytics for external auditors (2016) found that process mining could result in advantages both for the auditor, and for the client.

Advantages for the auditor are for example the scale and speed which are possible with data analytics. And for the client: also visualization, more insight into processes for example in the exceptions.



Moreover, data analytics enables auditors to improve substantive procedures and test of controls, in other words obtaining audit evidence, which therefore improves audit quality.

This all means that for the audit on the information systems, the demands and competencies will change. The audit becomes more data driven as mentioned in the previous sections, therefore more activities within the interim control phase will be handover from the financial auditor to the IT-auditor. As a result, the IT-auditor and the financial auditor have to be more cooperative. For this reason the financial auditor has to possess hands-on knowledge of several technical data analytics, and on the other side, possess strong communication competencies. Because they now have to take care that they execute these tests in a rigor and reliable manner, so the quality of these tests have to be sufficient, and secondly they have to be able to communicate the reliability and results of data-driven tests. This is nowadays not the case yet according to researchers van der Heijden and Bajnath (2015). This can be improved by standardizing these technical analytical methods, then the quality of the results are more reliable, and the analytical method is easier to apply for more clients.

## **2.4 CONCLUSION**

The first section of this research concerns explaining the context of this research. This entails the financial audit, and the application of process mining.

Company nowadays are obliged to formulate financial statements for their stakeholders, in order to inform them about their financial situation. These financial statements are subject to financial audits by financial auditors in order to mitigate the risk of misinformation. These financial audits can be executed by internal and external auditors, the main difference between these two is that the internal auditor audits the company in assignment of the management of the company and the external auditors audits the company in assignment of external stakeholders.

The financial audit entails obtaining a thorough understanding of the internal control of a company of all procedures and processes which lead up to financial reporting, this also includes obtaining an understanding of the relevant information systems. The execution of an external financial audit is in accordance with predefined principles, standards and procedures which depend on the local legal and regulatory requirements. Common processes which are in scope for the financial audit are purchasing, human resources, sales, production and financial processes.

All these processes have in common that much attention is devoted to the level of control in output of the process, and of the systems and employees who execute the process. The control of these processes is of interest of both internal and external stakeholders, especially in the audit with the goal to assure that these processes are in compliance with regulatory and law requirements and do not possess misinformation.

Process mining is defined as extracting information from event logs recorded by systems and/or applications which support the process in a company. An event log is a chronological record of a process activity which are saved to a file in the system or application. An event log consist out of data



which is entered by an employee of the company; input data, and contextual data which is automatically logged by the information system; metadata.

There are several perspectives and methodologies in process mining analysis, the process perspective looks into the execution of a process and the organizational perspective, into who performed a process. And as last, the case perspective which answers the question what happened in a specific transaction, within these perspectives several analysis are possible.

The process mining procedure has several phases. The first phase concerns extracting data from the systems and/or application which support the processes which are in scope for the audit, afterwards in the second phase, the event logs are translated into a standard format of an event log, thus in this phase you get the first overview of the process. Afterwards analysis are undertaken, such as process discovery and conformance checking. At last, the results are validated with the customer.

The traditional approaches seem to lack the ability to incorporate the changes happening in the auditing environment, the consequence of this is that the external financial auditors do not have sufficient knowledge to make a correct judgement about the fairness of the financial statement of a company.

Process mining allows an analysis on the whole population instead of a sample, this incorporates the change in business environment, of an increase in business transactions, and this is not covered in traditional auditing approaches. Furthermore, Process mining focuses on the business processes, rather than the value of transactions and its accumulations, which does not so it gives the auditor full insight into the manner of control business processes of the customer, the part which is executed by the system (automated steps of the process) and the part which is executed by employee of the customer.

Besides this incorporates process mining also contextual data besides entry data. This means that the financial auditor is now able to study data which independently entered by the customer which is researched, this increases the trustworthiness of the data as such.

The added value of process mining in the context of financial auditing process is that financial auditors are now able to have full insight into the deviations of a company's internal control, and besides this communicate this and advice the client company on improving it.

Another advantage of process mining is that the application does not interfere with the daily operation of the company which is researched. The auditing process as such also benefits from process mining; in the planning phase with understanding all processes and possible risks in scope, and in the execution phase with the test procedures. It especially assist to recognize unknown risks in the clients processes as a result of the amount of data that process mining can handle, here process mining comes into effect in the identify and assess risks phase.

This all also means that the competences and capabilities of the auditor will change in the future, as the financial audit becomes more data driven.

## **3 DATA QUALITY WITHIN PROCESS MINING**

### **3.1 INTRODUCTION**

In the upcoming section data quality within process mining is discussed. First, data quality is defined. Secondly quality dimensions are discussed and as last, identified data quality issues within process mining are listed along with suggestions made in literature of how to handle these identified issues.

### **3.2 DEFINITION**

In order to research the quality of data in process mining it is essential to define data quality in the first place, in order to acknowledge all facets or dimensions of data quality, in order to research this in a later stage.

Wand and Wang (1996) conducted a research on data quality dimensions in ontological foundations, they state that the notion of data quality is influenced by the use of data. This seems logical because the goal of one person which is using the data, might differ from another person, in other words, the interest could be different and therefore also the conditions where the data should adhere to. Therefore it is important to look from the perspective of the intended user when defining data quality. Another factor that influences the definition of quality is the context where the data is used, this could mean the business environment in this case, including its processes (Cai & Zhu, 2015). The perception that data quality is dependent on the user and environment of the data is also shared by other researchers such as The Total Data Quality Management group of MIT University led by Professor Richard Y. Wang has done in-depth research in the data quality area. They defined “data quality” as “fitness for use” (Wang & Strong, 1996, p. 6) where fitness is judged by the user of the data. And another example is the researcher Eckerson (2002, p11) who also defined data quality as “fitness or suitability of data to meet business requirements”. Thus looking at these definitions it is clear that the quality of data is foremost dependent on the user of the data and its environment or in another word, its context.

### **3.3 QUALITY DIMENSIONS**

Other researchers define quality data more rigor by defining its attributes, or dimensions (Eckerson, 2002). Quality dimensions are defined as a set of data quality attributes that represent a single aspect or construct of data quality.

Wand & Wang (1996) states that the quality dimensions are grounded in the ability of data in an information system within a company to represent the reality of that particular business environment without any insufficiencies. Therein they recognized four possible sources of deficiencies, which in a later stage are used to define the quality dimensions. The first possible source of deficiency is; incomplete representation, this occurs when the reality is not mapped exhaustively into the information system, which results in an information system which is not able to represent the reality as the data is incomplete. The second deficiency is ambiguous representation, this situation occurs when two different states of reality is mapped into one state of the information system, then the

information system is not able to assume which state represents the reality, thus the data in the information system is not reliable anymore.. The third source of deficiency is called: ‘Meaningless state’. It is not required of the information system to be mapped exhaustively to the reality, for this reason there might be states within the information system which are not mapped back to the reality; these states are called meaningless states, these meaningless states consist out of data which is not relevant for the reality to know. The previous deficiencies are a result of design deficiencies of the information system as such. The last source of deficiency is garbling which means that a state of reality is mapped back into the wrong state of the information system which could result in two situations: either the state of reality is mapped back to a meaningless state of the information system, the consequence of this situation is that the data is not able to be mapped back to reality or, the state of reality is mapped to an incorrect state of the information system, the consequence of this situation is that the state of the information system does not represent the reality anymore. This deficiency is a result of an operational deficiency, which is a typical human error.

The research on data quality dimensions in ontological foundations then is able to motivate five common data quality dimensions in the auditing context, namely; accuracy and precision, reliability, timeliness and currency, completeness and consistency out of these four intrinsic deficiencies (Wand & Wang, 1996).

A key researcher on data quality dimensions is Redman In one of his researches in 1990s data quality dimensions are developed in order to relate these dimensions to the processes in a company which introduce database processes (Huh, Keller, Redman, & Watkins, 1990). This research proposes completeness, consistency, accuracy and currency as the most important quality dimensions.

Completeness of data in this respect means that all relevant data in a particular set is present (Huh, Keller, Redman, & Watkins, 1990). Consistency of data is accomplished if it meets three conditions, these are; that the format of the data representation meets the format requirements and that two or more sets of data within a database do not conflict with one another, and as last, that the level of consistency in data sets correlate to the correctness of data. The research states that consistency arises with a standard edit approach in formatting the data sets.

Huh et. al (1990) defines that accuracy as a measure of correctness of the data. The correctness of the data largely depends on the application of data, and the attribute or state in question (Fox, Levitin, & Redman, 1994). This is often confused with precision in this respect, however, precision is focused on the reproducibility of data and currency as up-to-date data.

Another research on this subject by Fox, Levitin, & Redman (1994) states the dimensions discussed in the previous research are more extensive. In this research dimensions are grouped into similar into themes in order to cover every facet of data quality.

Accuracy, precision, and reliability (Fox, Levitin, & Redman, 1994) is the first theme of quality dimensions. In this theme the research broadened the perspective on the correctness of data, namely accuracy with precision, the reproducibility of data and reliability. Precision in this respect refers to

the level of detail or classification that the attribute or state possesses. Reliability in this research is defined as a measure, more specifically a probability of correctness of data.

The second theme concerns currentness, age, and timeliness (Fox, Levitin, & Redman, 1994). In this theme the research broadened the perspective on the currency of data with age and timeliness. Age refers to the time which is necessary to process and deliver the information, and the reporting interval of the system and timeliness assess the availability of data for decision making.

The third theme concerns completeness and duplication of data (Fox, Levitin, & Redman, 1994). In this theme the research broadened the perspective on the completeness of data. The notion of incompleteness has two predominant sources, the first is an empty state in the data as a result of an unknown value of an applicable attribute; thus a state that is empty in the information system, but is relevant for the user of the system, or a value of a non-applicable attribute; thus a state that is filled with a value in the information system, but is not relevant for the user of the system. The second source of incompleteness is a missing state as such, in a data set. Besides, missing and empty states in a data set, it could also contain states which are double, or irrelevant, and these are known as duplicate records.

The last theme concerns the consistency and integrity of data (Fox, Levitin, & Redman, 1994). In this theme the research broadened the perspective on the consistency of data with the related notion of integrity. Integrity in this study extends the theme with the security and concurrency of data.

Nowadays, data quality dimensions are still used in literature, for example in another study, executed by Cai & Zhu (2015) who researched the challenges in data quality and data quality assessment in the digitalization era and found similar dimensions of data quality which are commonly accepted and based on the current business needs. Besides dimensions this research also defines typical associated elements within quality dimensions.

**Table 1 Quality dimensions (Cai & Zhu, 2015).**

Quality dimension	Elements	Definition of elements
Availability	Accessibility	- Measure of difficulty to access data
	Timeliness	- Duration between data generation and utilization
Usability	Credibility	- Objective and subjective believability of the source of data
	Data definition/documentation	- Specification of data. This is brought into relation with usability of data by this research
	Metadata	- Metadata is brought into relation with better understanding of data by (end)users
Reliability	Accuracy	- Measure of correlation with reference value which is subject to the application of data

	Consistency	- Equivalency of data within databases
	Integrity	- Dependent on the application of data, overall measure data modification in an unauthorized and/or undetected manner
	Completeness	- Validity of the values of all components in a data set
Relevance	Fitness	<ol style="list-style-type: none"> <li>1. Measure of the amount of data which is accessed by the (end)user</li> <li>2. The level that the data meets the (end)user requirements</li> </ol>

Besides the stated elements above, the study also acknowledges another overall significant element of data quality in the digitalization era which is structure, this refers to the level of difficulty to transform semi-structured or unstructured data into a structured format, and readability which refers to the ability to communicate data.

Another current study about quality dimensions in the example in the research of Eckerson (2002) studied the impact of data quality on using data as a strategic resource for the company. The methodology of the research consist out of interviews with experts, surveys among business intelligence professionals and a literature review. He found in this study that the dimensions of data are the following:

1. Accuracy: Does the data accurately represent reality or a verifiable source?
2. Integrity: Is the structure of data and relationships among entities and attributes maintained consistently?
3. Consistency: Are data elements consistently defined and understood?
4. Completeness: Is all necessary data present?
5. Validity: Do data values fall within acceptable ranges defined by the business?
6. Timeliness: Is data available when needed?
7. Accessibility: Is the data easily accessible, understandable, and usable?

Rozinat et.al. (2007) and Van der Aalst et.al. (2011) researched the dimensions of quality in order to find an evaluation measure. They defined ‘fitness’, ‘precision’, ‘generalizability’, ‘ Structure, and simplicity’ as the dimensions which are relevant during process mining.

Fitness measures to which extend the extracted data model contains the same logs as in the source system(s). Thus, in the situation that all the logs which exists in the source system(s), are present in the extracted data model, then the model fits perfectly, for this reason can this quality dimension be categorized to concern the completeness of the data.

Precision and generalizability are two antonyms in this context. Precision addresses overly general models, and generalizability addresses models which are overly precise. Overly general models have a good fit, however the precision in the model is poor, this means that the model does not allow to

demonstrate additional traces which are possible theoretically, regardless of the fact if these traces actually occurred in the event log, thus in real life. Overly precise models, on the other hand, is the exact opposite, where the model demonstrates all additional traces which are possible theoretically, however this may not give an abstract view of the process which is needed to evaluate this in a workable manner. For this reason, is determining the correct level of precision and generalizability a decision which depends largely on the question to be answered. Thus the level of precision versus generalization must be in line with the end-user i.e. the person who needs the data in order to answer their research questions.

The last dimensions that known in literatere in process mining (Rozinat, de Medeiros, Günther, Weijters, & Aalst, 2007) (Van der Aalst, et al., 2011). Covers in their study is structure and simplicity. The structure of a process model concerns the modeling formalism. The quality dimension is measured in the perspective of the skills and capabilities that the person may or may not have in modeling language. A process model with more or less suitable structure for this reason presented by a modeling language which is understandable for the target audience of the model. Modeling language in this light is the vocabulary of the model, in other words the syntax (the notation of the language) and the semantics (its meaning) (Harel & Rumpe, 2004). Simplicity (Van der Aalst, et al., 2011) is a dimension that also incorporates the capability of human perception. A simpler model is better to understand by a human, than a complex one. Thus, the model is better understood by its stakeholders as the degree of simplicity increases. When comparing the quality dimensions with the aforementioned studies these dimensions overlap the ones found in ontological foundations. See table 2 for an overview of quality dimensions found in literature, along with a definition, applicable indicators and sources.

**Table 2 Overview of quality dimensions.**

Quality dimension	Definition of quality dimension	Indicator(s)	Sources
Accuracy and precision	The information system represents a real-world state as represented in the real world.	<ul style="list-style-type: none"> <li>- The level of detail that the data state possesses is in compliance with the level of data in the real world (reproducibility of data)</li> <li>- The correctness (accuracy) of data is in compliance with the need of the (end)user of the data</li> <li>- The level of precision (versus generalization) of the model is in line with the (end) user.</li> </ul>	(Wand & Wang, 1996; Eckerson, 2002; Huh, Keller, Redman, & Watkins, 1990; Rozinat, de Medeiros, Günther, Weijters, & Aalst, 2007; Fox, Levitin, & Redman, 1994; van Der Aalst, et al., 2011; Cai & Zhu, 2015)

Reliability	The data in the information system is trustworthy to display the correct information	<ul style="list-style-type: none"> <li>- The probability of the correctness of data is sufficient</li> <li>- Unambiguous representation</li> </ul>	(Wand & Wang, 1996; Fox, Levitin, & Redman, 1994; Cai & Zhu, 2015)
Timeliness and currency	The delay between a change of the real-world state and the resulting modification of the information system state	<ul style="list-style-type: none"> <li>- The data is accessible for the (end)users</li> <li>- The duration for processing and delivering the information to a user (age) is kept to a minimum</li> </ul>	(Wand & Wang, 1996; Eckerson, 2002; Huh, Keller, Redman, & Watkins, 1990; Fox, Levitin, & Redman, 1994; Cai & Zhu, 2015)
Completeness	Ability of information system to represent every meaningful state of the state of reality	<ul style="list-style-type: none"> <li>- The extracted data model contains the same logs as in the source information system(s)</li> </ul>	(Wand & Wang, 1996; Eckerson, 2002; Huh, Keller, Redman, & Watkins, 1990; Rozinat, de Medeiros, Günther, Weijters, & Aalst, 2007; Fox, Levitin, & Redman, 1994; van Der Aalst, et al., 2011; Cai & Zhu, 2015)
Consistency	Data values are expected to correlate in the same situation	<ul style="list-style-type: none"> <li>- The data is represented in compliance with format requirements</li> <li>- One or more data sets do not conflict with one another</li> <li>- There is correlation between consistency and accuracy of data</li> <li>- The data is integer, security and concurrency of data is in compliance with the need of the (end)user</li> </ul>	(Wand & Wang, 1996; Eckerson, 2002; Huh, Keller, Redman, & Watkins, 1990; Fox, Levitin, & Redman, 1994; Cai & Zhu, 2015)
<i>Structure</i>	Measure of the capability of human perception to understand use the model	<ul style="list-style-type: none"> <li>- The modeling formalism is in relation to the capabilities and skills of the (end) user.</li> </ul> <p>Related notion: specification of data definitions and documentation is brought into relation with usability</p> <p>Related notion: Metadata increases the contextual dimension of data, and therefore is brought into relation with the understanding of data for (end)user</p>	(Eckerson, 2002; Rozinat, de Medeiros, Günther, Weijters, & Aalst, 2007; van Der Aalst, et al., 2011; Cai & Zhu, 2015) Eckerson (2002)

### 3.4 DATA QUALITY WITHIN PROCESS MINING

The quality of process mining results depend heavily on the input, as the saying states: “garbage in, garbage out...”(Bose, Mans, & van der Aalst, Wanna Improve Process Mining Results?, 2013, p. 2).



Thus, one can say that the power of process mining is dependent on the event log (Jans, Alles, & Vasarhelyi, Process Mining of Event Logs in Auditing: Opportunities and Challenges, 2010). It turns out that, in practice, a simple event log is not readily available (Coney, 2016)

Bose et al. (2013) applied process mining in over 100 companies, and from these cases identified four categories of issues which concerned process characteristics and almost 30 additional classes of quality issues. These concern incomplete, noisy and imprecise event logs. Additionally, as processes in companies nowadays tend to be complex, and subject to an extensive range of deviations, event logs are increasingly fine-granular, heterogeneous and voluminous too.

Some contemporary process mining algorithms address these problems however Bose et al. (2013) stresses that more attention should be paid to the process mining data before applying these techniques.

A manifesto written by Van der Aalst et al. (2011) acknowledges the challenges in process mining regarding data quality, this is according to them for the reason that process mining still an emerging technology is. The process mining manifesto states that good benchmarks for quality criteria are to be developed as the quality of event logs need to adhere to a sufficient level (Van der Aalst, et al., 2011). These need to be trustworthy, and complete. Moreover safe in order to have adequate consistency, actually adhere to all quality dimensions defined in the previous section. This means that first of all, the data needs to be complete given a certain scope and have well defined semantics and secondly, the event logs need to be reliable, i.e. the events actually happened and the according attributes are correct and as last accessible for the end-user.

### **3.4.1 Identified data quality issues within process mining**

The issues which are identified process mining can be divided within two categories, these are process characteristics and the quality of an event log. The first category deals with the challenges which arise due to business processes and the underlying information system(s) deviations and the latter category deals with challenges which arise from the quality manifested in an event log.

#### **Quality issues as a result of process characteristics**

The first category concerns process characteristics, depending on how these metrics manifest in process mining data, one can face several challenges such as dealing with fine-granular events, case heterogeneity, voluminous data and concept drifts.

The first hurdle to deal with is voluminous data (Bose, Mans, & van der Aalst, 2013), this is a result of a wide range of contemporary information systems and applications within a company which produce immense amounts of data, and most of the time, on a low granularity level for every level within a system. The process mining techniques which are existing today are not (yet) able to manage this increasing amount of complex data, and also scoping becomes more difficult as the amount of data increases (Van der Aalst W. , 2011).

The process mining manifesto (Van der Aalst, et al., 2011) acknowledges challenges within the process mining process of extracting, merging and cleaning event data. The data may distributed



among several systems in a company (Van der Aalst, et al., 2011). The problem here is that there exist differences between the identifiers between the information systems. One system may identify a person by its name, while another system may identify a person by its employee number.

Another consequence of multiple systems is connecting these to each other, the challenge is the issue of correlation. The question here is to find of every event the corresponding case, as events are scattered among systems in an organization (Van der Aalst W. , 2011).

Secondary correlation occurs when an event is identified with activities (Van der Aalst W. , 2011) and these activities occur in parallel, then activities with the same definitions of different processes are hard to subdivide into processes as multiple options appear.

The second hurdle is case heterogeneity (Bose, Mans, & van der Aalst, 2013), this is as a result of the growing flexibility that a company needs to incorporate in their processes in order to meet customer needs. Besides customer needs, also the environment of a company may change over time due to changing legal or regulatory conditions or variation in supply and demand. For this reason, processes may possess a high amount of distinct scenarios, or in other words, traces which are possible. The hurdle here is that existing process mining techniques are not able to cope with case heterogeneity. A proposed solution given in literature is trace clustering, which separates the traces into subsets of homogenous cases. However, a disadvantage of this technique is that clustering these traces happens on a subjective basis.

The third hurdle is event granularity (Bose, Mans, & van der Aalst, 2013) (Van der Aalst, et al., 2011) (Van der Aalst W. , 2011). As mentioned in one of the previous paragraphs, as a result of a wide range of contemporary information systems and applications within a company, a large amount of low or mixed granularity level event data is produced. Low granularity data is most common in high-tech systems and in information systems in which event logs are linked to automated reports in supportive software. Process mining analysis with low level granularity event logs are difficult to comprehend for human perception and cognitive systems. The challenge here is to find the correct level of granularity for the (end) user of the analysis.

The dynamic processes of companies change over time, these are known as concept drifts (Bose, Mans, & van der Aalst, 2013), as the model changes over time in unforeseen ways, these changes are also present in the event logs accordingly. The manifest of such concept drift in process mining analysis is also known as second-order dynamics (Bose, van der Aalst, Žliobaitė, & Pechenizkiy, 2011). These changes are categorized in momentary and evolutionary changes. Evolutionary changes are process changes which are substantial different for example other activities are executed within the process, or other data may be used. Momentary changes are only changes in the process for a certain time period. These can be viewed as outliers in the mined process data. The Process mining manifesto (Van der Aalst, et al., 2011) also acknowledges this that processes or in other words events take place in a certain context within a company. Seasonal deviations in business processes may

influence event logs. Merging seasonal influences with process mining data could give better insight into the behavior.

Jans et al. (2014) who researched the event log building procedure states that structuring event data currently takes a lot of time and labor. The reason for this is that an information system is most of the time not based on processes, however more on relations where a huge range of documents which are somehow related are stored in different levels of detail. This becomes a challenge in merging the data which follows a certain process at one level of detail (Van der Aalst, et al., 2011). ERP systems of companies may be process or object centric. Object centric data does not relate to a certain process but to a certain object, which could be a product or container. Then the object centric data, which could consist of a certain product need to be combined with other data in order to analyze the whole process, for example the purchase to pay process. This is also acknowledged by another study on process mining applied on auditing from Van der Aalst et al. (2016), in here the researchers state that the integration of process mining is first and foremost dependent on fact if an ERP system is object or process oriented (van der Aalst, van Hee, van der Werf, & Verdonk, 2016).

Please see table 3 for an overview of all discussed identified issues as a result of process characteristics.

**Table 3 Identified issues as a result of process characteristics.**

Issue	Explanation	Consequences
Voluminous data	The data extracted is from an increasing wide range of contemporary, and/or legacy information systems and applications within a company are difficult to handle within process mining analysis	<ul style="list-style-type: none"> <li>- Existing process mining algorithms are not able to cope with case voluminous</li> <li>- Differences between the identifiers between different information systems</li> <li>- (Secondary) Correlation problems</li> </ul>
Case heterogeneity	Growing flexibility that a company needs to incorporate in their processes may possess a high amount of distinct scenarios which are difficult to handle within process mining analysis	<ul style="list-style-type: none"> <li>- Existing process mining algorithms are not able to cope with case heterogeneity</li> <li>- A high amount of scenarios/traces possible</li> </ul>
Granularity	The data extracted from contemporary, or legacy information systems and applications within a company possess a mixed level or low level granularity which are difficult to handle within process mining analysis	<ul style="list-style-type: none"> <li>- low level granularity event logs are difficult to comprehend for human perception and cognitive systems</li> <li>- Issues in finding the right level granularity for the (end)user of the analysis</li> </ul>

Concept drift	Dynamic processes of companies change over time which may influence the process mining model in an unforeseen manner	<ul style="list-style-type: none"> <li>- Outliers due to momentary changes</li> <li>- No insight in process due to deviations in business processes which may influence event logs unforeseen</li> </ul>
Object centric data	Object centric ERP systems does not relate to a certain process but to a certain object.	<ul style="list-style-type: none"> <li>- structuring and merging event data takes a lot of time and labor</li> </ul>

Contemporary business processes are an important factor in a process mining analysis as these are reconstructed with the aid of process mining (Werner, Gehrke, & Nüttgens, 2012) (ICAEW, 2016). So the data that the execution of business processes provide serve as a basis for the analysis.

In the previous paragraphs the experienced hurdles related to process characteristics within process mining are described, these were found as a result of contemporary business process characteristics.

In the following section more attention is given to the process, and process environment in which these problems manifest.

**Quality issues as a result of event log deficiencies**

Quality issues are firstly due to contemporary business process deviations, however if all these aforementioned challenges were to be solved, then there are still data quality issues possible within the event log, the following section concerns this second category. Four categories are defined here. These could for instance be a consequence of the snapshot taken for process mining, this means that when historic data is extracted from the information systems, then the data concerns a certain range as you take a particular timeslot of the applicable process. The problem here is that you are likely also to retrieve incomplete processes, which started or ended outside the boundaries of the timeslot. The consequence here is that the quality of an event log decreases (Van der Aalst W. , 2011).

**Missing data**

The first category concerns missing data (Bose, Mans, & van der Aalst, 2013) (Van der Aalst, et al., 2011), the issue here is that an event log misses mandatory information which is needed for process mining analysis. The source of this problem is the log process of the system(s). The issue of missing data within data mining does not stand alone and various methods are already developed in the past to eliminate missing values, however a general procedure within the field of process mining does not exist yet (Daniels & Feelders, 2000).

Missing data can take place in any kind of component of the event log. In the following paragraph every component is discussed in relation to the above category.

Missing cases occur when the case took place in reality but it has not been recorded in the log. The effect of missing cases is that not the full population of cases is analyzed, or the sample taken does not represent the population, therefore the results are not generalizable. One can reconcile these hurdles, however this takes considerable effort, for instance deriving the information needed from other sources or the use of interpolate timestamps (Van der Aalst, et al., 2011).

Besides missing cases, also missing case attributes can occur within process mining analysis. This can affect analysis which focus on the attributes of cases. The more attributes are missing in such analysis, the more unreliable the output becomes, as all cases with missing attributes are excluded from analysis.

Besides missing cases and its attributes, also another component of an event log such as the event itself could be missing. This means that the trace in the data which is extracted, does not contain all events which occurred in reality. These occur for example when the data extracted from the information systems has a certain timeslot, events which belong to traces within the timeslot can be missing, as they happened outside the time range of the slot. These incomplete traces might come up as an outlier i.e. noise, which can be solved by cleaning the data, furthermore there are no solutions to this problem yet.

Besides missing events, also missing event attributes can occur within process mining analysis. This occurs when values which correspond to the event are incomplete. This can affect the reliability of the analysis, as it affects results found in cases and events.

Missing relationships also arises as a quality issue. Missing relationships occur when the connection between events and cases are missing, as mentioned before, these could influence the generalizability of the overall analysis.

When one or more events do not contain timestamps, missing timestamps may form an issue to either the applicability of process mining analysis, or to the reliability of the results of such analysis (Van der Aalst W. , 2011). However, when the position of events within a case are assured to be correct (even without a timestamp), then this does not pose an issue. Missing timestamps can form an quality issue, in the case where the position of events within a case are not assured, then the position of events are difficult to place in an order which reflects real behavior. If the position of events are not known, process mining becomes difficult in forming a flow with process mining, and the results less valid as there is no guarantee that a particular order of events is a reproduction of reality.

Missing activity names of events can also occur, the effect of this issue is that during analyses it is not known which activities an event consist out of. This becomes a bigger problem when the case IDs are relying on the activity names. Such problems affect the validity and reliability of the analysis.

Missing resources i.e. the originator of an activity within event logs can severely affect process mining analysis with an organizational perspective, such as social network analysis and role analysis.

### ***Incorrect data***

The second category concerns incorrect data, the issue here is that the information available in the event log is logged incorrectly, for instance the timestamp is incorrect (Bose, Mans, & van der Aalst, 2013).

Incorrect data can take place in any kind of component of the event log. In the following paragraph every component is discussed in relation to the above category.

Incorrect cases occur when certain cases in a log are incorrect as they belong to another process, and not the process where they are assigned to. The source of this issue is found in incorrect logging of the information systems within a company which supports multiple business processes. These incorrect cases may demonstrate itself as noise i.e. outliers in an analysis.

Outliers are also referred to as noise, this occurs when the event logs show exceptional behavior. This problem has to be addressed when cleaning the data in the process mining process (Van der Aalst, et al., 2011).

Besides incorrect cases, also incorrect case attributes can occur within process mining analysis. This means that the values which correspond to the case its attributes are logged incorrectly. This can effect analysis which focus on cases in the event logs. There are two manners known to resolve this issue, the first is to exclude the cases which have incorrect values in their attributes, and the latter is to predict the correct value of the attribute (this is only possible when there are a statistical significant amount of correct cases available).

Besides incorrect cases and its attributes, also another component of an event log such as the event itself could be incorrect. This means that the event is logged in the system, while it has not taken place in real life. Also an event has attributes, the values which correspond with these attributes can also contain incorrect data. This affects mainly analysis which have a data perspective, such as discovery analysis.

Incorrect relationships arise as a quality issue when the connection between events and cases are logged incorrectly.

When some or all recorded timestamps are incorrect, which means that they do not log the correct time instance when behavior occurred in reality, a quality issue arises (Van der Aalst W. , 2011). These discrepancy in time can be because of clocks within systems which are not set correctly, for example, have a delay of five minutes. Moreover, when clocks within several information systems of a company are not synchronized with each other. The consequences of this incorrect timestamps can possibly be severe, as this could result in an analysis which is not reliable. Moreover, as cause and effect situations within processes are influenced as the positions of events are not ordered correctly.

Incorrect activity names can also occur, this occurs when activity names are logged incorrectly.

Incorrect resources i.e. originators of an activity occurs when these are logged incorrectly. These pose problems in analysis with an organizational perspective as these focus on resources/departments.

### ***Imprecise data***

The third category concerns imprecise data, the issue here is that the entries in the event logs are too common or rough. An example could be that the timestamp is not noted as an exact time instant but as the parts of the day. Then a low level granularity analysis is made unreliable (Bose, Mans, & van der Aalst, 2013).

Imprecise data can take place in several components of the event log. In the following paragraph every component is discussed in relation to the above category.

Imprecise relationships occur are a result of chosen definitions of cases, in other words, connections between events in these logs are not reliable anymore because the chosen definition of a case makes it difficult for the process mining algorithm. For example, when events receive exactly the same definition i.e. activity name it is not possible to correlate these with the original reference. Thus, imprecise activity names can influence relationships which are made between events, more precisely too coarse activity names could potentially offer problems during analysis of these event logs. The effect is this is that the results are less reliable and accurate.

Besides imprecise relationships and activity names, also imprecise case attributes can occur.

Imprecise case attributes occur when the values are too coarse or abstract. The effect during analysis focused on case attributes, is that the results are too imprecise or even not usable.

The quality of an event attribute can also decrease as a result of too coarse values. The effect of this is that the attributes are not properly usable for a profound analysis.

Too coarse components can also affect the timestamps. This can affect the reliability of the order of events (Van der Aalst W. , 2011). For example, if there are two events which have the date and no time indication, then it is for certain which event took place first, and which afterwards. There can also occur a mixed level of granularity within imprecise timestamps; one information system may log timestamps very precise with date and time, while others in the same company only log the date that the event took place.

As mentioned before, too coarse or abstract timestamps affect the reliability of the order of events, and therefore influences the precision of the position of events. This could result in relations between events which are partly or not existing in reality.

Quality issues of imprecise components of event logs could also affect the resource or originator of an event log. For example, when the resource only registers the department instead of the employee who is executing the activity. This could, as mentioned in the previous quality issues categories, influence the analysis with an organizational perspective, namely influence the precision such an analysis.

#### ***Irrelevant data***

The fourth and last category concerns irrelevant data, the issue here is that the data in the event log is irrelevant for the applicable analysis. The solution of this is to derive the relevant data by filtering or aggregation, however this is very time demanding and challenging (Bose, Mans, & van der Aalst, 2013)

There are two components within an event log which can be influenced through irrelevant data, these are cases and events. These components are discussed in the following paragraph.

Irrelevant cases occur when certain cases in an event are not relevant for a particular context of an analysis. Including cases which are not in scope of the analysis could possibly make the model more complex than necessary. Resolving the case of irrelevant cases is by removing or filtering these during analysis, however this is only possible when the cases contain enough information to make such judgement.

Irrelevant events occur when a set of event logs are not ready yet for analysis and still include irrelevant events for a certain scope analysis and therefore need to be filtered and/or aggregated for instance to find the correct level of granularity. This could be challenges as the inclusion and exclusion is based on the subjective perspective of the analyzer, and aggregation and filtering is only possible when the events contain enough information to make such judgement.

Please see table 4 for an overview of all discussed quality issues within an event log as a result of event log deficiencies.

**Table 4 Overview of all quality issues as a result of event log deficiencies.**

	Case	Case attribute	Event	Event attribute	Relationships	Position	Timestamps	Activity names	Resources
Missing data	Case which took place in reality is not recorded in log	Values which correspond to the case are incomplete	Trace in extracted data does not contain all events which occurred in reality.	Values which correspond to the event are incomplete	The connection between events and cases are missing	Uncertainty in ordering of events within cases	One or more events do not contain a time indication	One or more events do not contain an activity name	One or more events do not contain resource which executed the event
Incorrect data	Case is assigned to incorrect process	Values which correspond to the case are logged incorrect	Event which did not take place in reality is recorded in log	Values which correspond to the event are logged incorrect	Connection between events and cases are logged incorrectly	Incorrect ordering of events within cases	Incorrect time instance recorded when behavior occurred in reality	Incorrect activity names are recorded in log	Incorrect resources are recorded in log
Imprecise data		Recorded value is too coarse		Recorded value is too coarse	Uncertainty in placing connection between cases and events during analysis	Uncertainty in ordering of events within cases	Recorded value is too coarse or has mixed granularity	Recorded value is too coarse	Recorded value is too coarse or has mixed granularity
Irrelevant data	Certain cases in an event are irrelevant for a particular context of an analysis		Extracted logs include irrelevant events for a particular context of an analysis						



### 3.5 GUIDELINE

In the following section various literature is discussed which gives suggestions of how to handle the issues which are identified in the application of process mining. The structure is as following, first a small introduction of the identified issue will be given, and afterwards the suggestions in literature of handling this issue will be discussed.

#### 3.5.1 Voluminous data

The data extracted for process mining is from an increasing wide range of contemporary, and/or legacy information systems and applications within a company which are difficult to handle within process mining analysis, this leads to existing process mining algorithms are not able to process these voluminous data, differences between different identifiers between each system and correlation problems.

Suggestions made in literature to solve this issue of process mining algorithms which are not able to cope with voluminous data and its consequences are to progress research into process mining algorithms in order to solve challenges which are encountered using this tool (Van der Aalst & Weijters, 2004) (Van der Aalst, et al., 2011). Further investigation into the topic of improving process mining algorithms is outside the scope of this research.

Research suggests to handle the issue of too much data in a process mining analysis by scoping (Van der Aalst W. , 2011).

The goal of process mining is to seek information within operational processes which answer the research questions which are stated. The challenge within scoping in this potential issues is to retrieve such data from an immense variety and volume of data sources, which exists in information systems structures nowadays (Van der Aalst & Weijters, 2004) (Bose, Mans, & van der Aalst, Wanna Improve Process Mining Results?, 2013). Within the audit approach the auditor decides which financial processes are in scope from a risk perspective, information about the information systems which are supporting this the financial processes are however not or only marginal considered by the auditor (Werner, Gehrke, & Nüttgens, 2012). Also other researchers stressed the importance of scoping, Bose, Mans, and Van der Aalst (2013) suggested that the appropriate scope of an analysis should be based on the context of the analysis. For this reason, it can be assumed that both the key research questions, and the context of the analysis are important to acknowledge.

Jans, et al. (2012) argues that the first step in process mining a process analysis entails which lists all relevant information about the business process. Subject matter experts in the auditee organization are needed to fulfill this step of the process analysis.

Aside from the selection of processes which are in scope for the audit by the auditor, also the identifier (the process instance which is followed in the process mining model) needs to be identified. The choice of the identifier depends on the key questions which are posed by the auditor, in order to answer these key questions with the created process mining model.

Van der Aalst suggests that further into the procedure of process mining, in the cleaning and enrichment phase, correlation issues can be solved by adding heuristics of extra attributes to the event logs, this is only possible by analysts who possess enough domain knowledge of the systems and/or applications which support the process, and besides this enough context knowledge about the process execution (Van der Aalst W. , 2011).

Thus in conclusion, several researches suggests to handle the issue of too much data in a process mining analysis by scoping, scoping is done on the basis of the key research questions posed. Both context knowledge of the process in scope, and domain knowledge of the information systems is needed to scope properly. A process analysis previous to scoping can provide this knowledge. The auditor has to collaborate with the customer to retrieve this information. The chosen identifier after the process analysis, should be based on the key research questions posed. The process analysis can also assist in solving correlation problems in the cleaning and enrichment phase.

### **3.5.2 Case heterogeneity**

Growing flexibility that a company needs to incorporate in their processes may possess a high amount of distinct scenarios which are difficult to handle within process mining analysis. Consequences of this issue are that existing process mining algorithms are not able to cope with case heterogeneity and that a high amount of scenarios (traces) are possible which are difficult to comprehend for human perception.

Several suggestions are done to handle the issue of case heterogeneity. Van der Aalst (2011) studied this subject and suggested aggregation of process instances, this means grouping traces on the basis of process characteristics, such as a certain type of order. This requires a precise process model for reference, in other words knowing which process characteristics there are, and deciding on this knowledge how to aggregate these from which the event logs are retrieved and does not incorporate anomalies. Another manner to tackle this hurdle is without using a prescribed process model and only look at the event log data retrieved, and build the model based on the data flow, however this requires preprocessing of data by a domain expert or a dedicated event log generation within the system(s) and/or application(s) which support the process. However preprocessing of a domain expert, and a dedicated event log generation is not always feasible in the auditing context is not always feasible, as you have a range of customers with different processes and systems and/or applications (Lu, Fahland, & Van der Aalst, 2014). Therefore this potential issue calls for another approach. Diamantini, Genga, Potenza, and van der Aalst (2015) suggests a method which looks into each trace separately of the process model and graphs these in a manner capable for human understanding. Then, conformance checking is performed in order to replay the process mining model and discover anomalies, and by studying these get insight to the process flow and repair these detected events following a certain set of rules defined by applicable stakeholders.

In conclusion, several suggestions are done to handle the issue of case heterogeneity, such as grouping the event logs on the basis of process characteristics, however a precise prescribed model is needed in

order to know the process characteristics and this approach does not incorporate the recovery of anomalies. Another suggestion is made build the model based on the data flow, however this requires either preprocessing of a domain expert or a dedicated generation of event logs, which is not always feasible in the context of this research. Therefore another approach is needed, a method which firstly looks into each trace separately, and afterwards performs conformance analysis on these traces could be a feasible manner to handle this issue, as it supports the discovery of anomalies and does not require preprocessing of a domain expert or a dedicated event log generation.

### **3.5.3 Granularity**

The data extracted from contemporary, or legacy information systems and applications within a company possess a mixed level or low level granularity, in other words; level of detail, which are difficult to handle within process mining analysis. Consequences of this issue are low or mixed level granularity event logs are difficult to comprehend for human perception and cognitive systems and issues in finding the right level granularity for the (end) user of the analysis.

A separation can be made between the potential issue of mixed granularity and a too low or coarse level of granularity. Several suggestions are given in literature to solve low granularity issues such as grouping events on a higher level via semantics ontologies or correlated activities such as flattening the logs in information systems (Van der Aalst W. , 2011). However these solutions are most of the time cumbersome (some solutions are only partly automated) and/or subjective. Research suggests that better tools and methodologies are needed to address these issues (process mining manifesto). A mixed level of granularity means that between systems which support the processes which are in scope for the audit have different levels of granularity. Important here is that the correct level of granularity is achieved which is dependent on the (end) user of the analysis (see literature review about data quality dimensions).

Thus this depends on the subject of the key questions concern for example an activity, which has a coarse level, or attributes of an activity which requires a lower level of granularity.

In conclusion, several decisions have to be made after defining the desired granularity level on the basis of the key research questions of the process mining analysis, in the procedure of process mining. In case of a desired coarse level of granularity, the data which is on a low level of granularity can be aggregated in order to create a model which is understandable for human perception.

### **3.5.4 Concept drift**

Dynamic processes of companies change over time which may influence the process mining model in an unforeseen manner. Consequences of this issue are outliers due to momentary changes and no insight in process due to deviations in business processes which may influence event logs in an unforeseen manner.

There has not been many research on handling this issue known to the limits of this research. Gunther, Rinderle-Ma, Reichert, Van Der Aalst, and Recker (2008) suggest using process mining as a feedback loop in order to see how the process is changed in a business setting. This works as follows, in the

information system(s) and/or application(s) that support a process record event logs, about how the process is executed, this in turn gives the organization information on how the process is executed versus how it is designed. This gives the company information to if needed change their processes in order to accomplish full integration in how the process is designed, executed and analyzed. However this assumes that all necessary logs are recorded in a system, and this is most of the time not feasible yet. Other research on this subject is by Bose, van der Aalst, Žliobaitė, and Pechenizkiy (2011), suggests several approaches in order to detect and localize changes in a process. This research suggests to detect these concept drifts using two approaches, one is to use statistical tests in order to detect changes in the data set regarding how the process was executed by dividing a log in two subsequent event logs and see if there are changes in the traces, another method is to detect any changes using metrics, such as the relation type count, which looks at the changes in a set of relations between activities over a certain time period and see if there are changes in the traces.

In conclusion, the research which is known to the limits of this research suggests two approaches to detect any changes in the process mining model by using data metrics and statistical tests on the process mining data.

### **3.5.5 Object centric data**

In this situation, systems do not record event logs along a process but on an object. The consequence of this issue is that structuring and merging event data takes a lot of time and labor as the analyst has to merge the data of all objects such as invoices and orders into an assumed process flow.

There has not been many research on handling this issue known to the limits of this research however research thus state that the integration of process mining dependent on solving this issue, moreover as event logs most of the time are object centric (Van der Aalst, et al., 2011). Building an event log in this situation requires a lot of time and labor, as data about these objects need to be merged and structured in order to have a process flow as output. The information of these objects are merged manually into an assumed process, where there is the jeopardy of losing details, therefore the added value of process mining i.e. full insight in the execution of the processes, is to an extent lost. For example, segregation of duty analysis is still possible because the information of the objects, for example the approval of certain invoices, is still retrievable. However the completeness and trustworthiness are corroded for a process discovery analysis. For this reason, it is questionable if there is still added value in performing a process mining analysis for the audit. This depends again on the key questions asked by the auditor.

There are process mining algorithms available which incorporate this issue. Werner, Gehrke, & Nüttgens (2012) provided complementary knowledge by filling the gap between accounting, compliance and process mining. They provide an algorithm which mines the traces of financial transactions within ERP system. Thus, it can find all relevant documents of one financial transaction. This approach is suitable for object centric ERP systems within companies, in the contrary the

approach of Van der Aalst (2011) which uses an algorithm for process oriented systems, however as earlier mentioned, the research of process mining algorithms are outside the scope of this research. The process mining manifesto (2011) proposes to use create standard guidelines in process mining in order to improve performance of process mining between different vendors who provide process mining techniques. This may lead to little problems in converting the data, and no interpretation problems (Van der Aalst, et al., 2011). This logic may perhaps also apply of different vendors of information systems and/or applications. This could be an outcome for systems which possess non-customized core functionalities by employing for example standard extraction scripts.

In conclusion, there is an inherent constrain in the possible key research questions in the process mining analysis if there are systems and/or applications in scope which contain object centric data. A manner to handle this issue is to create standard guidelines for systems which possess non-customized core functionalities in order to mitigate problems due to converting the data into event logs and no interpretation problems.

### **3.5.6 Missing data**

Missing data is defined as missing mandatory data which is needed for performing the analysis.

When looking into literature, there is no technique known to the limits of this research in literature to recover missing data, as it is not possible to recover something that is not there in the first place.

Fayyad, Piatetsky-Shapiro and Smyth (1996) who studies an approach for retrieving knowledge from data mining techniques states concerning this subject that there should be a strategy present when using a data mining technique in order to deal with this issue, however does not give suggestions into what strategies might be possible.

The process mining manifesto (2011), states that it is often possible to retrieve this information, however with significant effort, however on the other hand the research also assumes that process mining data are never to be regarded as complete, moreover as processes are not only excluded to be applied in one organization but also crossing several organizations as processes are also integrated in a supply chain. In this situation, the data is more difficult to retrieve, as two organizations may not share the same approach to the approach, nor recording this in their IT infrastructure. It thereby states that one should always be aware of this fact when interpreting a process mining model.

In conclusion, research poses that the process mining model should never be regarding as complete, however gives a solution to handle this issue by retrieving any missing data.

### **3.5.7 Incorrect data**

Incorrect data is defined as that process mining data does not correlate with the execution of the process in reality.

When looking into literature, there is no technique or clear guideline known to the limits of this research in literature to solve the issue of incorrect data in the application of process mining.

The researchers Van der Aalst, et al., (2011) also acknowledges this problem and suggests to detect outliers and poses the question on how to handle these, however does not suggest any methods to

handle this issue. Van der Aalst & Weijters (2004) reviews process mining into context, and states the main issues. Herein noise is stated as both incorrect and incomplete data. Incorrect data is stated here as a result of process mining algorithms which do not map the information of recorded event logs in an correct manner, as a cause of for example case heterogeneity (see identified issues resulting from contemporary process characteristics). It suggests that only if domain experts are in place, these incorrect examples of process mining can be discovered, in other words validated.

In conclusion, there is more research needed to explore, handle and solve this issue of incorrect data. Literature published until now suggests that domain knowledge is needed to uncover situations where incorrect data occurs, by means of validating the model. Besides this, are outliers recognized as possible manifestations of incorrect data. Furthermore, that process mining issues as a result of contemporary process characteristics in this case, case heterogeneity should be solved in order to mitigate the risk of incorrect data.

### **3.5.8 Imprecise data**

Imprecise data is defined as data which is not at the correct detail level of the (end) user. For this reason the level of detail must comply with the key research questions asked in the process mining model.

There has not been many research on handling this issue known to the limits of this research in the process mining field. One suggestion for imprecise timestamps is to interpolate these manually (Van der Aalst, et al., 2011), interpolation of timestamps is a method to construct new data points, in this case a timestamp, manually to activities.

In conclusion, it is important that the level of detail complies with the key research questions asked in the process mining model, however more research in this section is needed in order to handle this issue properly in the application of process mining, as only a suggestion is given to handle imprecise timestamps.

### **3.5.9 Irrelevant data**

Irrelevant data is defined as a process mining model that consists out of data which is not needed to answer the research question.

There has not been many research on handling this issue known to the limits of this research in the process mining field.

Bose, Mans, and Van der Aalst (2013) suggests that properly scoping the process mining assignment mitigates the risk of irrelevant information within the data extraction process. The process of scoping within process mining is defined as selecting the activities and traces within an event log that are not interesting for analysis (Bezerra, Wainer, & van der Aalst, 2009) (Van der Aalst, et al., 2011), for this reason the scoping procedure, as mentioned before should be based one key research questions.

The researcher Van der Aalst (2011) addressed this challenge of deciding which information is relevant, or irrelevant and suggested that domain knowledge is needed about the processes which are scrutinized in order to collect and locate the required information needed and to scope this

accordingly. One manner to retrieve domain knowledge is, as earlier mentioned, by performing a process analysis in collaboration with the customer (Jans, Alles, & Vasarhelyi, Process mining of event logs in internal auditing: a case study, 2012) (Jans, Alles, & Vasarhelyi, A field study on the use of process mining of event logs as an analytical procedure in auditing, 2014) in order to contain the domain knowledge needed to handle this issue.

In conclusion, more research is needed to resolve this issue. Properly scoping the process mining assignment based on the key research questions could assist in handling this issue. Furthermore is domain knowledge needed to collect and locate information which is relevant in the process mining model, this could be retrieved by performing a process analysis in collaboration with the customer.

**3.5.10 Guideline**

The suggestions that are made in literature to handle the identified issues can be summarized in table 5. The suggestions made for the identified issues voluminous data and irrelevant data; granularity and imprecise; case heterogeneity and incorrect data overlap to a certain extent for this reason are the suggestions combined for the aforementioned identified issues.

**Table 5 Proposed guideline for the identified issues out of literature.**

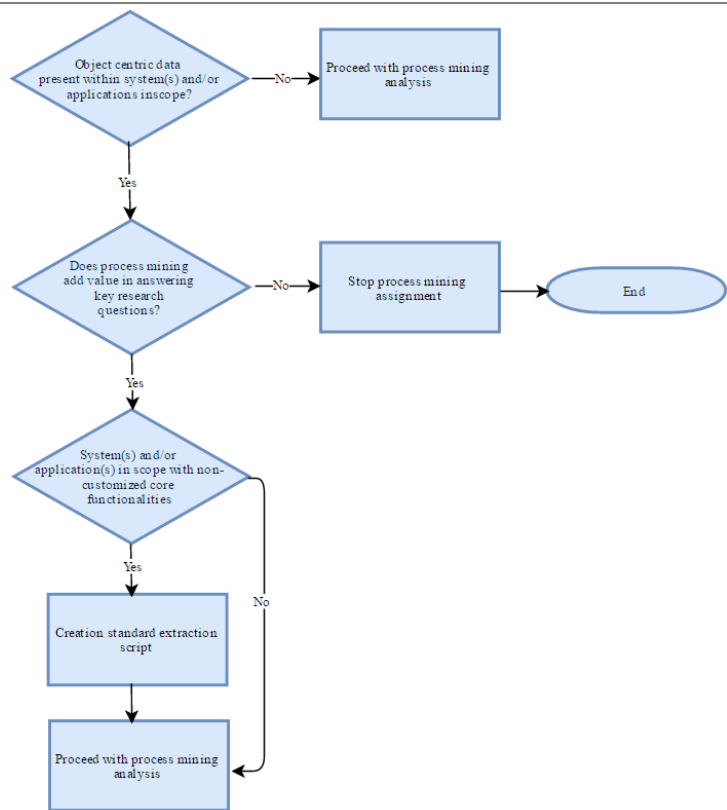
<b>Proposed guideline for the issues of voluminous data and irrelevant data:</b>	
<p>A process analysis will be performed in collaboration with the auditee in order to obtain context knowledge about the financial processes in scope and domain knowledge of the information systems. This process analysis entails which data is available in the recorded event logs.</p> <p>The process analysis can also assist in solving correlation problems further down in the cleaning and enrichment phase by providing information on the context of the processes which makes it easier for the analyst to add heuristics to the data.</p> <p>Scoping is accomplished according key research questions and the choice of identifier (the process instance which is followed in the process mining analysis) are made on the basis of the process analysis. This manner of scoping assists in collecting and locating only relevant data for the process mining model.</p>	<pre> graph TD     A[Performing process analysis in collaboration with the auditee] --&gt; B[Scoping process mining analysis on the basis of key research questions]     B --&gt; C[Defining unique identifier on the basis of the key questions and process analysis]             </pre>
<b>Proposed guideline for the issue of granularity and imprecise data:</b>	
<p>The desired level of granularity (level of detail) is defined based on the key research questions concerning the financial processes in scope, in order to find the right level of granularity for the end user of the analysis. In order to apply to correct level of granularity in the process mining model a plan of approach (PoA) is defined together with all relevant stakeholders, especially the end user.</p>	<pre> graph TD     A[Defining desired level of detail level based on key research questions] --&gt; B[Defining PoA in order to apply desired level of detail]             </pre>



**Proposed guideline for the issue of object centric data:**

The application of process mining with a workflow oriented data mining tool offers limited functionality when applied object centric data, for this reason, already in the scoping process one has to decide if process mining adds value in answering the key research questions.

When proceeding with the process mining analysis, the employment of standard extraction guidelines (scripts) are suggested in case of system(s) and/or application(s) in scope with non-customized core functionalities.

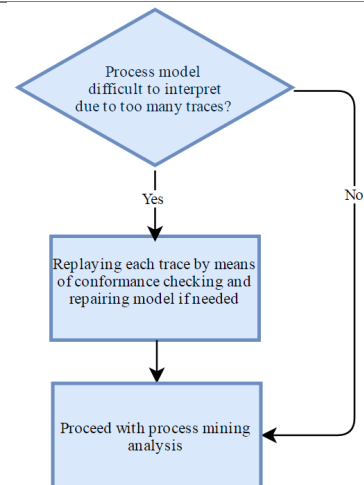


**Proposed guideline for the issue of case heterogeneity and incorrect data:**

In case the process model is difficult to interpret due to too many traces:

- Each trace is visualized in an process instance graph (a graph which visualizes each trace separately), in order to make it understandable for human perception
- Conformance checking (check between the process as described and how the process is executed) is performed in order to replay each trace and discover any anomalies

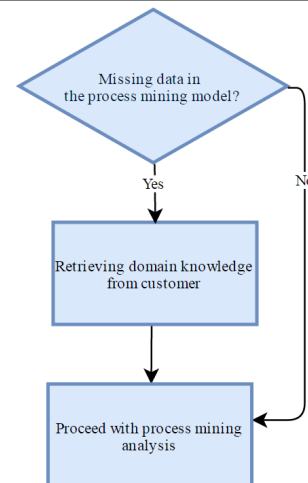
The traces with anomalies are studied in order to get insight in the process flow. Detected anomalies are repaired following a certain set of rules defined by applicable stakeholders.





**Proposed guideline for the issue of missing data:**

In case of missing data after data extraction, domain knowledge is often retrievable from the customer.

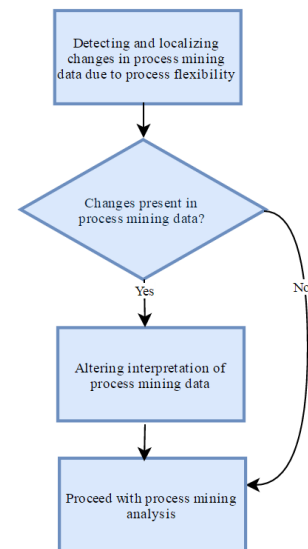


**Proposed guideline for the issue of concept drift:**

A check is performed on the process mining data in order to detect and localize any changes in the process.

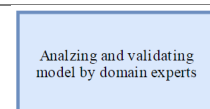
There are two approaches possible to detect and localize changes happened in the business process:

1. Use statistical tests in order to detect changes in the data set by dividing a log in two subsequent event logs and see if there are changes in the traces between these two data sets.
2. Detect any changes using metrics, such as the relation type count, which looks at the changes in relations between activities over a certain time period and from this knowledge finds any changes.



**Proposed guideline for the issue of incorrect data:**

Domain knowledge of the auditee is needed to uncover situations where incorrect data occurs, by means of validating the model.



Outliers in the process mining model are recognized as possible manifestations of incorrect data, these should be taken as warnings for incorrect data for the process mining analyst.

### 3.6 CONCLUSION

The definition of data quality is a construct is influenced by a several facets, it is clear that the quality of data is foremost dependent on the user of the data and its environment or in another word, its context. It is also possible to define quality with its dimensions, these are constructs which together define data quality.

Quality dimensions are grounded in the ability of data in an information system to represent the reality of that particular business environment without any insufficiencies. Therein they recognized four possible sources of deficiencies, which in a later stage are used to define the quality dimensions.

These deficiencies are; incomplete representation, ambiguous representation, meaningless state and garbling. These are then used to motivate intrinsic quality dimensions. See table 2 for an overview of quality dimensions found in literature which are found relevant in this thesis as they cooperate dimensions which are found in the data analytics, process mining and in the auditing context, along with a definition, applicable indicators and sources.

The impact of quality issues is huge as a result of the digitalization to both system makers and users, this leads to company damages, loss of integrity in the system or application, a decrease in data utilization, moreover within decision making.

Data quality issues within process mining are divided within two categories; contemporary process characteristics and event log deficiencies. The first category deals with the challenges which arise due to the contemporary business processes deviations and their underlying information system(s) and the latter category deals with challenges which arise from the quality manifested in an event log.

The consequences of these quality issues in an event log are far-reaching. The analysis may suffer a loss in generalizability and reliability and besides this, in the case when an analysis focuses on a component which has suffered extensively from these quality issues in an event log, even not possible anymore. Several solutions for these problems are suggested in literature, however these are in many situations, very time demanding, challenging and subjective.

The literature concerning handling the identified data quality issues in the application of process mining is very limited, for this reason more research is needed in order to make progress in the tools and methodologies of process mining. However some suggestions for handling these issues are made in literature and these are summarized in the following in table 5.

With the literature review in this chapter the research questions, of which data quality issues exist in the application of process mining is answered. Besides this are suggestions given in literature to handle these issues, these are combined a guideline. This answered the third research question partly, of how to handle data quality issues. However, it is still not known which data quality issues influence process mining in the context of financial auditing, and if this proposed guideline to handle these identified issues is working i.e. is relevant to solve these issues, and if this guideline is feasible to execute in the process mining procedure.

## 4 METHODOLOGY

### 4.1 INTRODUCTION

The following chapter concerns the methodology. The structure of the chapter is as following, first the research design is discussed, along with the process of this research. Afterwards the propositions which lead the multi-case interviews and the data collection methods are stated and afterwards the reliability and validity measures that are taken in this research are discussed.

### 4.2 RESEARCH DESIGN

The upcoming section concerns the research design, in this section a logical plan is presented in order to connect existing literature with research questions, and finally to its conclusion.

#### 4.2.1 Research questions

How can data quality issues which influence the application of process mining in the auditing context be handled?

The following sub-questions are defined in order to answer the research question of this research:

1. Which data quality issues are known to arise in the application of process mining?

This question is answered using a literature review.

2. Do the arisen data quality issues influence the application of process mining in financial auditing?

This question is answered using case study interviews with participants who perform process mining in the context of financial auditing.

3. How do you handle quality issues which have a relevant influence on the application of process mining in the auditing context?

This question is answered by creating a guideline with the use of literature, and this is verified by means of a personally administered questionnaire.

#### 4.2.2 Unit of analysis

The unit of analysis of this study is referred to as the system of action (Tellis, 1997) is the process mining procedure as defined by the organization of the interviewee during the case. Herein the difference lies between internal or external auditing, as in external financial auditing the procedure of process mining is executed by the IT auditor and the financial auditor. The financial auditor scopes the process mining analysis and gives advice on the basis of the process mining model, and the IT auditor prepares the actual process mining model.

#### 4.2.3 Process

The design of this process is in the following manner. The first phase is exploratory, in this phase a literature review is executed in order to identify quality issues within process mining, afterwards a holistic multi-case study within the auditing context. The goal of this phase is to have a full insight in quality issues within the application of process mining in the auditing context. Secondly, a guideline is designed which aggregates the findings of the exploratory phase in a guideline format. Afterwards these aggregated findings are validated with experts in the process mining field.

In the literature review extensive readings are performed in order to identify any variable which may explain the problem at hand, furthermore create a basic framework for further investigation.

A case study is chosen as a research method here as case studies are especially helpful in exploring processes in a manner which is not yet fully covered by academic studies, hence to answer how questions about contemporary events (Meyer, 2001).

Coney (2016) states that sharing experiences and herewith developing a standard approach could overcome the challenges within the application of process mining therefore, in order to listen in on these experiences a holistic multi-case approach is selected in this study in order to have both comparison and contrast between the cases as well as a deep and rich look into each case. The cases are sampled in a qualitative manner, which means that these are purposively selected instead of randomly in order to search for richness in information (Meyer, 2001) in order to tap the interviewees which are assumed to possess all relevant knowledge to answer the case study questions. A replication approach is used in a multi-case study (Yin, 2013), please see figure 2. The first step in this case study method is to develop theory, done with the aid of existing literature. Secondly, case selection takes place along with designing a data collection protocol. Every case study is considered as full study of the unit of analysis as such, the case report which follows indicates how a particular proposition was demonstrated. In the situation that a discovery occurs in one of the case studies the theory and propositions are altered if needed. Afterwards individual case reports are written, these are then compared in order to draw conclusions regarding replications and/or differences among results. These conclusions are validation by subject matter experts within the context of external financial auditing.

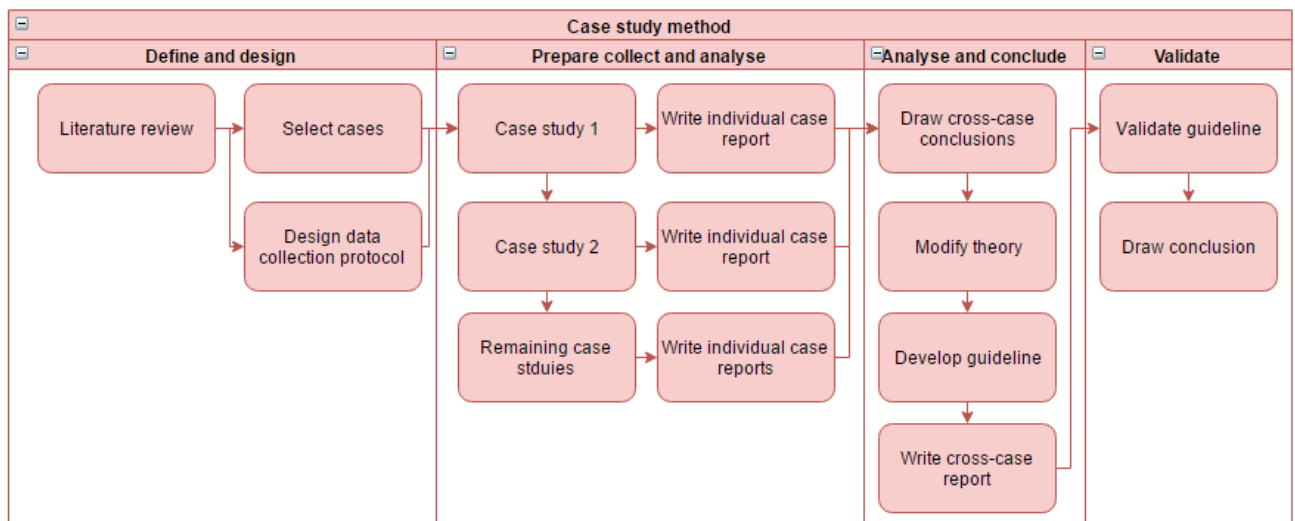


Figure 2 Replication approach of a multi-case study (Yin, 2013).

### 4.3 PROPOSITIONS

The mode of generalization within case studies is analytic, which means that existing theory is used as a framework with which to compare the empirical results of the study and besides this assist in defining the appropriate data design and collection methods (Yin, 2013). Practically, this means that

propositions are made in order to direct attention to the subjects which need to be examined in order to answer the research questions.

Data quality issues occur according to literature, both as a result of contemporary process characteristics and in the event log deficiencies itself. However there has not been research if these data quality issues have an impact to in the context of auditing. Issues, as a matter or situation which could be harmful, if these are not able to be handled sufficiently, in other words that it does not impact the action which is intended, in this case process mining. The goal of this research, amongst other things, is to verify if these quality issues have an impact in the auditing context, in other words, have an influence on the action of using process mining in the audit. For this reason the following propositions are defined:

1. Data quality issues as a result of contemporary process deviations have an influence on the application of process mining in the auditing context?
2. Data quality issues as a result of event log deficiencies have an influence on the application of process mining in the auditing context?

#### **4.4 RELIABILITY AND VALIDITY**

The upcoming section will contain criteria for judging the quality of the set research design. The approach of Yin (2009) is followed in establishing tactics in this research in order to ensure reliability and validity in this research.

First of all, multiple sources are used such as secondary sources i.e. literature and primary sources i.e. case study interviews in order to ensure validity. Besides this, is the report reviewed by key informants. Moreover is the replication approach research design set in the previous section known to stimulate validity of a research. Furthermore is a chain of evidence established in order to let the reader of this research get insight into the derivation from research sources and evidence towards the establishment of conclusions upon the research questions. Chain of evidence means that there exist clear links among the questions asked, the data collected and the conclusion of the research. For this reason, the report itself will include sufficient evidence of this research and also include a description of how this evidence was collected.

Reliability is assured in this research with the tactic of accomplishing a study database. A case study database entails a formal assembly of evidence from multiple sources of data.

Within a multi-case research it is essential according to a study of Yin (2013) that the protocol of the case study should include the following elements:

- An overview of the case study project objectives, and presentations about the topic under study i.e. the introduction and the literature review of this research
- Field procedures/reminders about procedures, credentials for access to data sources of those sources, therefore electronic database credentials are possessed by the researcher, and besides this access is granted within the network of the company where this research is executed (please see appendix A for the procedure)

- Case study questions/the questions that the investigator must keep in mind during data collection (please see appendix B for the basic structure of a case study interview)
- A guide for the case study report/the outline and format for the report are made in compliance with the report approaches of Yin (2009)

The inclusion of rules and procedures in a protocol of a case study increases the reliability (Tellis, 1997).

Besides this does the triangulation of evidence which is formed achieve reliability in the data collection process of the case studies and the evidence as such (Tellis, 1997) (Yin, 2013).

## **4.5 DATA COLLECTION**

### Introduction

The data collection is based on the research design and research questions of this research.

#### **4.5.1 Literature review**

For case studies, the review of theory is an essential part of the research. The goal of this literature review is to identify and assess present knowledge about the topic at hand, and develop a strong framework which determines what data to collect, and the strategies for analyzing this data in the later stages of this research (Yin, 2013).

Adding the secondary data sources in the literature review is accomplished in the following manner: the first step is to gather all sources for the review. These sources entail for instance: textbooks, journals, unpublished manuscripts and reports.

These are all found via electronic databases such as Google Scholar and the library of Tilburg University. Keywords used are process mining, data quality issues, data quality and external financial auditing. The data is evaluated on the basis of the relevance of the issues taken into account in the source, the importance or impact of a text book or article (measured by the number of citations) and the year of the publications, in here a mix of recent publications in the context of the problem statement and older publications on which current research is grounded is taken as a guiding principle. The documentation of literature is in compliance with American Psychological Association (APA) style in order to cite references and citations in a publicly accepted manner.

#### **4.5.2 Case study Interviews**

The interviewees are informed before the interview about the goal of the research and the structure of the interview (please see appendix A). The sampling of case interviews are done purposely in order to seek information richness across perspectives on process mining (Meyer, 2001). Besides this are the interviewees purposely selected on their knowledge on the topic at hand. As process mining is an emerging technology, a threshold of one year experience in the practical application of process mining is maintained. Please see table 6 for the distribution of interviewees.

**Table 6 Distribution of interviewees (multi-case interviews).**

ID	Function	Industry
C1	External IT Auditor	Financial services for small and medium sized enterprises (SMEs)
C2	External IT Auditor	Financial services for SMEs
C3	External IT Auditor	Financial services for SMEs
C4	External IT Auditor	Financial services for SMEs
C5	External financial auditor	Financial services for SMEs
C6	External financial auditor	Financial services for SMEs
C7	Internal auditor	Insurance
C8	Internal auditor	Services

Please see appendix B for the basic structure of the case study interview. For the validation of potential issues, a forced 2 point rating scale is used in order to avoid a neutral midpoint option and be able to analyze the results. In order to get insight into the issue if it is found relevant in the application of process mining, then the interviewees are asked how they handle the issue and the measure of occurrence.

These interview questions are validated by a subject matter expert within both context and subject of this research. The basic structure commences with descriptive questions about the interviewee, afterwards semi-closed questions are asked in order to collect data needed to conclude the propositions.

The audio of the interview are taped (with approval of participant), rather than noted, and the participants anonymity is preserved in order to decrease bias of the researcher and besides this to maintain a level of accuracy and richness of data.

### 4.5.3 Validation

For validation the researcher Lynn (1986) states that three experts are sufficient to make inferences based on the personally administered questionnaire. The validation of the guideline is fully executed by subject matter experts of the external financial auditing please see table 7. As process mining is an emerging technology, a threshold of one year experience in the practical application of process mining is maintained.

**Table 7 Distribution of participants (personally administered questionnaires).**

ID	Function	Industry
V1	External IT auditor	Financial services for SMEs
V2	External IT auditor	Financial services for SMEs
V3	External financial auditor	Financial services for SMEs

The constructs which are tapped are relevancy and feasibility (Refers to the ease and practicality of applying the guideline). Relevancy is measured using an item rating scale suggested by Davis (1992), feasibility is measured by 2 questions i.e. easy to administer and interpretability (Jordan & Turner, 2008). All questions contain rating scales with an even number of points are used in order to avoid a neutral midpoint option (Gilbert & Prion, 2016).

## 4.6 DATA ANALYSIS

In the end, the data collected should be linked to the propositions in order to see if these reflect with the propositions set for this study and find explanations for these findings. Data analysis within the case studies refers to studying, categorizing and tabulating the evidence in order to infer conclusions and provide inferences on the stated propositions (Yin, 2013).

The validation of the guideline is measured by the item content validity index. This index is used in order to quantify validity of an assessment by subject matter experts (Gilbert & Prion, 2016). This can be computed in two ways, namely by measuring the content validity on item level and on the overall scale level (Gilbert & Prion, 2016). In order to avoid an index which only expresses a proportion which is in agreement with the item analysis, and therefore possesses a standard error of the proportion, all experts must agree on each individual item level. In other words, the content validity context on item level should be 1.00 in order to be considered as a reasonable representation. For the content validity index on scale level, literature defines 0.80 or higher as an acceptable score for it to be considered as a reasonable representation (Polit & Beck, 2006).

## 4.7 CONCLUSION

This chapter concerned the methodology of this research. First the research design is discussed along with the process of this research, secondly the propositions are discussed, and afterwards the data collection methods. As last, the reliability and validity measures that are taken in this research were discussed.



## 5 RESULTS AND ANALYSIS

### 5.1 INTRODUCTION

The following chapter concerns the results and analysis of this research. The results of the multi-case interviews are stated in different views (from the perspective of the functional groups and cross case). Afterwards the propositions are discussed along with the validation of the guidelines made in order to handle the identified issues in literature.

### 5.2 MULTI-CASE STUDY INTERVIEWS

In the following section the results can be found of the multi-case study interviews. These are structured in the following professional groups; external IT auditors, external financial auditor and internal auditors. First the results of the multi-case interviews are discussed, afterwards a small analysis is stated. Please see appendix C for the summarized results of the case study interview of every professional group.

#### 5.2.1 External IT auditors

The following section will entail the results and analysis in the perspective of external IT auditors. Case C1-C4 are interviews which were undertaken with IT auditors, they are responsible preparing the process mining model for the external financial auditors.

#### Results

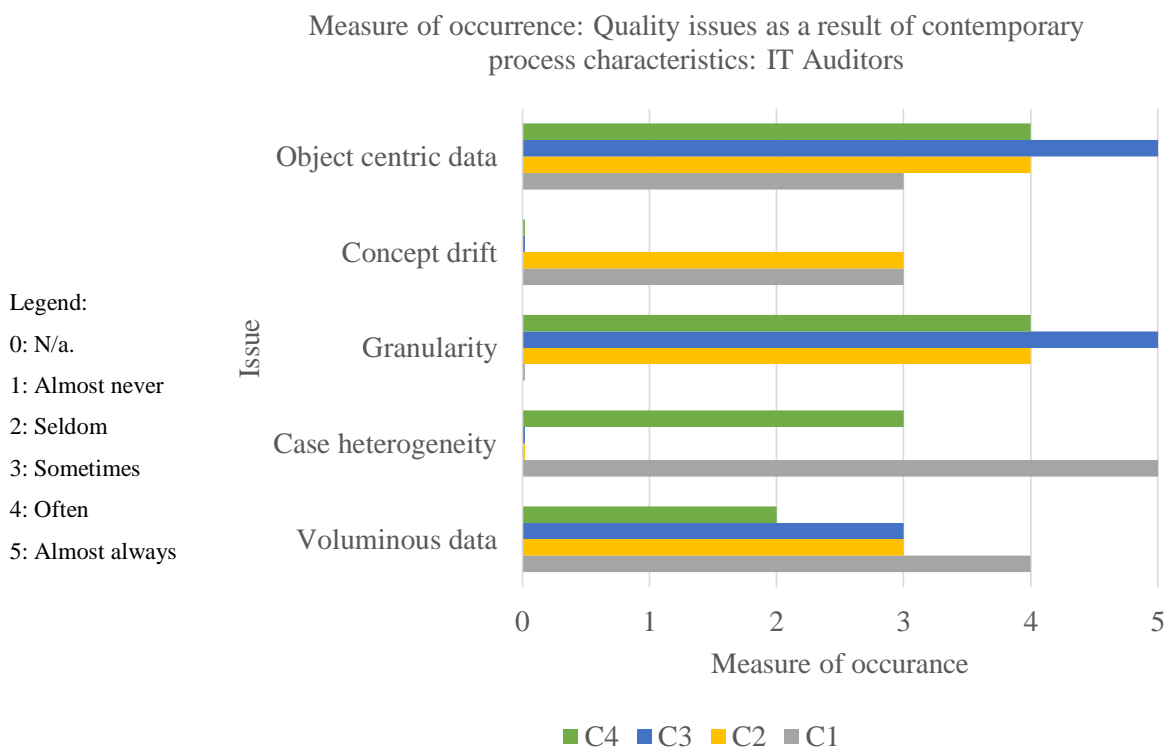
First the results of the external IT auditors are defined in table 8. First the issues which are found influencing the process mining application as a result of process characteristics and event log deficiencies are discussed.

**Table 8 Results multi-case study IT auditors.**

<b>Case ID</b>	<b>Issues as a result of process characteristics</b>	<b>Issues as a result of event log deficiencies</b>
<b>C1</b>	Almost all issues (except for object granularity) which are found in literature have been found as influencing the application of process mining in this case.	Almost all issues (except for object irrelevant) which are found in literature have been found as influencing the application of process mining in this case.
<b>C2</b>	Almost all issues (except for case heterogeneity) which are found in literature have been found as influencing the application of process mining in this case.	Almost all issues (except for irrelevant data) which are found in literature have been found as influencing the application of process mining in this case.
<b>C3</b>	Object centric data, granularity and voluminous data are found as influencing the application of process mining in this case. The issues of	Almost all issues (except for imprecise data) which are found in literature have been found as influencing the application of process mining in this case.

	concept drift, and case heterogeneity has been found to be irrelevant in the application of process mining.	
C4	Almost all issues (except for concept drift data) which are found in literature have been found as influencing the application of process mining in this case.	Imprecise is found as influencing the application of process mining in this case. The issues of missing data, incorrect data and irrelevant has been found to be irrelevant in the application of process mining.

Figure 3 shows the measure of occurrence of the issues as a result of process characteristics if it is found as an issue which influences the application of process mining in the auditing context, and figure 4 shows the measure of occurrence of the issues as a result of event log deficiencies if found as an issue which influences the application of process mining in the auditing context. The values of these tables are defined as, 1; almost never, 2; seldom, 3; sometimes, 4; often, and 5; almost always influencing the application of process mining. If the issue in a certain case has not been found as influencing the application of process mining in the auditing context, no value is given to the measure of occurrence, for this reason this remains; 0.



**Figure 3 Measure of occurrence: Quality issues as a result of contemporary process characteristics: IT Auditors.**

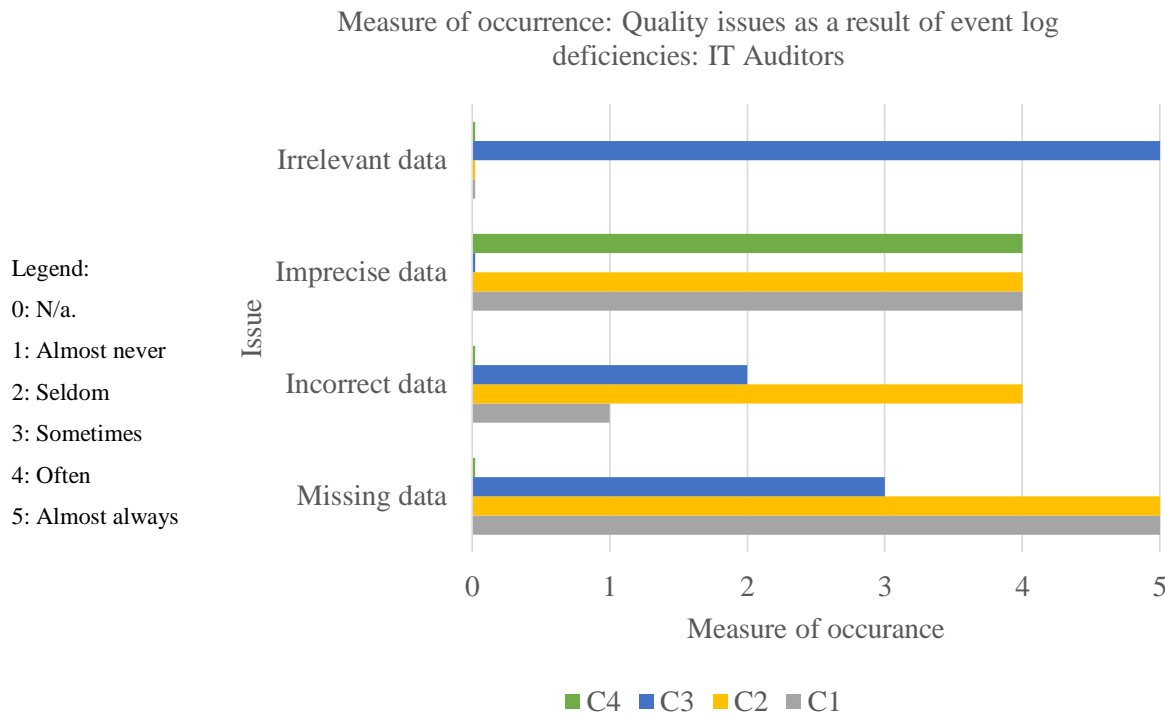


Figure 4 Measure of occurrence: Quality issues as a result of event log deficiencies: IT Auditors.

### Analysis

The first results which are discussed in the following section are the data quality issues as a result of process characteristics and afterwards the results of the data quality issues as a result of event log deficiencies are discussed within the professional group of external IT auditors.

Object centric data has been found as a relevant issue, which influences the application of process mining in the auditing context by all external IT auditors. According to interviewee C2 (see appendix D) is the reason for relevancy of this issue in this context as a result that most systems and/or applications in the processes which are in scope for the audit are supported by object centric oriented systems.

Concept drift has been found as a relevant issue which influences the application of process mining in the auditing context by two out of four IT auditors. A reason for irrelevancy of this issue in the auditing context is given by C3 (please see appendix D), according to this external IT auditor do these processes which are in scope for the audit not change often.

Granularity has been found as a relevant issue which influences the application of process mining in the auditing context by three out of four external IT auditors.

Case heterogeneity has been as a relevant issue which influences the application of process mining in the auditing context by two out of four external IT auditors. A reason for irrelevancy of this issue in the auditing context is given by C3 and C4 (please see appendix D), according to this IT auditor do

these processes which are in scope for the audit do not possess as much flexibility due to the reason that these need to be controlled in a sufficient manner by the company for the audit practice.

Voluminous data has been found as a relevant issue, which influences the application of process mining in the auditing context by all external IT auditors.

According to interviewee C2 (see appendix D) is the reason for relevancy of this issue in this context as a result that ideally you need a large timeslot of a whole year in the audit. Besides this, are according to interviewees C3 and C4 (see appendix D) the amount of systems and/or applications in place which support the process of the company an important factor, where voluminous data becomes a relevant issue if more than one system and/or application is applicable.

Irrelevant data has been found as a relevant issue which influences the application of process mining in the auditing context by one out of four external IT auditors. According to interviewee C1 (see appendix D) is the reason for irrelevancy of this issue because these kind of irrelevant data is easily removed from the process mining model, thus therefore it does not influence the application of process mining. Besides this, also according to interviewee C6 (see appendix D) the influence of this issue too little for it to influence the application of process mining in the auditing context.

Imprecise data has been found as a relevant issue which influences the application of process mining in the auditing context by three out of four external IT auditors.

Incorrect data has been found as a relevant issue which influences the application of process mining in the auditing context by three out of four external IT auditors. A reason for irrelevancy of this issue in the application of process mining in the auditing context is given by interviewee C4 (see appendix D), who states never to have experienced incorrect data yet to be an influence to the application of process mining. Another interviewee C2 (see appendix D) comments on this issue that this issue is only influencing the application of process mining as a result of a human error, as the interviewee never has not experienced incorrect data as result of system errors yet to the application of process mining.

Missing data has been found as a relevant issue which influences the application of process mining in the auditing context by three out of four external IT auditors. A reason for irrelevancy of this issue in the application of process mining in the auditing context is given by interviewee C4 (see appendix D), who states never to have experienced missing data yet.

### **5.2.2 External financial auditors**

The following section will entail the results and analysis in the perspective of external financial auditors.

Case C5 and C6 are interviews which were undertaken with external financial auditors, they are responsible for scoping the process mining assignment and advising on the basis of the process mining model.

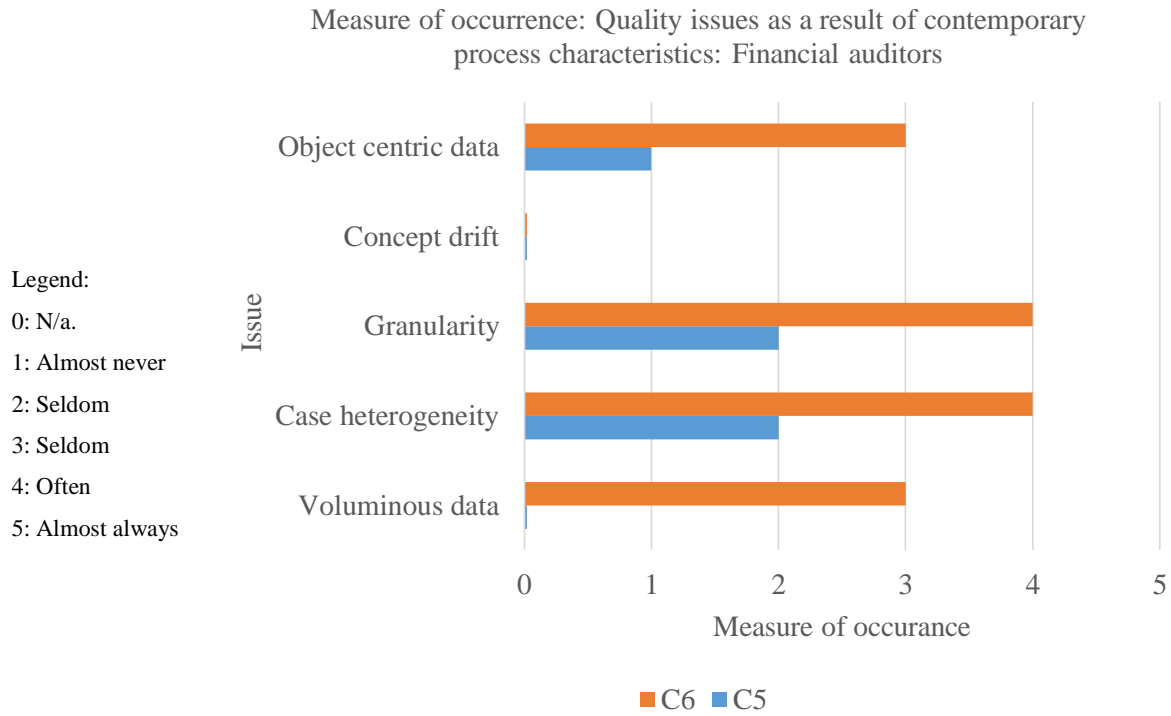
**Results**

First the results of the external financial auditors are defined in table 9. First the issues which are found influencing the process mining application as a result of process characteristics and event log deficiencies are discussed.

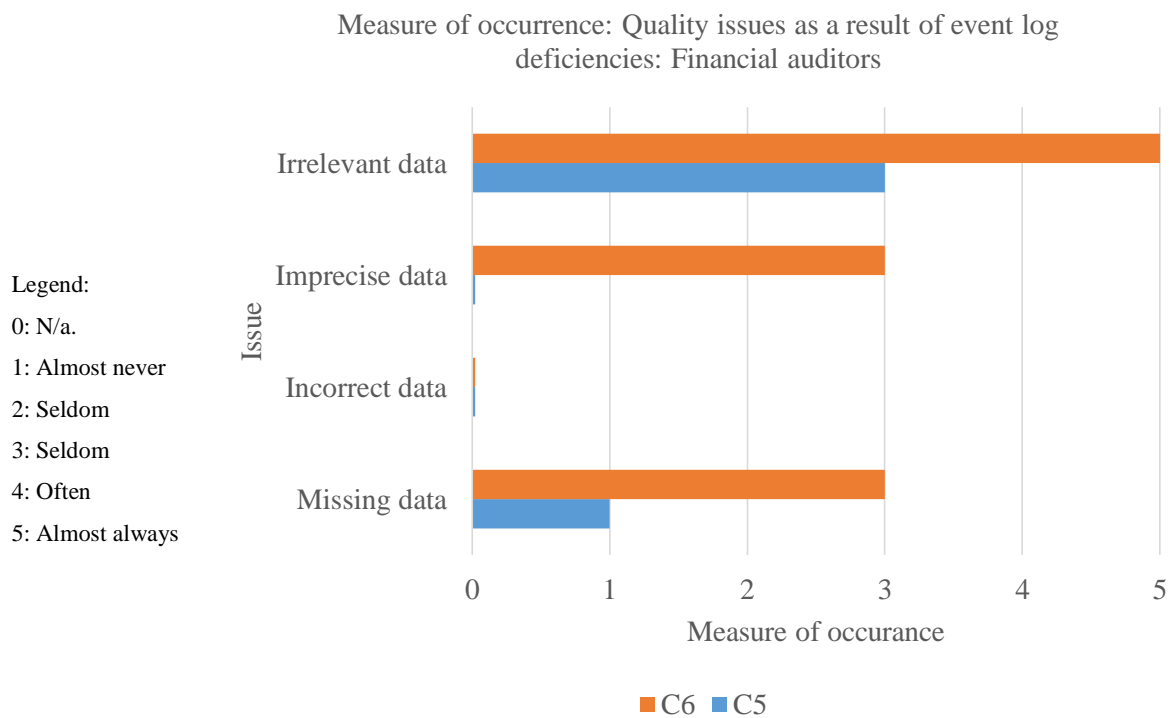
**Table 9 Results multi-case study financial auditors.**

<b>Case ID</b>	<b>Potential issues as a result of process characteristics</b>	<b>Potential issues as a result of event log deficiencies</b>
<b>C5</b>	Object centric data, granularity and case heterogeneity are found as influencing the application of process mining in this case. The issues of concept drift, and voluminous data has been found to be irrelevant in the application of process mining.	Missing data, irrelevant data are found as influencing the application of process mining in this case. The issues of imprecise data and incorrect data has been found to be irrelevant in the application of process mining.
<b>C6</b>	Almost all issues (except for object concept drift) which are found in literature have been found as influencing the application of process mining in this case.	Almost all issues (except for incorrect data) which are found in literature have been found as influencing the application of process mining in this case.

Figure 5 shows the measure of occurrence of the issues as a result of process characteristics if it is found as an issue which influences the application of process mining in the auditing context, and figure 6 shows the measure of occurrence of the issues as a result of event log deficiencies if found as an issue which influences the application of process mining in the auditing context. The values of these tables are defined in the same manner as in the previous section i.e. 1; almost never, 2; seldom, 3; sometimes, 4; often, and 5; almost always influencing the application of process mining. If the issue in a certain case has not been found as influencing the application of process mining in the auditing context, no value is given to the measure of occurrence, for this reason this remains; 0.



**Figure 5 Measure of occurrence: Quality issues as a result of contemporary process characteristics: Financial auditors.**



**Figure 6 Measure of occurrence: Quality issues as a result of event log deficiencies: Financial auditors.**

**Analysis**

The second group of results which are discussed in the following section are the data quality issues as a result of process characteristics and afterwards the results of the data quality issues as a result of event log deficiencies are discussed within the professional group of external financial auditors.

Object centric data has been found as a relevant issue, which influences the application of process mining in the auditing context by all external financial auditors. According to interviewee C6 (see appendix D) is the reason for relevancy that object centric data influences the trustworthiness of the process mining analysis.

Concept drift has been found as an irrelevant issue which influences the application of process mining in the auditing context by all external financial auditors. A reason for irrelevancy of this issue in the auditing context is given by C5 (please see appendix D) , according to this external financial auditor do these processes which are in scope for the audit not change often, however this could be a problem in other organizations. The other external financial auditor C6 (please see appendix D) states that, even when changes occurred in the process, as external financial auditor, you are aware of these changes before process mining analysis.

Granularity has been found as a relevant issue which influences the application of process mining in the auditing context by all external financial auditors.

Case heterogeneity has been found as a relevant issue which influences the application of process mining in the auditing context by all external financial auditors.

Voluminous data has been found as a relevant issue, which influences the application of process mining in the auditing context by one out of two external financial auditors. However, the external financial auditor C5 (see appendix D) who found this issue as irrelevant for influencing the application of process mining in the auditing context comments that instead of one year of data, one month of data is used to have an overview which is understandable for human understanding.

Irrelevant data has been found as a relevant issue which influences the application of process mining in the auditing context by all external financial auditors.

Imprecise data has been found as a relevant issue which influences the application of process mining in the auditing context by one out of two external financial auditors. Interviewee C5 (please see appendix D) that external IT auditors solve this issue, however the other external financial auditor C6 (please see appendix D) does not share the same opinion.

Incorrect data has been found as an irrelevant issue which therefore does not influences the application of process mining in the auditing context by all external financial auditors. Interviewee C5 (please see appendix D) that external IT auditors solve this issue.

Missing data has been found as a relevant issue which influences the application of process mining in the auditing context by all external financial auditors.

**5.2.3 Internal auditors**

The following section will entail the results and analysis in the perspective of internal auditors.

Case C7 and C8 are interviews which were undertaken with internal auditors, they are responsible for the whole process mining procedure.

**Results**

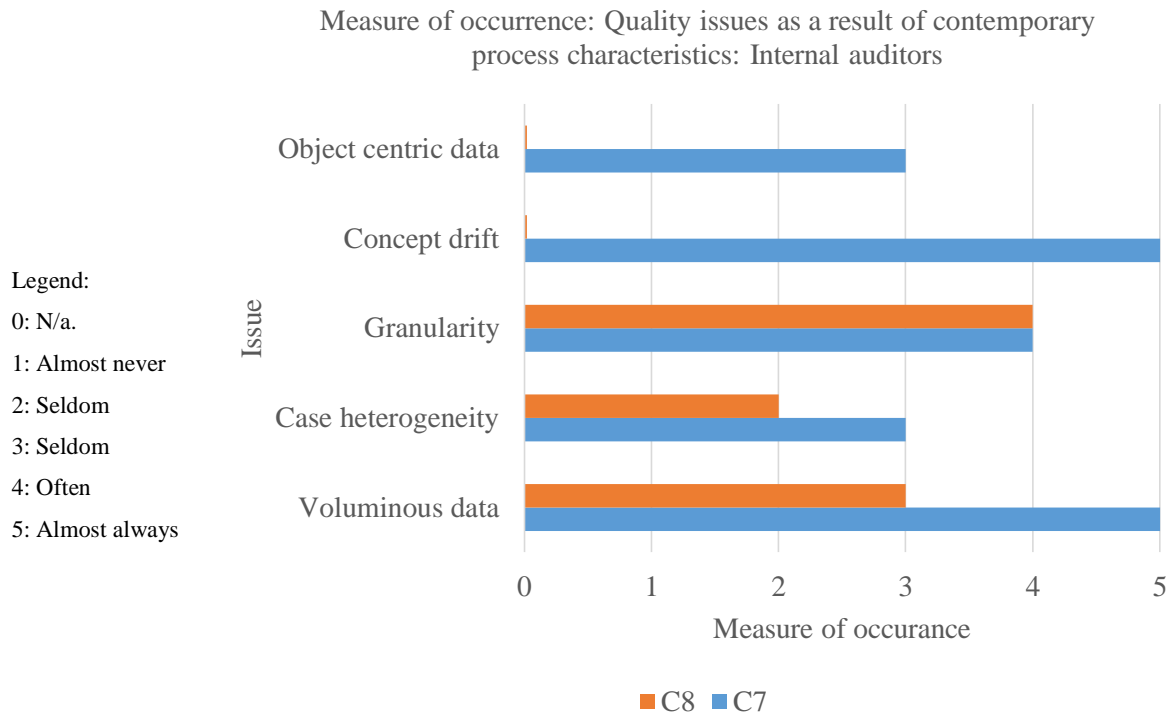
First the results of internal auditors are defined in table 10. First the issues which are found influencing the process mining application as a result of process characteristics and event log deficiencies are discussed.

**Table 10 Results multi-case study internal auditors.**

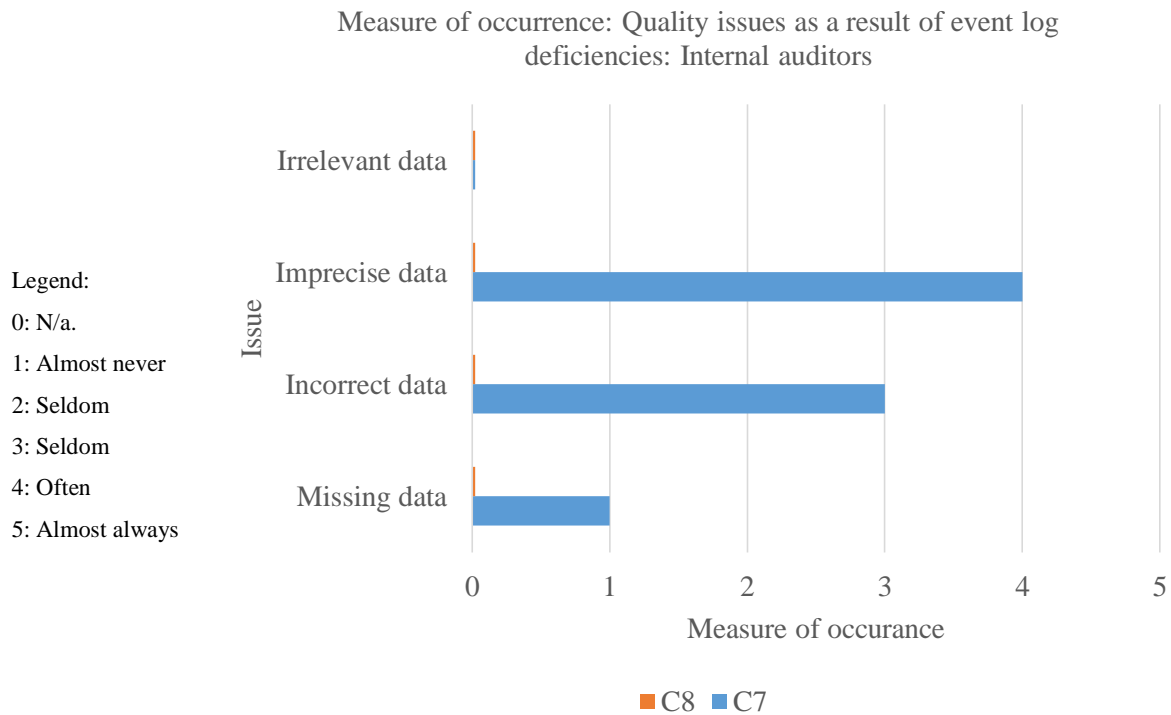
<b>Case ID</b>	<b>Potential issues as a result of process characteristics</b>	<b>Potential issues as a result of event log deficiencies</b>
<b>C7</b>	All issues which are found in literature have been found as influencing the application of process mining in this case.	Missing data, incorrect data and imprecise data are found as influencing the application of process mining in this case. The issue of irrelevant data has been found to be irrelevant in the application of process mining.
<b>C8</b>	Almost all issues (except for object centric data) which are found in literature have been found as influencing the application of process mining in this case.	None of the issues resulting from event logs are found as influencing the application of process mining in this case.

Figure 7 shows the measure of occurrence of the issues as a result of process characteristics if it is found as an issue which influences the application of process mining in the auditing context, and figure 8 shows the measure of occurrence of the issues as a result of event log deficiencies if found as an issue which influences the application of process mining in the auditing context. The values of these tables are defined in the same manner as in the previous section i.e. 1; almost never, 2; seldom, 3; sometimes, 4; often, and 5; almost always influencing the application of process mining. If the issue in a certain case has not been found as influencing the application of process mining in the auditing context, no value is given to the measure of occurrence, for this reason this remains; 0.





**Figure 7 Measure of occurrence: Quality issues as a result of contemporary process characteristics: Internal auditors.**



**Figure 8 Measure of occurrence: Quality issues as a result of event log deficiencies: Internal auditors.**

**Analysis**

The last group of results which are discussed in the following section are the data quality issues as a result of process characteristics and afterwards the results of the data quality issues as a result of event log deficiencies are discussed within the professional group of internal auditors.

Object centric data has been found as a relevant issue, which influences the application of process mining in the auditing context one out of two internal auditors. According to interviewee C8 (see appendix D) is the reason for irrelevancy that their systems are workflow oriented.

Concept drift has been found as an relevant issue which influences the application of process mining in the auditing context by one out of two internal auditors, according internal auditor C7 (see appendix D) influences this issue the application of process mining only in case of technical changes in the system itself.

Granularity has been found as a relevant issue which influences the application of process mining in the auditing context by all internal auditors, also is case heterogeneity been found as a relevant issue which influences the application of process mining in the auditing context by all internal auditors.

Voluminous data has been found as a relevant issue, which influences the application of process mining in the auditing context by all internal auditors.

Irrelevant data has been found as an irrelevant issue which does not influence the application of process mining in the auditing context by all internal auditors.

Imprecise data has been found as a relevant issue which influences the application of process mining in the auditing context by one out of two internal auditors.

Incorrect data has been found as a relevant issue which influences the application of process mining in the auditing context by one out of two internal auditors.

Missing data has been found as a relevant issue which influences the application of process mining in the auditing context by one out of two internal auditors. According to interviewee C7 (see appendix D) is process mining not applied in case of missing data. Interviewee C8 (see appendix D), states that the issue is irrelevant in the application of process mining as their systems are self-regulating, this means that the issue is solved within a certain time duration and therefor does not influence the application of process mining in their organization.

### 5.3 CROSS CASE ANALYSIS

#### 5.3.1 Introduction

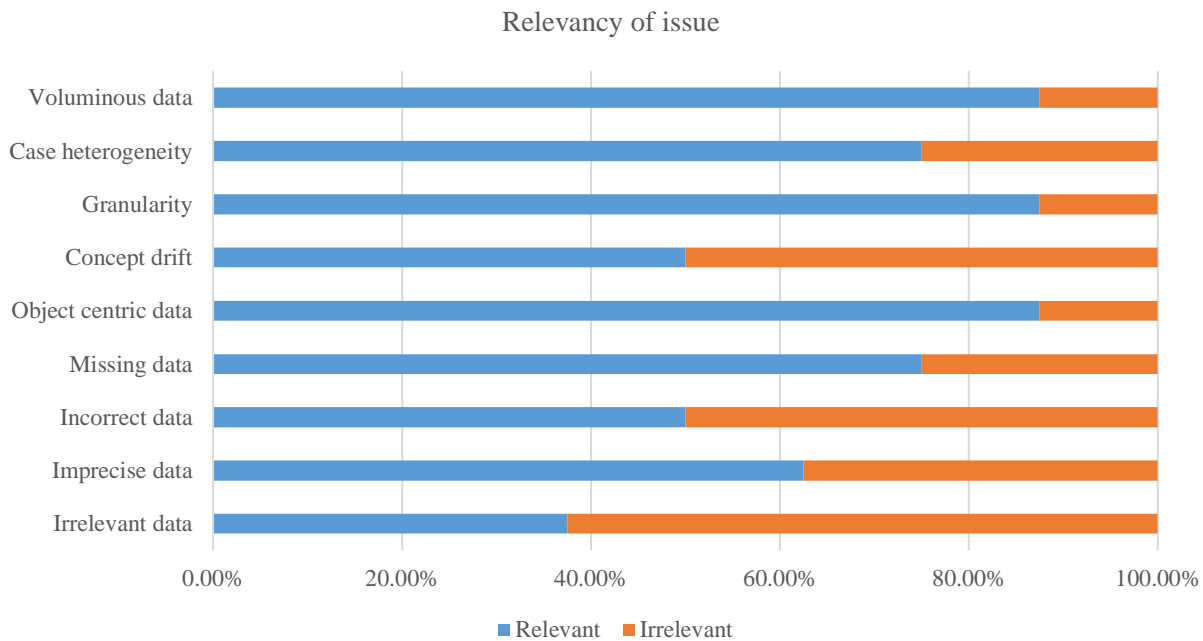


Figure 9 Relevancy of issue.

Figure 9 shows which part (in percentage) of the cases found the identified issue to be relevant or irrelevant as influencing the application of process mining in the auditing context.

Figure 10 shows the average measure of occurrence between different professions if the identified issue is found as relevant in influencing the application of process mining in the auditing context.

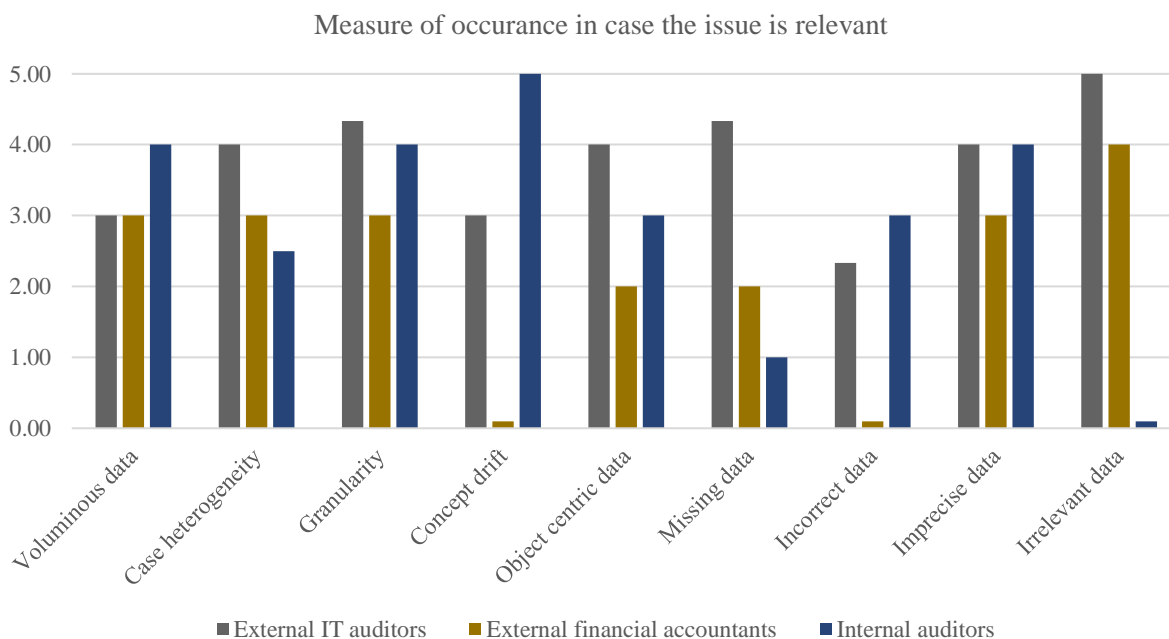


Figure 10 Measure of occurrence in case the issue is relevant according functional groups.

**Voluminous data**

Voluminous data has been found as a relevant issue, which influences the application of process mining in the auditing context in 87.50 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors is not identical, interviewee C2 (see appendix D) identifies the identifier to use in the process mining in the scoping process according to the research questions, interviewee C4 (see appendix D) finds the common identifier by finding out how the systems communicate which each other, however they almost all agree that it an approach to this issue is to understand the context of the process (see appendix D). The financial external auditor interviewee C6 is not aware of these approaches, while they are working on the same process as the external IT auditors. The internal auditor interviewees C7 and C8 (see appendix D ) both have a different manner of approaching this issue, internal auditor C7 approaches the issue in a similar manner as the external IT auditors, by obtaining context knowledge of the process, and scoping the assignment together with the customer on a data level. The internal auditor C8 (see appendix D) is familiar with the unique identifier which is connected to every case in the system.

**Case heterogeneity**

Case heterogeneity has been found as a relevant issue, which influences the application of process mining in the auditing context in 75.00 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors is not identical, (please see appendix D) the interviewee C1 suggests to filter on relevant traces based on the characteristics of activities, and validate these afterwards with the customer. Interviewee C4 suggests to group activities in different process mining models, the external financial auditors share this approach (see appendix D). The internal auditors have another approach which, interviewee C7 decides on the basis of the research questions if the process mining model is this valuable to execute, and interviewee C8 also looks into the research questions, and decides on this basis which traces are relevant.

**Granularity**

Granularity (the level of detail) has been found as a relevant issue, which influences the application of process mining in the auditing context in 87.50 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors is according C3 and C4 (see appendix D) to collaborate with the end-user of the analysis to find the correct level of granularity. According to interviewee C2, a too low level of detail is easily solved by grouping variables, and a too high level of detail is treated as the issue missing data. The approach of finding the correct level of granularity in collaboration with the end-user while keeping the trustworthiness of the model (see appendix D). The approach of internal auditors this issue is to find the correct level of detail on the basis of the research questions, which is similar as to the approach of the external IT and financial auditors.

**Concept drift**

Concept drift has been found as a relevant issue, which influences the application of process mining in the auditing context in 50.00 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors is similar (see appendix D) , understanding the context of the processes in the preparation phase and validate the model afterwards with the customer. According interviewee C2 (see appendix D) a statistical test is performed to uncover any changes in the model. The approach to this issue of the internal auditor C7 (see appendix D) is similar by understanding the context of the processes in the preparation phase.

**Object centric data**

Concept drift has been found as a relevant issue, which influences the application of process mining in the auditing context in 87.50 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors not identical, interviewee C1 and C4 (see appendix D) approach the problem by deciding if process mining is still possible on the basis of the research questions posed, the other external IT auditors (interviewees D) pose to merge and connect the merge and connect the object centric data (see appendix D). The external financial auditor C6 is not aware of any approach to handle this issue, the external financial auditor C5 approaches the issue similar to the external IT auditors by deciding if process mining is still valuable to possible. The internal auditors C7 (see appendix D) approach this issue in a similar manner by firstly deciding id process mining analysis is still valuable, and if valuable, secondly by researching the processes and interpreting the object centric data to form a process flow.

**Missing data**

Missing data has been found as a relevant issue, which influences the application of process mining in the auditing context in 75.00 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors C2 and C3 (see appendix D) is similar, in collaboration with the relevant stakeholders (customer, external financial auditor) it is discussed if the missing data can be retrieved in another manner, and besides this according to C3 discuss how this impacts the process mining analysis. Furthermore according to external auditor C2 (see appendix D) is discussed together with the customer is willing to alter event logging in the future in order to prevent missing data of future process mining analysis. The external financial auditors share the same manner of approach as the external IT auditors (see appendix D). The internal auditor C7 has another approach which entails a feasibility test previous to process mining, if this test concludes essential missing data, then the process mining analysis is not executed.

**Incorrect data**

Incorrect data has been found as a relevant issue, which influences the application of process mining in the auditing context in 50.00 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors, where two out of three cases (see appendix D) proposes a check to reveal any incorrect data, however there is no common test known which is performed within this procedure. The internal auditor C7 (see appendix D)

performs a test together with the customer on the process mining analysis together with source system together with the customer.

#### **Imprecise data**

Imprecise data has been found as a relevant issue, which influences the application of process mining in the auditing context in 62.50 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditors is as following:

Interviewee C4 (see appendix D) proposes that first of all is looked in the possibility that the activities are already ordered in the correct order. According to interviewees C1, C2 and C4 (see appendix D) dependent on the research questions and propose to interpolate timestamps. Interviewee C2 adds to this issue that there is also the possibility of free text fields which do not possess any structure, these are not analyzable. Internal auditor proposes a similar approach to this issue by first of all looking if the data does not contain the order, if a deeper look into the data is taken or interpolating timestamps in validation with the customer.

#### **Irrelevant data**

Imprecise data has been found as a relevant issue, which influences the application of process mining in the auditing context in 37.50 % of the cases. The manner of approach for handling this issue if it is found as influencing the application of process mining by external IT auditor C3 (see appendix D) is in the same manner as voluminous data. The external financial auditor's approach of C5 (see appendix D) to handle this issue is in the same manner as heterogeneity and granularity is handled. The other external financial auditor C6 (see appendix D) proposes to retrieve context knowledge of the process in collaboration with customer, and use this as a basis to decide which information is relevant.

## **5.4 PROPOSITIONS**

The following section will handle the propositions that helped guiding the multi-case interviews. For both proposition the conclusions are based on the following guideline:

- If two or more cases judge the identified issue from literature as influencing the application of process mining in the auditing context the proposition is accepted.
1. Data quality issues as a result of contemporary process deviations have an influence on the application of process mining in the auditing context?

**Table 11 Results proposition 1.**

Item	Potential issue	C1	C2	C3	C4	C5	C6	C7	C8	Conclusion
1	Voluminous data	X	X	X	X/-	-	X	X	X	Accepted
2	Case heterogeneity	X	-	-	X	X	X	X	X	Accepted
3	Granularity	-	X	X	X	X	X	X	X	Accepted
4	Concept drift	X	X	-	-	-	-	X	-	Accepted
5	Object centric data	X	X	X	X	X	X	X	-	Accepted

2. Data quality issues as a result of event log deficiencies have an influence on the application of process mining in the auditing context?

**Table 12 Results proposition 2.**

Item	Potential issue	C1	C2	C3	C4	C5	C6	C7	C8	Conclusion
1	Missing data	X	X	X	-	X	X	X	-	Accepted
2	Incorrect data	X	X	X	-	-	-	X	-	Accepted
3	Imprecise data	X	X	-	X	-	X	X	-	Accepted
4	Irrelevant data	-	-	X	-	X	X	-	-	Accepted

## 5.5 GUIDELINE VALIDATION

The following section contains the results and analysis of the validation of the guideline. First the results are discussed of the personally administered questionnaire which was executed to validate the guideline. Afterwards these results are analyzed in order to draw a conclusion.

### 5.5.1 Results validation

Please see appendix E for the personally administered questionnaire which was executed with three subject matter experts. The validation of the guideline is measured by using the content validity index. All experts must agree on each individual item level, therefore the item content validity index (CVI) should be 1.00 in order to validate the guideline.

Please see appendix F for the results of the personally administered questionnaire. Table 13 summarizes these results in one table. In this table the item CVI results are given per construct which is measured i.e. relevancy and feasibility (easy to administer and easy to interpret).

**Table 13 Item CVI of relevancy and feasibility per item.**

Guideline for identified issue	Item CVI Relevancy	Item CVI Feasibility		Conclusion
		Easy to administer	Easy to interpret	
Voluminous data, irrelevant data	1.00	1.00	1.00	Accepted
Heterogeneity, incorrect data	0.67	0.67	1.00	Rejected
Concept drift	1.00	1.00	1.00	Accepted
Object centric data	0.67	0.33	1.00	Rejected
Granularity, imprecise data	1.00	0.67	1.00	Rejected
Missing data	0.67	0.67	1.00	Rejected
Incorrect data	1.00	1.00	1.00	Accepted

### 5.5.2 Analysis validation

The conclusion of the results of the validation of the questionnaire is given in table 13. The guideline proposed for voluminous and irrelevant data, concept drift and incorrect data are found as relevant for handling the issue at hand and feasible to execute in the process mining procedure.

The proposed guideline for handling granularity and imprecise data is found to be relevant in handling the issue, however not easy to administer in the process mining procedure. The other proposed guidelines for heterogeneity and incorrect data, object centric data and missing data are found not relevant to handle the issue at hand, and not easy to administer in the process mining procedure.

## 5.6 CONCLUSION

This chapter concerned the results and analysis of the results of this research. First the results and analysis of the multi-case interview are stated from different points of view. Afterwards the stated propositions of this research are concluded. Besides this are the results of the validation of the guideline discussed. The following chapter will discuss these results and analysis.



## **6 DISCUSSION**

### **6.1 INTRODUCTION**

The following chapter concerns the discussion of this research. This entails first of all concerns a discussion of the results from the perspective of the research questions, which are posed in the introduction, and methodology of this research. Afterwards limitations of this research are given, along with the recommendations.

### **6.2 DISCUSSION**

The following section entails the discussion of the research, the structure of this section is based upon the research questions posed in this research.

The problem statement of this research is how data quality issues which influence the application of process mining in the auditing context can be handled. This problem statement can be answered using three sub-research questions. First of all, the question is answered which quality issues are known to arise in the application of process mining. To answer this question, a literature review is executed. The issues which are identified in literature can be divided within two categories, these are process characteristics and event log deficiencies. The first category deals with the challenges which arise due to contemporary company processes and the underlying information system(s) deviations and the latter category deals with issues which arise from deficiencies of the quality manifested in an event log.

Quality issues which are identified due to contemporary company processes and the underlying information system(s) deviations are voluminous data, case heterogeneity, granularity, concept drift and object centric data.

The issue of voluminous data entails that the data extracted for process mining is from an increasing wide range of contemporary, and/or legacy information systems and applications within a company are difficult to handle within process mining analysis, this leads to existing process mining algorithms which are not able to process these voluminous data, differences between different identifiers between each system and correlation problems.

The issue of case heterogeneity entails that the growing flexibility that a company needs to incorporate in their processes may possess a high amount of distinct scenarios which are difficult to handle within process mining analysis. Consequences of this issue are that existing process mining algorithms are not able to cope with case heterogeneity and that a high amount of scenarios (traces) are possible which are difficult to comprehend for human perception.

The issue of granularity entails that the data extracted from contemporary, or legacy information systems and applications within a company possess a mixed level or low level granularity, in other words; level of detail, which are difficult to handle within process mining analysis. Consequences of this issue are low or mixed level granularity event logs are difficult to comprehend for human

perception and cognitive systems and issues in finding the right level granularity for the (end) user of the analysis.

The issue of concept drift entails that dynamic processes of companies change over time which may influence the process mining model in an unforeseen manner. Consequences of this issue are outliers due to momentary changes and no insight in process due to deviations in business processes which may influence event logs in an unforeseen manner.

The issue of object centric data entails that object centric oriented systems do not record event logs along a process but on an object. The consequence of this issue is that structuring and merging event data takes a lot of time and labor as the analyst has to merge the data of all objects such as invoices and orders into an assumed process flow.

Quality issues which are identified from deficiencies of the quality manifested in an event log are missing data, incorrect data, imprecise data and irrelevant data.

The issue of missing data entails missing mandatory data which is needed for performing the analysis, incorrect data entails data in the process mining model that does not correlate with the execution of the process in reality, imprecise data entails data which is not at the correct detail level of the (end) user and irrelevant data entails process mining data that is not needed the answer the key research questions.

The second research question is if these identified research questions influence the process mining application in the financial auditing context. This question was answered using a multi-case interview methodology with participants who perform process mining analysis in the context of financial auditing. These interviews are guided by theoretical propositions (please see methodology).

Case interviews are performed with actors both in internal and external auditing, the result of these interviews are that all aforementioned issues which are identified in literature to an extent influence the process mining application within the financial auditing context.

The third research questions which is posed is how to handle these identified and verified problems within the financial auditing context. To answer this question a literature review was performed in order to find out which suggestions are made to handle these identified quality issues, and afterwards these suggestions are formed into a guideline. One cannot presume that these guidelines are relevant to solve these identified issues, and moreover if executing this guideline is feasible in the process mining procedure within the financial auditing context. For this reason the guideline is verified by means of a personally administered questionnaire with three actors in the external auditing context. The item content validity index is used to measure the findings on relevancy and feasibility.

The suggestions that are made in literature to handle these identified issues are limited. The following suggestions are made upon the aforementioned identified issues (Some suggestions overlap to a certain extent for the identified issues, therefore are these combined in the guideline):

**Proposed guideline for the issues of voluminous data and irrelevant data:**

- A process analysis in collaboration with the auditee in order to obtain context knowledge about the financial processes in scope and domain knowledge of the information systems.
- Scoping is accomplished according key research questions and the choice of identifier (the process instance which is followed in the process mining analysis) are made on the basis of the process analysis.

**Proposed guideline for the issue of granularity and imprecise data:**

- The desired level of granularity (level of detail) is defined based on the key research questions concerning the financial processes in scope, in order to find the right level of granularity for the end user of the analysis.
- A plan of approach is defined together with all relevant stakeholders, especially the end user in order to find the correct level of granularity.

**Proposed guideline for the issue of object centric data:**

- In the scoping process one has to decide if process mining adds value in answering the key research questions.
- The employment of standard extraction guidelines (scripts) are suggested in case of system(s) and/or application(s) in scope with non-customized core functionalities.

**Proposed guideline for the issue of case heterogeneity and incorrect data:**

- Visualizing each trace separately in order to make it understandable for human perception.
- Conformance checking is performed in order to replay each trace and discover any anomalies.

**Proposed guideline for the issue of missing data:**

- Retrieve any missing data of the customer.

**Proposed guideline for the issue of concept drift:**

- A statistical test is performed on the process mining data in order to detect and localize any changes in the process.

**Proposed guideline for the issue of incorrect data:**

- Validation of the process mining model by the customer.

These guidelines are verified on relevancy for handling this issue, and feasibility for executing this guideline in the process mining procedure as earlier mentioned. The guideline proposed for voluminous and irrelevant data, concept drift and incorrect data are found as relevant for handling the issue at hand and feasible to execute in the process mining procedure.

The proposed guideline for handling granularity and imprecise data is found to be relevant in handling the issue, however not easy to administer in the process mining procedure. The other proposed guidelines for heterogeneity and incorrect data, object centric data and missing data are found not relevant to handle the issue at hand, and not easy to administer in the process mining procedure.

### 6.3 LIMITATIONS

In the following section, the limitation of this research are discussed. First, this is discussed in the perspective of this topic which is researched, and afterwards in the perspective of the research methodology.

Process mining is an emerging technology, this means that the academic literature about this topic still is lacking from variety and immersion. The academics who discuss this topic, process mining are not versatile. The geographical boundary of most research on process mining in this context are executed in Europe by a few researchers, such as the researches done by Jans in Belgium (Jans, Alles, & Vasarhelyi, A field study on the use of process mining of event logs as an analytical procedure in auditing, 2014) (Jans, Alles, & Vasarhelyi, Process Mining of Event Logs in Auditing: Opportunities and Challenges, 2010) (Jans, Alles, & Vasarhelyi, Process mining of event logs in internal auditing: a case study, 2012) (Jans, Alles, & Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, 2013), and van der Aalst in the Netherlands (Van der Aalst W. , 2011) (van der Aalst, van Hee, van der Werf, & Verdonk, 2016). This means that there is not sufficient variety in researches. Also the immersion of research of process mining, especially of data quality in process mining is not sufficient enough. There exists few academic literature about handling data quality issues within process mining, this limited the research on making a guideline for handling of verified data quality issues influencing the application in the auditing context.

Besides this are there limitations bound to the research methodology which is used in this research. Multiple-case studies are used in this research in order to find out which identified issues influence the process mining application in the financial auditing context. The issue of generalization of results of case studies is a well-known issue appeared in literature (Tellis, 1997). For this reason multiple-case studies are executed from different functional groups within the financial auditing context i.e. internal auditors, IT auditors and financial auditors. However these functional groups, for example the internal auditors, do not include internal auditors from all relevant industries, and the IT and financial auditors are taken from the financial services industry for SMEs, thus not for large sized companies. This limitation is due to the constraint of time and availability of sources in the scope of this research execution.

### 6.4 RECOMMENDATIONS

The following section consist out of recommendations these are based upon the discussion and limitations previously discussed in this chapter.

From the perspective of the results which are discussed in this chapter, the recommendation for future research are to progress in academic research on handling the identified data quality issues within process mining, in the financial auditing context.

From the perspective of the limitations of this research which are discussed in this chapter, more varied and immersed research should be undertaken within the topic of process mining.

An issue of conducting multi-case studies is the generalizability of the findings. For this reason, more research should be undertaken to achieve external validity of results. This could be accomplished by broadening the scope of this research with more varied cases, representing each relevant industry.

## 7 CONCLUSION

The following section concerns the conclusion of this research. First the topic of this research will be discussed and afterwards a summary on the results of the problem statement of the research, along with the research questions.

As mentioned in the introduction and methodology, the problem statement of this research is: How can data quality issues which influence the application of process mining in the auditing context be handled? Therefore, first of all, the question should be answered which quality issues arise in the application of process mining. Within literature nine data quality issues are found which can be categorized within two categories, these are; data quality issues as a result of contemporary company process deviations, and data quality issues as a result of event log deficiencies.

Data quality issues as a result of contemporary company process deviations concerns the issues named voluminous data; data extracted is from an increasing wide range of contemporary, and/or legacy information systems and applications within a company are difficult to handle within process mining analysis, case heterogeneity; growing flexibility that a company needs to incorporate in their processes may possess a high amount of distinct scenarios which are difficult to handle within process mining analysis, granularity; the data extracted from contemporary, or legacy information systems and applications within a company possess a mixed level or low level granularity which are difficult to handle within process mining analysis, concept drift; dynamic processes of companies change over time which may influence the process mining model in an unforeseen manner and object centric data; structuring and merging event data takes a lot of time and labor for object centric oriented systems.

Data quality issues as a result of event log deficiencies concerns the issues named missing data; missing mandatory data which is needed for performing the analysis, incorrect data; process mining data does not correlate with the execution of the process in reality, imprecise data; data which is not at the correct detail level of the (end) user and irrelevant data; process mining model that consists out of data which is not needed to answer the research question.

The following question is if these identified data quality issues influence the application of process mining in the financial auditing context. This question is raised because the processes which are in scope for the financial audit possess a certain control within a company in order to comply with regulatory and law requirements, and are for this reason not comparable with any other process where process mining is applied. This question is answered using a multi-case interview with participants who perform process mining in the auditing context. The result of this multi-case interview is that all identified issues in literature influence the application of process mining in the auditing context.

The last research question to answer the problem statement of this research is how to handle these identified data quality issues in the application of process mining in the auditing context. This question is answered by creating a guideline with the use of literature, and verifying it with experts on the field of process mining in the auditing context. The conclusion of this is that there is not yet

enough academic research undertaken to adequately handle the all identified issues that influence the application of process mining. Therefore, more progress in academic research has to be undertaken. The proposed guideline for handling the identified issues: voluminous, irrelevant data, concept drift and incorrect data are found as relevant for handling the issue at hand and feasible to execute in the process mining procedure. This entails for the issue voluminous data conducting a process analysis in collaboration with the customer and scoping of the process mining analysis based on key research questions and the process analysis. For the issue concept drift this entails a statistical test in order to uncover any changes within the company process. As last, for incorrect data this means validating the process mining model by the customer.

## 8 REFERENCES

- (2012). In I. Stuart, *Auditing and Assurance Services: An Applied Approach* (pp. 126-139). New York, NY: McGraw-Hill.
- (2014). In B. W. B.V.B.A., *BDO Audit manual* (pp. 28-36).
- Agrawal, R., Gunopulos, D., & Leymann, F. (1998). Mining Process Models from Workflow Logs. *Sixth International Conference on Extending Deatabas Technology*, 469-483.
- Akkerman, S., Admiraal, W., Brekelmans, M., & Oost, H. (2006). Auditing Quality of Research in Social Sciences. *Quality & Quantity*.
- Bezerra, F., Wainer, J., & van der Aalst, W. (2009). Anomaly detection using process mining. *Enterprise, business-process and information systems modeling*, 149-161.
- Bose, R., Mans, R., & van der Aalst, W. (2013, April). Wanna Improve Process Mining Results? *Computational Intelligence and Data Mining (CIDM)*, 127-134.
- Bose, R., van der Aalst, W., Žliobaitė, I., & Pechenizkiy, M. (2011). Handling concept drift in process mining. *International Conference on Advanced Information Systems Engineering*, 391-405.
- Bozkaya, M., Gabriels, J., & van der Werf, J. (2009). Process diagnostics: a method based on process mining. *Information, Process, and Knowledge Management*, 22-27.
- Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*(14), 1-10. doi:<http://doi.org/10.5334/dsj-2015-002>
- Chen, H., Chiang, R., & Storey, V. (2010). Business intelligence research. *MIS Quarterly*, 34(1), 201-203.
- Coney. (2016, 3 17). *Casestudy Audit firms: Integration of process mining*. Retrieved from Technische Universiteit Eindhoven: [http://www.win.tue.nl/ieeetfpm/doku.php?id=shared:process\\_mining\\_case\\_studies](http://www.win.tue.nl/ieeetfpm/doku.php?id=shared:process_mining_case_studies).Definition
- Daniels, H., & Feelders, A. (2000, August). Methodological and practical aspects of data mining. *Information & Management*(37), 271-281. doi:10.1016/S0378-7206(99)00051-8
- Davis, L. (1992). Instrument review: Getting the most from your panel of experts. *Applied Nursing Research*(5), 194-197.
- de Kok, P. (2016, 8 28). *De ware liefde van de Accountant 3.0*. Retrieved from Accountant: [www.accountant.nl/artikelen/2016/8/dewareliefdevandeaccountant3.0/?ctx=switchcomments-take\(6\)#](http://www.accountant.nl/artikelen/2016/8/dewareliefdevandeaccountant3.0/?ctx=switchcomments-take(6)#)
- Deckers, F., & Van Kollenburg, J. (2016). *Elementaire theorie accountantscontrole*. Winschoterdiep: Noordhoff Uitgevers B.V.
- Diamantini, C., Genga, L., Potena, D., & van der Aalst, W. (2015). Towards Process Instances Building for Spaghetti Processes. *SEBD*, 256-263.



- Eckerson, W. W. (2002). Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute*.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*(17(3)), 37.
- Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information processing & management*(30(1)), 9-19.
- Gehrke, N., & Mueller-Wickop, N. (2010). Basis Principles of Financial Process Mining A Journey through Financial Data in Accounting Information Systems. *AMCIS*, 289.
- Gilbert, G., & Prion, S. (2016). Making Sense of Methods and Measurement: Lawshe's Content Validity Index. *Clinical Simulation In Nursing*(12 (12)), 530 - 531.
- Gunther, C., Rinderle-Ma, S., Reichert, M., Van Der Aalst, W., & Recker, J. (2008). Using process mining to learn from process changes in evolutionary systems. *International Journal of Business Process Integration and Management*(3(1)), 61-78.
- Harel, D., & Rumpe, B. (2004). Meaningful modeling: what's the semantics of " semantics"? *Computer*(37(10)), 64-72.
- Huh, Y., Keller, F., Redman, T., & Watkins, A. (1990). Data quality. *Information and software technology*(32 8)), 559-565.
- ICAEW. (2016, 11 3). *Data analytics for external auditors*. Retrieved from ICAEW:  
<https://www.icaew.com/-/media/corporate/files/technical/iaa/tecpln14726-iaae-data-analytics--web-version.ashx>
- Investopedia. (2016, 11 3). *Generally Accepted Accounting Principles - GAAP*. Retrieved from Investopedia: Generally Accepted Accounting Principles - GAAP
- Jans, M., Alles, M., & Vasarhelyi, M. (2010, February). Process Mining of Event Logs in Auditing: Opportunities and Challenges. *SSRN*. Retrieved from [ssrn.com/abstract=1578912](https://ssrn.com/abstract=1578912)
- Jans, M., Alles, M., & Vasarhelyi, M. (2012). Process mining of event logs in internal auditing: a case study. Retrieved from  
<https://uhdspace.uhasselt.be/dspace/bitstream/1942/14227/1/Process%20Mining%20Case%20Study%20EAA.pdf>
- Jans, M., Alles, M., & Vasarhelyi, M. (2013). The case for process mining in auditing: Sources of value added and areas of application. *International Journal of Accounting Information Systems*.
- Jans, M., Alles, M., & Vasarhelyi, M. (2014). A field study on the use of process mining of event logs as an analytical procedure in auditing. *Accounting Review*, 1751-1773. doi:10.2308/accr-50807
- Jordan, J., & Turner, B. (2008). The feasibility of single-item measures for organizational justice. *Measurement in Physical Education and Exercise Science*(12(4)), 237-257.

- Koopman, A., & de Kok, P. (2014). Process mining: Leuk voor de liefhebber of noodzaak? *Informatie*, 43-48.
- Lim, E.-P., Chen, H., & Chen, G. (2012). Business Intelligence and Analytics: Research Directions. *ACM Trans.Manage. Inf. Syst.* 3, 4, Article 17 (January 2013), 17-26.  
doi:<http://dx.doi.org/10.1145/2407740.2407741>
- Lu, X., Fahland, D., & Van der Aalst, W. (2014). Conformance checking based on partially ordered event data. *In International Conference on Business Process Management*, 75-88.
- Meyer, C. (2001). A case in case study methodology. *Field methods*(13(4)), 329-352.
- Polit, D., & Beck, C. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health*(29(5)), 489-497.
- Prima Kurniati, A., Kusuma, G., Agung, G., & Wisudiawan, A. (2015). DESIGNING APPLICATION TO SUPPORT PROCESS AUDIT USING PROCESS MINING. *Journal of Theoretical and Applied Information Technology*(Vol 80. No. 3), 473.
- Rozinat, A., de Medeiros, A., Günther, C., Weijters, A., & Aalst, W. (2007, September). The need for a process mining evaluation framework in research and practice. *International Conference on Business Process Management*, 84-89.
- Sekaran, U., & Bougie, R. (2011). *Research method for business: A skill building approach*. New York: John Wiley & Sons, Inc.
- Tellis, W. M. (1997). Application of a case study methodology. *The qualitative report*(3(3)), 1-19.
- Van der Aalst, W. (2011). *Process Mining*. Berlin Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-19345-3
- Van der Aalst, W., & Weijters, A. (2004). Process mining: a research agenda. *Computers in industry*(53(3)), 231-244.
- Van der Aalst, W., Adriansyah, A., De Medeiros, A., Arcieri, F., Baier, T., Blicke, T., . . . (2011, August). Process mining manifesto. *International Conference on Business Process Management*, 169-194.
- van der Aalst, W., van Hee, K., van der Werf, J., & Verdonk, M. (2016, 11 3). *Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor*. Retrieved from Technische Universiteit Eindhoven: <http://www.wis.win.tue.nl/~wvdaalst/publications/p593.pdf>
- van der Heijden, B., & Bajnath, S. (2015). IT-AUDIT IN 2015 AND BEYOND: DATA DRIVEN! *Compact*\_3, 26-31. Retrieved 11 3, 2016, from [https://www.compact.nl/articles/it-audit-in-2015-and-beyond-data-driven/?zoom\\_highlight=IT+audit+in+2015+en+verder](https://www.compact.nl/articles/it-audit-in-2015-and-beyond-data-driven/?zoom_highlight=IT+audit+in+2015+en+verder)
- van Raak, J., & Thürheimer, U. (2016, September). Opportunities to improve the measurement of audit quality: a call for collaboration between the profession and academics. *MAB*(9), 352-358.
- Wand, Y., & Wang, R. (1996, November). Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*(Vol.39, No. 11), 86-95.

- Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*(12(4)), 5-33.
- Werner, M., Gehrke, N., & Nüttgens, M. (2012). Business process mining and reconstruction for financial audits. *Proceedings of the Annual Hawaii International Conference on System Sciences*. ResearchFate.
- Yin, R. K. (2013). *Case study research: Design and methods*. Sage publications.

## 9 APPENDIXES

### APPENDIX A - INTERVIEW PROTOCOL - INVITATIONS

The following invitations are send per email to contact persons to invite participants to cooperate in the multi-case interviews, and afterwards in the personally administered questionnaire.

Multi-case interview invitation (translated from Dutch to English)

Dear NAME,

I am currently busy with my thesis about the data quality within process mining. For this reason I would like to conduct interviews with people who have experience with the application of process mining.

The goal of my thesis is to verify identified data quality issues out of literature in the financial auditing context, and with this information create guidelines to handle these (verified) quality issues.

The target group for these interviews are internal, IT and financial auditors. For this reason, I would like to invite you to participate in an interview.

Could you please assist me in this research?

Thanks in advance for your time and input!

Personally administered questionnaire invitation (translated from Dutch to English)

Dear NAME,

I have recently interviewed you about the identified data quality issues in process mining in the financial auditing context. With the aid of this information I created guidelines according to suggestions made in literature. I need to validate this guideline with experts in the context of financial auditing. I would like to invite you to participate in a personally administered interview. This will take up to 30 minutes of your time.

Could you please assist me in this research?

Thanks in advance for your time and input!

## APPENDIX B – BASIC STRUCTURE MULTI-CASE INTERVIEWS

Do these issues have a relevant influence on the application of process mining?					
<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>	Relevant	Irrelevant	If relevant, could you please give a short explanation how you handle these issues?
1.	What is your current function?				
2.	In which sector do you operate?				
3.	How many years do you have experience with process mining?				
4.	<i>Voluminous data</i>	<i>The data extracted is from an increasing wide range of contemporary, and/or legacy information systems and applications within a company which are difficult to handle within process mining analysis</i>	- Existing process mining algorithms are not able to cope with case voluminous - Differences between the identifiers between different information systems - (Secondary) Correlation problems	0	0
5.	<i>Case heterogeneity</i>	<i>Increasing flexibility that a company needs to incorporate in their processes may result in process without a clear structure or flow</i>	- Existing process mining algorithms are not able to cope with case heterogeneity - No structure within process flow	0	0

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>	Relevant	Irrelevant	If relevant, could you please give a short explanation how you handle these issues?
<b>6.</b> <i>Granularity</i>	<i>The data extracted from contemporary, or legacy information systems and applications within a company possess a mixed level or low level granularity which are difficult to handle within process mining analysis</i>	<ul style="list-style-type: none"> <li>- <i>low level granularity event logs are difficult to comprehend for human perception and cognitive systems</i></li> <li>- <i>Issues in finding the right level granularity for the (end)user of the analysis</i></li> </ul>	0	0	
<b>7.</b> <i>Concept drift</i>	<i>Dynamic processes of companies change over time which may influence the process mining analysis in an unforeseen manner</i>	<ul style="list-style-type: none"> <li>- <i>Outliers due to momentary changes</i></li> <li>- <i>No insight in process</i></li> </ul>	0	0	
<b>8.</b> <i>Object centric data</i>	<i>Systems do not record event logs along a process but on an object which is harder to analyze</i>	<ul style="list-style-type: none"> <li>- <i>structuring and merging event data takes a lot of time and labor</i></li> </ul>	0	0	

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

How often do you experience consequences of these issues within the application of process mining? If the issue was experienced as irrelevant in the previous questions, then the answer here is N/a.

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>	-1- Almost never	-2- Seldom	-3- Sometimes	-4- Often	-5- Almos t always	N/a
<b>9.</b> <i>Voluminous data</i>	<i>The data extracted is from an increasing wide range of contemporary, and/or legacy information systems and applications within a company which are difficult to handle within process mining analysis</i>	<ul style="list-style-type: none"> <li>- <i>Existing process mining algorithms are not able to cope with case voluminous</i></li> <li>- <i>Differences between the identifiers between different information systems</i></li> <li>- <i>(Secondary) Correlation problems</i></li> </ul>	0	0	0	0	0	0
<b>10.</b> <i>Case heterogeneity</i>	<i>Increasing flexibility that a company needs to incorporate in their processes may result in process without a clear structure or flow</i>	<ul style="list-style-type: none"> <li>- <i>Existing process mining algorithms are not able to cope with case heterogeneity</i></li> <li>- <i>No structure within process flow</i></li> </ul>	0	0	0	0	0	0

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>	-1- Almost never	-2- Seldom	-3- Sometimes	-4- Often	-5- Almos t always	N/a
<b>11.</b> <i>Granularity</i>	<i>The data extracted from contemporary, or legacy information systems and applications within a company possess a mixed level or low level granularity which are difficult to handle within process mining analysis</i>	- <i>low level granularity event logs are difficult to comprehend for human perception and cognitive systems</i> - <i>Issues in finding the right level granularity for the (end)user of the analysis</i>	0	0	0	0	0	0
<b>12.</b> <i>Concept drift</i>	<i>Dynamic processes of companies change over time which may influence the process mining analysis in an unforeseen manner</i>	- <i>Outliers due to momentary changes</i> - <i>No insight in process</i>	0	0	0	0	0	0
<b>13.</b> <i>Object centric data</i>	<i>Systems do not record event logs along a process but on an object which is harder to analyze</i>	- <i>structuring and merging event data takes a lot of time and labor</i>	0	0	0	0	0	0
<b>14.</b>	<i>Are there other process characteristics which potentially have a relevant influence on the application of process mining?</i>							



## DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

Do these issues have a relevant influence on the application of process mining?							
<i>Issue</i>	<i>Example</i>	Relevant	Irrelevant	If relevant, could you please give a short explanation how you handle these issues?			
<b>15.</b>	<i>Missing data</i>	<i>A process step took place in reality but was not recorded in the event log</i>	0	0			
<b>16.</b>	<i>Incorrect data</i>	<i>Due to a system error, recorded process steps are assigned to the wrong process human error</i>	0	0			
<b>17.</b>	<i>Imprecise data</i>	<i>Timestamps only record the day of the recorded process step</i>	0	0			
<b>18.</b>	<i>Irrelevant data</i>	<i>Certain processes in the analysis are not needed to answer the research question</i>	0	0			
How often do you experience consequences of these issues within the application of process mining? If the issue was experienced as irrelevant in the previous questions, then the answer here is N/a.							
<i>Potential issue</i>	<i>Example</i>	-1- Almost never	-2- Seldom	-3- Sometimes	-4- Often	-5- almost always	N/a
<b>19.</b>	<i>Missing data</i>	<i>A process step took place in reality but was not recorded in the event log</i>	0	0	0	0	0
<b>20.</b>	<i>Incorrect data</i>	<i>Due to a system error, recorded process steps are assigned to the wrong process</i>	0	0	0	0	0
<b>21.</b>	<i>Imprecise data</i>	<i>Timestamps only record the day of the recorded process step</i>	0	0	0	0	0
<b>22.</b>	<i>Irrelevant data</i>	<i>Certain processes in the analysis are not needed to answer the research question</i>	0	0	0	0	0
<b>23.</b>	Are there other event logging issues which have a relevant influence on the application of process mining?						

## APPENDIX C – SUMMURIZED RESULTS MULTI-CASE INTERVIEWS

Question description \Case ID	C1			C2			C3			C4		
<b>Current function</b>	Junior IT Auditor			Manager IT audit			Junior IT auditor			Junior IT auditor		
<b>Sector</b>	Financial Services industry			Financial services industry			Financial services industry			Financial services industry		
<b>Experience with process mining</b>	1 year practice			2 years practice			1.5 years practice			1.5 - 2 years		
<b>Issue</b>	Relevancy for application	Measure of occurrence	Manner of approach of relevant issue	Relevancy for application	Measure of occurrence	Manner of approach of relevant issue	Relevancy for application	Measure of occurrence	Manner of approach of relevant issue	Relevancy for application	Measure of occurrence	Manner of approach of relevant issue
<b>Voluminous data</b>	Relevant	Often	Merge and connect data, validate connection afterwards with customer.	Relevant, especially for the audit as ideally you need a large time slot of a whole year.	Sometimes	Scope data with the aid of research questions in preparation phase e.g. which identifier to use. Use standardized scripts for	Relevant	Sometimes	Retrieve the context of the process steps together with the customer by following one trace.	Both relevant as irrelevant, depends on the amount of systems that the process support.	Seldom	Find the common identifier, by finding out how the systems communicate which each other.

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

						extraction of data. Understand context of process. Merge automated process steps if needed.						
<b>Case heterogeneity</b>	Relevant	Almost always	Understand context of process in preparation phase. Filter on relevant traces based on the characteristics of activities, and validate analyzes with customer afterwards.	Irrelevant, not applicable in audit practice.	N/a.	Scoping the assignment in the preparation phase on the cornerstone activities mitigates the risk of too many traces	Irrelevant, Purchase and sales processes subject to the audit are controlled in a sufficient manner within a company	N/a.		Relevant	Sometimes	Grouping orders in different process mining models.
<b>Granularity</b>	Irrelevant, for lower	N/a.	Occurs frequently,	Relevant, too high	Often	Too low level of	Relevant	Almost always	After the first initial	Relevant	Often	Collaborate with the end-

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

	mixed level of granularity		however is easily fixed by filtering out too much detail in the analysis. Too high level granularity is treated as missing data.	level of data.		detail is easily solved by grouping details. Too high level of data is relevant problem as you encounter missing data.			analysis, the model is altered to the correct level of granularity, this happens in collaboration with the accountant and knowledge of the context of the audit.			user to find the correct level of granularity in the process mining model.
<b>Concept drift</b>	Relevant	Sometimes	Understand context of process in the preparation phase and add extra fields if changes are applicable. Take a small timeslot to mitigate the	Relevant	Sometimes	Understand context of process in preparation phase. Divide data set into 2 timeslots and compare these, and validate afterwards	Irrelevant	N/a.	Purchase and sell processes do not change often.	Irrelevant	N/a.	No comment.

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

			risk of process changes. Validate with customer afterwards.			with customer or with subject matter expert within company.						
<b>Object centric data</b>	Relevant	Sometimes	Quit process mining analysis	Relevant	Often, IS relevant for the audit are mostly object centric.	Understand context of process in preparation phase and object centric data. Afterwards merge object centric data and produce interpretation of process flow.	Relevant	Almost always	Have knowledge of the application, and use a structured script to extract the data of the application in scope.	Relevant	Often	Scope the key questions to the analysis which are possible with object centric data.
<b>Other relevant process characteristics issues?</b>	The IT auditor is unaware of process scope and output, this could be solved by communicating and agreement of scope to customer in the scoping phase.			No.			The background processes of a system can produce an immense size of data which pollute the data with irrelevant information.			No.		

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

<b>Missing data</b>	Relevant	Almost always	Filter on completed processes with Follower function.	Relevant	Almost always	Discuss with customer which data is missing, and decide if this data is relevant for analysis. Discuss if customer is willing to alter event logging for the future.	Relevant	Sometimes	In collaboration with the accountant and the customer. With the customer in order to retrieve any missing data if possible, and with the accountant to discuss which impact this may have on the process mining analysis.	Irrelevant	N/a.	I have not experienced this yet, we could retrieve the data which was missing in most cases.
<b>Incorrect data</b>	Relevant	Almost never	No comment.	Relevant, only due to human errors	Often	A check with source system is executed in validation phase.	Relevant	Seldom	Checked if the process was logged in a “logical manner” and	Irrelevant	N/a.	I have not experienced this yet, any outliers found in the data could be

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

									afterwards corrected any incorrect data.			explained by the customer.
<b>Imprecise data</b>	Relevant	Often	Dependent on research question. Interpolate timestamps in case of imprecise timestamps.	Relevant	Often	Understand context of process in preparation phase. Dependent on research question. Interpolate timestamps in case of imprecise timestamps. Free text fields do not have any structure and therefore not analyzable.	Irrelevant	N/a.	No comment.	Relevant	Often	Event registration could be in the correct order, interpolation of timestamps by following the process logically. However this impacts the trustworthiness of analysis.
<b>Irrelevant data</b>	Irrelevant	N/a.	You are able to remove irrelevant	irrelevant	N/a.	No comment.	Relevant	Almost always	In the same manner as voluminous	Irrelevant	N/a.	Problem has little impact on the process

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

			information easily in ETL process and by filter in the analysis.						data, and filter out any unnecessary information.			mining analysis.
<b>Other relevant event log issues?</b>	No.			The quality of the recorded data is very important, this is most of the time manually entered. So you are dependent on the person who entered the data, this is an inherent constraint.					You have to know what the variable means in a certain table within an application, otherwise it might be dangerous with interpreting the data. You need domain knowledge of the customer to know this, of how they use the system. This is a more or less a problem with object centric data. This becomes a larger problem if you have customized systems, to know what certain variables means.			No.



**APPENDIX D – MULTI-CASE INTERVIEW TRANSCRIPTS**

The following transcripts contain the results of the multi-case interviews. These are translated from Dutch to English.

Transcript – ID = C1

Actor

Interviewer What is your current function?

C1 Junior IT Auditor

Interviewer In which sector do you operate?

C1 Financial Services

Interviewer How many years do you have experience with process mining?

C1 Approximately 1 year, with several focuses.

Accountants do not know how to use process mining in their analysis because they are not yet experienced with the tool, another hurdle is that they find it hard to establish confidence in the data that this tool produces, in order to use it in the yearly audit on the financial statement.

Interviewer Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?

C1 Yes, It is hard when working with for example two different systems in one company, to make the connection between those systems. You cannot make a good analysis, if you do not have this under control. For example, in system A is not every event recorded, however in system B every event is recorded. Then you can see in system B that segregation of duty is applied, however you cannot see in system A if the same application is done here. So it is crucial that the connection between the systems is OK. In reality it is common that there are more than 1 systems in place for the process(es) in scope, and with every system added, more process steps are added too. The connection which is made is validated with the customer.

Interviewer Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?

C1 Yes, it is a relevant problem for the analysis. If you see this case, these are all process steps. The amount of traces here are too much to make sense of it. So you have to possess knowledge of the process before analyzing it in order to know which process steps are relevant and which are not. The way to handle this issue is to cutoff activities with a filter on the basis of the characteristics of the activities. For example if an activity is automatically done by the system, it is not relevant because it will always be executed in any situation. However the downside of this solution is that you make risk losing relevant data in your analysis. Cutting off activities in Disco is pretty easy, and you are able to trace the steps taken in filtering afterwards, and here is validation with the customer also a normal way of handling this issue in the validation process step of process mining.

Interviewer Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?

C1 No I think it does not have a relevant influence on the application of process mining. This happens a lot with automatically recorded event logs by a system, that these have a low level of detail. The details which are not relevant for analysis are easily filtered out of the analysis. So this happens a lot, however I do not see this as a potential problem, because it is solved easily.

Interviewer Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?

C1 How to handle with this issue is to take a small timeslot of data, then the chance of changes within the process itself is smaller. If you aware of for example several changes in the process due to for example seasons. Then you are able to add extra fields in the analysis, for every season. Then you are able to see per season what the process is like. To become aware of the changes in the process, you validate the process with the customer.

- Interviewer Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C1 My vision for the future process mining analysis, is to quit with the intention to do process mining analysis for the systems if this is the case. This is because the added value of process mining is lost due to object centric data in systems. The goal of process mining is that you discover how the process in reality works, and in case of an object centric you would have to give your own interpretation of how the process flows. You will see for example how many hours are invoiced in a system, but you will never see how the actual process was executed in reality. Besides this is handling object centric data also very time consuming. So it is relevant.
- Interviewer How often do you experience consequences of voluminous data within the application of process mining?
- C1 Often, because most of the time, several systems are in place for the processes in scope. The challenge here is that you need to make the connection between those systems, which is hard, or even impossible in some cases. Because there are differences between those systems, for this reason there are also two different processes.
- Interviewer How often do you experience consequences of case heterogeneity within the application of process mining?
- C1 Almost always
- Interviewer How often do you experience consequences of granularity within the application of process mining?
- C1 N/a
- Interviewer How often do you experience consequences of concept drift within the application of process mining?
- C1 Sometimes
- Interviewer How often do you experience consequences of object centric data within the application of process mining?
- C1 Half of the time, so sometimes
- Interviewer Are there other process characteristics which potentially have a relevant influence on the application of process mining?
- C1 In the BDO defined process, for process mining. The process mining analysis is sold by the accountant, and as IT auditor you get the assignment of the analysis and the deadline. Also the customer in this process has a powerful position in what it wants to know or see in the process mining analysis. So as an IT auditor you never know what your final input is or in other words, when the analysis is completed. This can be handled by giving the customer the scope of the process analysis beforehand.
- Interviewer Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C1 This is relevant. For example with processes which started or ended outside the time scope of the data, you can handle this with the follower function in the tool Disco. This filters on completed processes.
- Interviewer Does incorrect data, have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C1 I have not experienced this. However this is relevant. We do not perform tests with a sample of the data if this correlates with the source system, we only look at what is logical and what not.
- Interviewer Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C1 This is relevant, how do we handle with this issue? You order the process steps yourself in a logical manner. This is a fairly easy solution, however this potential issue could still jeopardize your analysis. Because you do make a premise with ordering the process steps, for this reason you lose details in your analysis and some validation in your analysis. However how much details and validation you lose is depended on the research question of the analysis, you still know how many persons for example authorized an order.

- Interviewer Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C1 No not relevant, already in the ETL process you are able to delete unnecessary data, for example the address details of a suppliers, and in the tool Disco you are able to filter any irrelevant data out of the analysis.
- Interviewer How often do you experience missing data within the application of process mining?
- C1 Almost always
- Interviewer How often do you experience incorrect data within the application of process mining?
- C1 Almost never
- Interviewer How often do you experience imprecise data within the application of process mining?
- C1 Often
- Interviewer How often do you experience irrelevant data within the application of process mining?
- C1 N/a.
- Interviewer Are there other event logging issues which have a relevant influence on the application of process mining?
- C1 No.

Transcript – ID = C2

Actor

- Interviewer What is your current function?
- C2 Manager IT audit
- Interviewer In which sector do you operate?
- C2 Financial Services
- Interviewer How many years do you have experience with process mining?
- C2 Two years
- Interviewer Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 Relevant, how we handle this? In the preparation phase, in the scoping process we pose the question in which part of the process we would like to look into. Which identifier we need for example. For many standard systems we have scripts in place in order to extract only the data we need. When using these scripts, we still see a large amount of process steps. So before we start analyzing we look into the data if we do not have too much data. For example with automatically recorded event logs, when you see that there is a large amount of recorded events in one second, you know that these can be reduced to only one event log, because a person can only do one thing in one second.
- Interviewer Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 In reality in the audit practice, you do not see this problem very often. This can be a relevant problem, however we handle this in our process mining analysis. Only when we deal with customized systems we see an increase in traces. However, here we also deal with this issue in the same manner as with voluminous data. So we already look into this issue I the preparation phase, during the scoping of the assignment when you select the cornerstone activities of the process. So afterwards we do not have the issue that the process mining algorithm is not able to cope with incorporating all traces.
- Interviewer Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 Yes this the most relevant problem for us. Too much detail in event logs is not a problem for us, in the preparation phase we group the event logs to an appropriate level for us to analyze. For example, with a purchase order, when you have all changes of this purchase order recorded, then you group these changes and name them; ‘purchase order changes’ for the proceeding analysis. On the other hand, with a too high level of event logs, you still are able to perform a process mining analysis, however the analysis will not go very deep into the process, because you do not have the relevant information. This is a relevant problem for us.

- Interviewer Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 This is relevant, this happens in reality. We do check this with the customer, but sometimes the customer is also not aware of this. So partly we can handle this in the preparation phase with checking for changes with the customer. However we still, after the first analysis, can stumble upon changes in the process. This can be checked by taking two arbitrary timeslots of data which is in scope, and check and measure if there are any changes visible between these two timeslots. Besides this, on the technical side, definitions can change, or activities can change in the system. The customer is in this case most of the time not aware of these problems. However if we do see this in the analyses we validate these change with the application or system manager. So it is a problem, but we are able to handle this.
- Interviewer Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 Relevant, happens a lot in reality. Many of the systems which customers have in place are not workflow oriented. We have to retrieve the context of this information at the customer. How we handle this? Well you have the data of the invoice, and from the order and the movement of goods etc., on the basis of this information you interpret what has happened in reality. So you retrieve from all transaction data, what really happened in the process. This takes considerable time the first time, however this is not the case if you get experienced with these systems, then it does not take more time than usual with a workflow oriented system. However you do take premises in the analysis and therefore lose detail, this is important for the trustworthiness of the analysis as such.
- Interviewer How often do you experience consequences of voluminous data within the application of process mining?
- C2 Sometimes, this is a larger problem in the audit, because we use the period of a whole year for an audit of the financial statement of a company.
- Interviewer How often do you experience consequences of case heterogeneity within the application of process mining?
- C2 N/a
- Interviewer How often do you experience consequences of granularity within the application of process mining?
- C2 Often
- Interviewer How often do you experience consequences of concept drift within the application of process mining?
- C2 Sometimes
- Interviewer How often do you experience consequences of object centric data within the application of process mining?
- C2 Often, this is perhaps for the reason that systems which are relevant for the audit on financial statements tend to be more object centric in reality.
- Interviewer Are there other process characteristics which potentially have a relevant influence on the application of process mining?
- C2 No.
- Interviewer Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 Relevant for the application for process mining. There is not a clear cut solution for this issue, as data is missing. One thing that we do is communicate with the customer and check first of all, if the data that we are missing is relevant for analysis and secondly discuss if they are able and willing to log the missing data for future.
- Interviewer Does incorrect data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 I have not experienced incorrect data due to a system error. However it is difficult to find the correct context of an activity which has been recorded in the source system. In the validation phase, the process mining analysis data is compared by the data in the source system. So this is a relevant problem, however not due to a system error, but because

- of human error in data entry. Systems errors are for example duplicate event logs, however this is easy to fix. Human errors are not easy to fix.
- Interviewer Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 Relevant, you are missing detail info. It is a hurdle for process mining info. The process mining tool does not know in which sequence the activities are executed if these only carry a date of execution in the event log, and several activities are logged on that specific date. How we handle this, is that we order the activities ourselves as indicated in the process descriptions. However this is dangerous because you make premises on how process steps are executed.  
Another problem are free text fields, for the reason that these fields do not have a structure and therefore hard to analyze. These fields possess activities which hand written. In this case you do have a correct level of granularity, however there is no structure in this fields to systematically analyze this.
- Interviewer Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C2 This is not really an issue, you always have irrelevant data within a dataset.
- Interviewer How often do you experience missing data within the application of process mining?
- C2 Almost always. Depends on which analysis you want to do, as how far the data allows it.
- Interviewer How often do you experience incorrect data within the application of process mining?
- C2 Often. When you see compare the process mining data with reality. However this is more due to the fact of understanding the context in which the recorded event logs are produced. A human error can for example be, recorded an incorrect step, or using an incorrect function to record a process step.
- Interviewer How often do you experience imprecise data within the application of process mining?
- C2 Often. This can involve a mixed level of granularity between systems.
- Interviewer How often do you experience irrelevant data within the application of process mining?
- C2 N/a.
- Interviewer Are there other event logging issues which have a relevant influence on the application of process mining?
- C2 The quality of the recorded data is very important, this is most of the time manually entered. So you are dependent on the person who entered the data, this is an inherent constraint.

Transcript – ID = C3

Actor

- Interviewer What is your current function?
- C3 JR. IT Auditor
- Interviewer In which sector do you operate?
- C3 Financial services
- Interviewer How many years do you have experience with process mining?
- C3 1.5 year
- Interviewer Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C3 If you have one system, then it is not relevant, however this issue becomes relevant if you have several systems which support a process and for this reason have to connect these systems with each other. Thus this is a relevant problem. How to handle this? Retrieve the context of the process steps together with the customer, by for example following an order together with the customer in order to validate the process. This happens predominantly in the validation phase of the process mining process, after the first initial analysis of process mining.
- Interviewer Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C3 Irrelevant, because I do purchase and sales processes, because these are subject to the audit. These processes are controlled in a sufficient manner, therefore are there a limit of traces



- possible. It does take experience to do this, so you do have to know where you have to look into.
- Interviewer Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C3 Relevant, how do we handle this? First you do an initial analysis and then the process mining tool discovers a certain model, then you look into the model and see which “garbage” you can throw out by using a filter, which is irrelevant for the audit. This happens in collaboration with the accountant and knowledge of the context of the audit. Most of the time, low level of data is a duplication of data.
- Interviewer Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C3 Processes which we analyze for the audit do not change that much, purchase and sales processes. So in this perspective is this problem not recognized. Thus irrelevant.
- Interviewer Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C3 Relevant problem for me, as this requires a lot of work for me. How do I handle this, this is very complicated, you have to know in the application where the data that you need is situated. For example for one event log, you need 18 different tables of data. You need to define the logic of a certain application and use this knowledge in process mining. So we use a structured script to extract the data, and then are still able to mine the process.
- Interviewer How often do you experience consequences of voluminous data within the application of process mining?
- C3 Not so often for me, because I have customers which have a process which is supported by one system, however now I do have such a customer with this issue, so Sometimes, 3.
- Interviewer How often do you experience consequences of case heterogeneity within the application of process mining?
- C3 N/a.
- Interviewer How often do you experience consequences of granularity within the application of process mining?
- C3 Almost always.
- Interviewer How often do you experience consequences of concept drift within the application of process mining?
- C3 N/a.
- Interviewer How often do you experience consequences of object centric data within the application of process mining?
- C3 Almost always.
- Interviewer Are there other process characteristics which potentially have a relevant influence on the application of process mining?
- C3 The background processes of a system can produce an immense size of data which pollute the data with irrelevant information.
- Interviewer Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C3 With the systems I work with, which log all activities it is not a relevant problem. However there are systems which I worked which I worked with, which do not log all activities, this is an inherent constraint, thus relevant. How do we handle this, we collaborate with the accountant and customer how to handle this. This could be retrieved of the customer, and for the accountant it is important to discuss which impact this may have on the output of the process mining model in the audit.
- Interviewer Does incorrect data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C3 No, I work predominantly with SAP, in this system, I have not yet experienced this, and however I did work with systems which did produce incorrect data, so a relevant problem. This was due to a system error, this was corrected. So I looked at what the process should produce in the event log, and saw that some steps were not relevant.

Interviewer Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C3 No, I did not have problems with timestamps or imprecise data. Thus irrelevant.  
 Interviewer Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C3 Yes very relevant, I handle this in the same matter as voluminous data.  
 Interviewer How often do you experience missing data within the application of process mining?  
 C3 Sometimes.  
 Interviewer How often do you experience incorrect data within the application of process mining?  
 C3 Seldom.  
 Interviewer How often do you experience imprecise data within the application of process mining?  
 C3 N/a.  
 Interviewer How often do you experience irrelevant data within the application of process mining?  
 C3 Almost always.  
 Interviewer Are there other event logging issues which have a relevant influence on the application of process mining?  
 C3 That some entries in a certain table are giving information about the context of the information visualized in a table, you have to know beforehand which information this is, and otherwise you might interpret the table in an incorrect manner. So you have to know what the variable means in a certain table. You have to know what data you have, otherwise it might be dangerous with interpreting the data. You need domain knowledge of the customer to know this, of how they use the system. This is a more or less a problem with object centric data. This becomes a larger problem if you have customized systems, to know what certain variables means.

Transcript – ID = C4

Actor  
 Interviewer What is your current function?  
 C4 JR. IT Auditor  
 Interviewer In which sector do you operate?  
 C4 Financial services  
 Interviewer How many years do you have experience with process mining?  
 C4 1.5 - 2 years  
 Interviewer Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C4 If one system supports the process then this issue is not relevant, however if you have more than one system which supports the financial process, this problem becomes relevant. For example correlation becomes a problem. How do we handle this? Find the common identifier, the systems also communicate which each other, so by knowing how these systems communicate which each other, you will find the common identifier. So this could be a relevant problem if there is a relevant problem if more than one system support the system. Thus both irrelevant as relevant, depends on the situation.  
 Interviewer Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C4 Very relevant, if there is little structure in how the process as business as usual is executed. How do we handle this: For example we group the type orders, type A and B, and show the traces of these orders, then you will have a process mining model type A and one of type B. There must be some kind of structure in a process.  
 Interviewer Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C4 This is a relevant problem, because the accountant has to be able to interpret the process mining model on the correct level of detail. Therefore it is for the IT auditor important to

- have this in place and what is expected from us. How do we deal with this: We collaborate with the end-user to find the correct level of granularity in the process mining model.
- Interviewer Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C4 Irrelevant, because the process mining data will give you the answer in how the process has changed.
- Interviewer Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C4 Relevant problem. In my experience most cases where with object centric data. The disadvantage of this, is that you cannot discover the process steps. For example, if you merge the data, and you do not know of other objects which should be taken into account, you have a loss of detail, and cannot discover how the process has taken place. Therefore it is in this case important to scope the process mining assignment to the analysis which are possible.
- Interviewer How often do you experience consequences of voluminous data within the application of process mining?
- C4 Seldom
- Interviewer How often do you experience consequences of case heterogeneity within the application of process mining?
- C4 Sometimes.
- Interviewer How often do you experience consequences of granularity within the application of process mining?
- C4 Often.
- Interviewer How often do you experience consequences of concept drift within the application of process mining?
- C4 N/a.
- Interviewer How often do you experience consequences of object centric data within the application of process mining?
- C4 Often.
- Interviewer Are there other process characteristics which potentially have a relevant influence on the application of process mining?
- C4 No.
- Interviewer Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C4 Depends is the information you miss is vital to do a process mining analysis. For example if you miss the registration of goods, or if the price changes are logged. Thus this depends on the goal of the process mining analysis. In general this has not been relevant, because we could retrieve the data in most cases. I have not experienced an influence of this issue.
- Interviewer Does incorrect data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C4 Depends on which information is incorrect, depends again on the goal of the process mining analysis. I have not experienced this yet, if we found outliers, these are then explained by the customer. Thus irrelevant.
- Interviewer Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C4 Relevant, how do we handle this: First of all, we hope that the system orders the event registration in the correct order, and otherwise, we interpolate timestamps by following the process logically. However this impacts the analysis if you want to retrieve information about in which order the process steps are executed.
- Interviewer Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C4 Irrelevant, small impact on process mining analysis.
- Interviewer How often do you experience missing data within the application of process mining?
- C4 N/a.
- Interviewer How often do you experience incorrect data within the application of process mining?



C4 N/a.  
 Interviewer How often do you experience imprecise data within the application of process mining?  
 C4 Often.  
 Interviewer How often do you experience irrelevant data within the application of process mining?  
 C4 N/a.  
 Interviewer Are there other event logging issues which have a relevant influence on the application of process mining?  
 C4 No.

Transcript – ID = C5

Actor

Interviewer What is your current function?  
 C5 Senior assistant  
 Interviewer In which sector do you operate?  
 C5 Financial Services  
 Interviewer How many years do you have experience with process mining?  
 C5 Two years  
 Interviewer Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C5 Not relevant, because we now only analyze one month, following this guideline you have significantly less data in your set. On the basis of the data of this month, you decide which items in the data you want to scrutinize for the whole year.  
 Interviewer Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C5 Relevant. How do we handle this? I zoom into the level that is appropriate for human perception, and adjust this level of detail to look into the amount of transactions. Based on these amounts I am able to decide whether or not a conclusion can be made. For example, when you deal with a process which is executed for the goal of blocking purchase orders which have not enough margin. When you look into this process, a spaghetti model comes forward, with a lot of different traces. Then I will look into the amount of transactions in every trace, and only select traces which possess a significant amount of traces. IT audit assist here in grouping the data into categories of for example orders which are more easily to interpret.  
 Interviewer Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C5 Yes is relevant, it takes considerable time and effort to get insight to a model with an inappropriate level of granularity. The process mining tool, allows zooming in and out while the trustworthiness of the data stays intact. IT audit assist in this process step, to zoom into the relevant scope of the assignment and keep the trustworthiness of the data intact.  
 Interviewer Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C5 Depends on the type of organization. I did a process mining analysis in a trade organization. This is purely purchase and sell, and has less influence of law or regulatory changes. Thus, I have not experienced this in my case, but I do acknowledge this to be relevant in other types of organizations.  
 Interviewer Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C5 Relevant. If you are missing details due to object centric data, you may be prohibited to perform some kind of analysis, such as segregation of duty. I think this is the most important issue for data quality. I would go forward with the assignment, however the trustworthiness of the data becomes doubtful as many manual changes has been made. I think I will discuss this with the IT audit department, if the process mining analysis is valuable, if I can lean on it, for me to spend my time and effort on.

Interviewer	How often do you experience consequences of voluminous data within the application of process mining?
C5	N/a.
Interviewer	How often do you experience consequences of case heterogeneity within the application of process mining?
C5	2, this becomes less as we become more experienced with process mining.
Interviewer	How often do you experience consequences of granularity within the application of process mining?
C5	2
Interviewer	How often do you experience consequences of concept drift within the application of process mining?
C5	N/a.
Interviewer	How often do you experience consequences of object centric data within the application of process mining?
C5	Almost never
Interviewer	Are there other process characteristics which potentially have a relevant influence on the application of process mining?
C5	What is missing in the tool now, is that the value in currency of the transactions are missing. I am able to make better decisions on the basis of this information for example know which transactions are a risk for material misstatement.
Interviewer	Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C5	Relevant, How do I handle this? I will request all needed from the customer if I experience missing cases in my analysis. However without the issue of missing data I am able to perform a process mining analysis in a faster fashion.
Interviewer	Does incorrect data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C5	I have not yet experienced incorrect data. How do I handle this, I will check this with a walk through test, this is done with the mined process flow and the data in the source system.
Interviewer	Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C5	I have not yet experienced imprecise data.
Interviewer	Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C5	Relevant problem. How do I handle this? In the same manner that I deal with granularity and voluminous data.
Interviewer	How often do you experience missing data within the application of process mining?
C5	Almost never This is what IT audit already solves for us.
Interviewer	How often do you experience incorrect data within the application of process mining?
C5	N/a. This is what IT audit already solves for us.
Interviewer	How often do you experience imprecise data within the application of process mining?
C5	N/a. This is what IT audit already solves for us.
Interviewer	How often do you experience irrelevant data within the application of process mining?
C5	Sometimes, as on one hand you need to check all traces in the process, and on the other hand is it easy to zoom out or filter out irrelevant information.
Interviewer	Are there other event logging issues which have a relevant influence on the application of process mining?
C5	No.

Transcript – ID = C6

Actor	
Interviewer	What is your current function?
C6	Senior manager audit & assurance

Interviewer	In which sector do you operate?
C6	Financial services
Interviewer	How many years do you have experience with process mining?
C6	2 years.
Interviewer	Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C6	Voluminous data issue is relevant, so you have to connect and filter information. This is done by the IT auditors. However this is also the added value of process mining, that you have the whole population in process mining.
Interviewer	Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C6	Relevant problem, how do we handle this: We filter for example the orders in collaboration with the customer on the type of order, then you are able to have an overview which is sufficient for human perception.
Interviewer	Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C6	Relevant. To exactly know how to understand the process mining model, and what the level of detail is i.e. what is being followed in the process mining model.
Interviewer	Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C6	Irrelevant, because as an accountant I am already aware of changes in the processes which are in scope for the audit, through interviews with the customers.
Interviewer	Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C6	IT Audit makes the analysis, however from my perspective this would be a problem because it influences the trustworthiness of the process mining analysis.
Interviewer	How often do you experience consequences of voluminous data within the application of process mining?
C6	Sometimes.
Interviewer	How often do you experience consequences of case heterogeneity within the application of process mining?
C6	Often.
Interviewer	How often do you experience consequences of granularity within the application of process mining?
C6	Often.
Interviewer	How often do you experience consequences of concept drift within the application of process mining?
C6	N/a.
Interviewer	How often do you experience consequences of object centric data within the application of process mining?
C6	No comment, IT Audit handles this issue
Interviewer	Are there other process characteristics which potentially have a relevant influence on the application of process mining?
C6	Scoping is difficult, for example if you extract data for a month, then the orders in the data do not include the full process of each order, because some orders may have started previous of that month, and still are following the process in this particular month. So this is complicated.
Interviewer	Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C6	Relevant, mostly because the customer does not log everything in the system. How do I handle this, this depends on the added value of process mining in the audit, thus I collaborate with the customer about what is possible for next year.
Interviewer	Does incorrect data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C6	Irrelevant.

- Interviewer Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C6 Relevant.  
 Interviewer Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C6 Relevant, how do I handle this, I use filters to delete irrelevant data from the process mining map, however to know what is relevant or irrelevant, you need to know the meaning and context of each process step together with the customer. We take one trace and then go together through the flow of the process. This takes considerable time and effort.  
 Interviewer How often do you experience missing data within the application of process mining?  
 C6 Sometimes.  
 Interviewer How often do you experience incorrect data within the application of process mining?  
 C6 N/a.  
 Interviewer How often do you experience imprecise data within the application of process mining?  
 C6 No comment. IT Audit handles this issue.  
 Interviewer How often do you experience irrelevant data within the application of process mining?  
 C6 Almost always.  
 Interviewer Are there other event logging issues which have a relevant influence on the application of process mining?  
 C6 The amount of transactions are incorrect, these do not add up in the end, it seems that some transactions are missing in the process mining model.

Transcript – ID = C7

- Actor  
 Interviewer What is your current function?  
 C7 Internal auditor  
 Interviewer In which sector do you operate?  
 C7 Insurance  
 Interviewer How many years do you have experience with process mining?  
 C7 2 years  
 Interviewer Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C7 Very relevant. How do I handle this? I take care that I know exactly what data is in the data set, for example which identifier you have in every system. Take irrelevant data out of the data set, if needed together with the customer. So, scope on data level.  
 Interviewer Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C7 Relevant, however this depends on the question, if the side tracks are part of the process or not. If your process has a lot of side tracks, your process does not lend itself to process mining analysis, however if you want to concentrate on analyzing these sidetracks it becomes another story.  
 Interviewer Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C7 Yes it is very relevant, for example if your level of granularity is too high, then you are missing data. The consequence of this is that you do not know the context of the data. How do I handle this? I decide previous to the analysis which data is relevant for me and for the customer to answer the research question. The correct level of detail depends on the process as well as the research question.  
 Interviewer Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
 C7 Relevant, because otherwise you analyze the incorrect data. How do I handle this? Previous to the analysis, by discussing changes with the customer, know what data is in your data set. On both a functional and technical level, and which relation they have. For example there

might be a change in the functioning of the process, but this has relatively no influence on the technical side of the system. Technical changes have more influence on process mining than functional changes. So only technical changes are relevant for the application of process mining.

- Interviewer Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C7 Very relevant, unless you need information about the object, however this is not the added value of process mining. So you need to know which connection there is between the objects, if there is any. So you need to research the context of this object, and afterwards go the customer to find data about the object in question.
- Interviewer How often do you experience consequences of voluminous data within the application of process mining?
- C7 Almost always
- Interviewer How often do you experience consequences of case heterogeneity within the application of process mining?
- C7 Sometimes
- Interviewer How often do you experience consequences of granularity within the application of process mining?
- C7 Often
- Interviewer How often do you experience consequences of concept drift within the application of process mining?
- C7 Almost always
- Interviewer How often do you experience consequences of object centric data within the application of process mining?
- C7 Sometimes
- Interviewer Are there other process characteristics which potentially have a relevant influence on the application of process mining?
- C7
- Date and time of systems do not run in sync, the result of this are inconsistencies such as the order of activities which is incorrect in the data.
  - Making the process mining analysis explainable towards the customers, let them understand how process mining works.
- Interviewer Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C7 Very relevant, we are not able to analyze data which is missing. This is relevant when it concerns  
Timestamps, activities, case, this data is essential for process mining and relevant if these are missing. How do I handle this? If the data is retrievable from the customers then you could solve this in this manner, and otherwise you are not able to solve this issue.
- Interviewer Does incorrect data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C7 Is relevant, it is not a representation of reality  
How do I handle this? Every step of the dataset is tested with the customer and with the source systems. Thus including the customer in the preparation phase of process mining.
- Interviewer Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C7 Relevant. How do I handle this? Sometimes the systems indicates that the event logs do not possess timestamps, however when you take a deeper look into the system, one level lower, there is a time indication.  
Or the system takes only the date, then we have to interpolate timestamps into the data in communication with the customer.
- Interviewer Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
- C7 Irrelevant.
- Interviewer How often do you experience missing data within the application of process mining?
- C7 Almost never, because we do not apply process mining if we miss relevant data.

- Interviewer C7 How often do you experience incorrect data within the application of process mining?  
Sometimes
- Interviewer C7 How often do you experience imprecise data within the application of process mining?  
Often
- Interviewer C7 How often do you experience irrelevant data within the application of process mining?  
N/a.
- Interviewer C7 Are there other event logging issues which have a relevant influence on the application of process mining?
- Charset (character set), (set of symbols and encodings), the manner that data is stored or extracted. The charset must be correct, otherwise you will get unreadable signs in your data. This happens often in my experience.
  - Is the event log usable? For example semantics, everything is logged in 1 cell. Then you need to go back to the customer for the context of this cell.
  - Privacy; are you allowed to analyze everything? Reform of EU data protection rules are implemented in May 2018.

Transcript – ID = C8

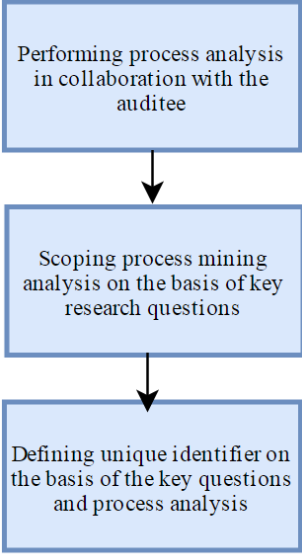
Actor

- Interviewer C8 What is your current function?  
Internal auditor
- Interviewer C8 In which sector do you operate?  
Service industry B2B
- Interviewer C8 How many years do you have experience with process mining?  
Two years, our goal is continued monitoring.
- Interviewer C8 Does voluminous data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
Relevant, how do we deal with this? We have a unique key for every case that is going through the process, even if it is going to another data warehouse, it carries the same unique key. Therefore we do not have a problem of different with different identifiers, and of correlation.
- Interviewer C8 Does case heterogeneity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
Relevant, how do we handle this? Our process is set in a way that employee are forced to follow the flow. For this reason are exceptions for us extremely interesting. There is certainly the danger to drown in the information. In order to solve this we are constantly in discussion about what is relevant for us, and what is not; where should we focus on to answer the research? Important here is to know the context of the process you are researching.
- Interviewer C8 Does granularity have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
Relevant, depends on research question, and therefore the type of analysis. If you need more than the basic information. For example who authorizes payments in the process, this is question for which more detailed information is needed.
- Interviewer C8 Does concept drift have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
Irrelevant problem for us, the changes of the process in the process mining tool appear afterwards. However, this is still not a huge problem, as we focus on the outliers which come out of the analysis. If we research these outliers, the change will become apparent.
- Interviewer C8 Does object centric data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?  
Irrelevant problem. We have two systems in our process which is relevant for the audit. These systems are connected with each other. As mentioned before, every case in our system gets a unique ID, which follows the whole process.



Interviewer	How often do you experience consequences of voluminous data within the application of process mining?
C8	Sometimes, half of the times
Interviewer	How often do you experience consequences of case heterogeneity within the application of process mining?
C8	2
Interviewer	How often do you experience consequences of granularity within the application of process mining?
C8	4
Interviewer	How often do you experience consequences of concept drift within the application of process mining?
C8	N/a.
Interviewer	How often do you experience consequences of object centric data within the application of process mining?
C8	N/a.
Interviewer	Are there other process characteristics which potentially have a relevant influence on the application of process mining?
C8	The largest problem for me are the amount of data, and if you have all the relevant data to answer your research question. We have a good insight, however we still handle this with trial and error.
Interviewer	Does missing data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C8	This problem does occur, however not a lot. How do we handle this? If we have missing data, we log a ticket, and the IT helpdesk decides on the next step to take. For example, there is a bug in the system that prevents events to be recorded, or some changes in the system has to be made in order to record the events. Irrelevant, it is daily monitored, and for this reason occurs seldom.
Interviewer	Does incorrect data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C8	Our system self-regulating so, a lot of fields are automatically filled and some fields which require details from employee are predefined by the system on how to fill these in (for example, this field requires 5 characters). Besides this are several application controls in place which prevent incorrect information in the system. So for this reason no relevant problem.
Interviewer	Does imprecise data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C8	No relevant problem for us.
Interviewer	Does irrelevant data have a relevant influence on the application of process mining? If relevant, how do you handle this issue?
C8	We retrieve our data with the aid of research questions, so we do not have irrelevant data in our process mining analysis.
Interviewer	How often do you experience missing data within the application of process mining?
C8	N/a.
Interviewer	How often do you experience incorrect data within the application of process mining?
C8	N/a.
Interviewer	How often do you experience imprecise data within the application of process mining?
C8	N/a.
Interviewer	How often do you experience irrelevant data within the application of process mining?
C8	N/a.
Interviewer	Are there other event logging issues which have a relevant influence on the application of process mining?
C8	No.

**APPENDIX E - QUESTIONNAIRE: GUIDELINE VALIDATION (21 QUESTIONS)**

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>
<p><b><i>Voluminous data and irrelevant data</i></b></p>	<p>The data extracted is from an increasing wide range of systems and/or legacy information systems and applications within a company which are difficult to handle within process mining analysis. For example: Process steps are executed in parallel in different systems, for this reason is the process mining algorithm not able to order these in the process mining model.</p> <p>Irrelevant data is defined as a process mining model that consists out of data which is not needed to answer the research question For example: Certain processes in the analysis are not needed to answer the research question</p>	<ul style="list-style-type: none"> <li>- Existing process mining algorithms are not able to cope with case volumes</li> <li>- Differences between the identifiers between different information systems</li> <li>- (Secondary) Correlation problems</li> </ul>
<p><b><i>Proposed guideline</i></b></p>		<p><i>Flow chart</i></p>
<ul style="list-style-type: none"> <li>- A process analysis will be performed in collaboration with the auditee in order to obtain context knowledge about the financial processes in scope and domain knowledge of the information systems. This process analysis entails which data is available in the recorded event logs.</li> <li>- The process analysis can also assist in solving correlation problems further down in the cleaning and enrichment phase by providing information on the context of the processes which makes it easier for the analyst to add heuristics to the data.</li> <li>- Scoping is accomplished according key research questions and the choice of identifier (the process</li> </ul>		 <pre> graph TD     A[Performing process analysis in collaboration with the auditee] --&gt; B[Scoping process mining analysis on the basis of key research questions]     B --&gt; C[Defining unique identifier on the basis of the key questions and process analysis]     </pre>



<p>instance which is followed in the process mining analysis) are made on the basis of the process analysis. This manner of scoping assists in collecting and locating only relevant data for the process mining model.</p>	
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

<b>Relevance of guideline</b>	
<b>Q1:</b>	To what extent do you find the proposed guideline to be relevant for handling this issue?
<b>0</b>	Not relevant
<b>0</b>	Somewhat relevant
<b>0</b>	Quite relevant
<b>0</b>	Highly relevant

<b>Feasibility of guideline</b>	
<b>Q2:</b>	To what extent do you agree that the proposed guideline to be easy to administer in the process mining process?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree
<b>Q3:</b>	To what extent do you agree that the proposed guideline to be interpretable?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>
<p><b>Case heterogeneity and incorrect data</b></p>	<p>Growing flexibility that a company needs to incorporate in their processes may possess a high amount of distinct scenarios which are difficult to handle within process mining analysis. For example: The process is executed in many different ways in order to react to the customer, for this reason many traces are possible in the process mining model, even more than human perception can handle.</p> <p>Incorrect data is defined as process mining data that does not correlate with how the process was executed in reality.</p>	<ul style="list-style-type: none"> <li>- Existing process mining algorithms are not able to cope with case heterogeneity</li> <li>- A high amount of scenarios/traces possible</li> </ul>
<b>Proposed guideline</b>		<b>Flowchart</b>
<p>In case the process model is difficult to interpret due to too many traces:</p> <ul style="list-style-type: none"> <li>- Each trace is visualized in an process instance graph (a graph which visualizes each trace separately), in order to make it understandable for human perception</li> <li>- Conformance checking (check between the process as described and how the process is executed) is performed in order to replay each trace and discover any anomalies</li> <li>- The traces with anomalies are studied in order to get insight in the process flow. Detected anomalies are repaired following a certain set of rules defined by applicable stakeholders.</li> </ul>		<pre> graph TD     A{Process model difficult to interpret due to too many traces?} -- Yes --&gt; B[Replaying each trace by means of conformance checking and repairing model if needed]     A -- No --&gt; C[Proceed with process mining analysis]     B --&gt; C     </pre>

<b>Relevance of guideline</b>	
<b>Q4:</b>	To what extent do you find the proposed guideline to be relevant for handling this issue?
<b>0</b>	Not relevant
<b>0</b>	Somewhat relevant
<b>0</b>	Quite relevant

<b>0</b>	Highly relevant
----------	-----------------

<b>Feasibility of guideline</b>	
<b>Q5:</b>	To what extent do you agree that the proposed guideline to be easy to administer in the process mining process?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree
<b>Q6:</b>	To what extent do you agree that the proposed guideline to be interpretable?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>
Concept drift	Dynamic processes of companies change over time which may influence the process mining model in an unforeseen manner. For example: The process is changed at some point in time due to a change in execution of the process, if the analyst is not aware of this change, the process mining model might be interpreted incorrectly.	<ul style="list-style-type: none"> <li>- Outliers due to momentary changes</li> <li>- No insight in process due to deviations in business processes which may influence event logs in an unforeseen manner</li> </ul>
<b>Proposed guideline</b>		<b>Flowchart</b>
<p>A check is performed on the process mining data in order to detect and localize any changes in the process.</p> <p>There are two approaches possible to detect and localize changes happened in the business process:</p> <ol style="list-style-type: none"> <li>3. Use statistical tests in order to detect changes in the data set by dividing a log in two subsequent event logs and see if there are changes in the traces between these two data sets.</li> <li>4. Detect any changes using metrics, such as the relation type count, which looks at the changes in relations between activities over a certain time period and from this knowledge finds any changes.</li> </ol>		<pre> graph TD     A[Detecting and localizing changes in process mining data due to process flexibility] --&gt; B{Changes present in process mining data?}     B -- Yes --&gt; C[Altering interpretation of process mining data]     C --&gt; D[Proceed with process mining analysis]     B -- No --&gt; D     </pre>

**Relevance of guideline**

<b>Q7:</b>	To what extent do you find the proposed guideline to be relevant for handling this issue?
<b>0</b>	Not relevant
<b>0</b>	Somewhat relevant
<b>0</b>	Quite relevant
<b>0</b>	Highly relevant

**Feasibility of guideline**

<b>Q8:</b>	To what extent do you agree that the proposed guideline to be easy to administer in the process mining process?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree
<b>Q9:</b>	To what extent do you agree that the proposed guideline to be interpretable?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree

DATA QUALITY WITHIN PROCESS MINING IN THE AUDITING CONTEXT

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>
<b>Object centric data</b>	Systems do not record event logs along a process but according to an object. For example: The event logs are not recorded following a process flow, but following objects such as invoices or orders, and these need to be merged in order to get insight into the whole process flow.	- structuring and merging event data takes a lot of time and labor as the analyst has to merge the data of all objects such as invoices and orders in an assumed process flow
<b>Proposed guideline</b>		<b>Flowchart</b>
<p>The application of process mining with a workflow oriented data mining tool offers limited functionality when applied object centric data, for this reason, already in the scoping process one has to decide if process mining adds value in answering the key research questions.</p> <p>When proceeding with the process mining analysis, the employment of standard extraction guidelines (scripts) are suggested in case of system(s) and/or application(s) in scope with non-customized core functionalities.</p>		<pre> graph TD     D1{Object centric data present within system(s) and/or applications in scope?}     D2{Does process mining add value in answering key research questions?}     D3{System(s) and/or application(s) in scope with non-customized core functionalities}     B1[Proceed with process mining analysis]     B2[Creation standard extraction script]     B3[Proceed with process mining analysis]     B4[Stop process mining assignment]     E([End])      D1 -- No --&gt; B1     D1 -- Yes --&gt; D2     D2 -- No --&gt; B4     D2 -- Yes --&gt; D3     D3 -- Yes --&gt; B2     D3 -- No --&gt; B3     B2 --&gt; B3     B4 --&gt; E     </pre>

<b>Relevance of guideline</b>	
<b>Q10:</b>	To what extent do you find the proposed guideline to be relevant for handling this issue?
<b>0</b>	Not relevant
<b>0</b>	Somewhat relevant
<b>0</b>	Quite relevant
<b>0</b>	Highly relevant

<b>Feasibility of guideline</b>	
<b>Q11:</b>	To what extent do you agree that the proposed guideline to be easy to administer in the process mining process?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree
<b>Q12:</b>	To what extent do you agree that the proposed guideline to be interpretable?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree

<i>Issue</i>	<i>Explanation</i>	<i>Consequences</i>
<b><i>Granularity, imprecise data</i></b>	<p>The data extracted from contemporary, or legacy information systems and applications within a company possess a mixed level or low level granularity, in other words detail, which are difficult to handle within process mining analysis. For example: The process model has a mixed level of detail between certain systems which support the financial processes in scope.</p> <p>Imprecise data is defined as data which is not at the correct level of detail of the (end) user. For example: Timestamps only record the day of the recorded process step</p>	<ul style="list-style-type: none"> <li>- Low level granularity event logs are difficult to comprehend for human perception</li> <li>- Issues in finding the right level granularity for the (end)user of the analysis</li> </ul>
<b><i>Proposed guideline</i></b>	<b><i>Flowchart</i></b>	
<p>The desired level of granularity is defined based on the key research questions concerning the financial processes in scope, in order to find the right level of granularity for the end user of the analysis. In order to apply to correct level of granularity in the process mining model a plan of approach (PoA) is defined together with all relevant stakeholders, especially the end user.</p>	<pre> graph TD     A[Defining desired level of detail level based on key research questions] --&gt; B[Defining PoA in order to apply desired level of detail]             </pre>	

<b>Relevance of guideline</b>	
<b>Q13:</b>	To what extent do you find the proposed guideline to be relevant for handling this issue?
<b>0</b>	Not relevant
<b>0</b>	Somewhat relevant
<b>0</b>	Quite relevant
<b>0</b>	Highly relevant

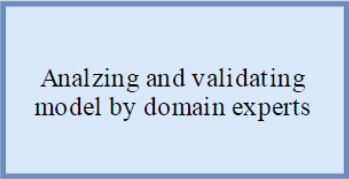


<b>Feasibility of guideline</b>	
<b>Q14:</b>	To what extent do you agree that the proposed guideline is easy to administer in the process mining process?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree
<b>Q15:</b>	To what extent do you agree that the proposed guideline is interpretable?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree

<i>Issue</i>	<i>Explanation</i>
Missing data	Missing data is defined as mandatory data which is not present in the process mining model, but is needed for the process mining analysis. For example: A process step took place in reality but was not recorded in the event log.
<i>Proposed guideline</i>	<i>Flowchart</i>
In case of missing data after data extraction, domain knowledge is often retrievable from the customer.	<pre> graph TD     A{Missing data in the process mining model?} -- Yes --&gt; B[Retrieving domain knowledge from customer]     B --&gt; C[Proceed with process mining analysis]     A -- No --&gt; C     </pre>

<b>Relevance of guideline</b>	
<b>Q16:</b>	To what extent do you find the proposed guideline to be relevant for handling this issue?
<b>0</b>	Not relevant
<b>0</b>	Somewhat relevant
<b>0</b>	Quite relevant
<b>0</b>	Highly relevant

<b>Feasibility of guideline</b>	
<b>Q17:</b>	To what extent do you agree that the proposed guideline to be easy to administer in the process mining process?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree
<b>Q18:</b>	To what extent do you agree that the proposed guideline to be interpretable?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree

<i>Issue</i>	<i>Explanation</i>	
Incorrect data	<p><i>Process mining data does not correlate with how the process was executed in reality.</i></p> <p><i>For example: Due to a system error, recorded process steps are assigned to the wrong process.</i></p>	
<i>Proposed guideline</i>	<i>Flowchart</i>	
<p>Domain knowledge of the auditee is needed to uncover situations where incorrect data occurs, by means of validating the model.</p> <p>Outliers in the process mining model are recognized as possible manifestations of incorrect data, these should be taken as warnings for incorrect data for the process mining analyst.</p>		

<b>Relevance of guideline</b>	
<b>Q19:</b>	To what extent do you find the proposed guideline to be relevant for handling this issue?
<b>0</b>	Not relevant
<b>0</b>	Somewhat relevant
<b>0</b>	Quite relevant
<b>0</b>	Highly relevant

<b>Feasibility of guideline</b>	
<b>Q20:</b>	To what extent do you agree that the proposed guideline to be easy to administer in the process mining process?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree
<b>Q21:</b>	To what extent do you agree that the proposed guideline to be interpretable?
<b>0</b>	Strongly disagree
<b>0</b>	Disagree
<b>0</b>	Agree
<b>0</b>	Strongly agree

## APPENDIX F – RESULTS VALIDATION

The following tables contain the results of the personally administered questionnaire.

### Relevancy

**Table 14 Results: Relevancy.**

Questionnaire question number	V1	V2	V3	Number in agreement	Item CVI
1	X	X	X	3	1.00
4	-	X	X	2	0.67
7	X	X	X	3	1.00
10	X	X	-	2	0.67
13	X	X	X	3	1.00
16	X	X	-	2	0.67
19	X	X	X	3	1.00

### Feasibility

**Table 15 Results: easy to administer.**

Questionnaire question number	V1	V2	V3	Number in agreement	Item CVI
2	X	X	X	3	1.00
5	-	X	X	2	0.67
8	X	X	X	3	1.00
11	-	X	-	1	0.33
14	X	-	X	2	0.67
17	-	X	X	2	0.67
20	X	X	X	3	1.00

**Table 16 Results: easy to administer.**

Questionnaire question number	V1	V2	V3	Number in agreement	Item CVI
3	X	X	X	3	1.00
6	X	X	X	3	1.00
9	X	X	X	3	1.00
12	X	X	X	3	1.00
15	X	X	X	3	1.00
18	X	X	X	3	1.00
2	X	X	X	3	1.00