Applying Process Mining to Analyse Business Process Performance in the Physical Asset Management Environment

by

Berno Theo Greyling

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Engineering Management in the Faculty of Engineering at Stellenbosch University

Department of Mechanical and Mechatronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Dr. J.L. Jooste

December 2015

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: ....
BT. Greyling

Date: December 2015

Abstract

Applying Process Mining to Analyse Business Process Performance in the Physical Asset Management Environment

BT. Greyling

Department of Mechanical and Mechatronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MScEng (EngMan)

December 2015

Physical Asset Management (PAM) has become a vital part of asset intensive organisations in recent years. The rise of standards associated with Asset Management (AM) and the competitive nature of large organisations has resulted in a demand for performance improvements in terms of the value created by physical assets. This thesis addresses this need by calling on an interdisciplinary approach to help improve the processes that support these physical assets within an organisation's PAM strategy execution environment. Improvement is facilitated by the use of Business Process Management (BPM) and Workflow Management (WfM) principles in conjunction with process mining. While BPM and WfM provide the tools, process mining generates the metrics and models to be improved.

This thesis starts by examining the literature of PAM and how processes form part of the enablers that support execution functions of a PAM strategy. Linkage between these enablers and physical assets creating value is then examined. The review finds that there is a lack of structured and mathematical analysis involved with PAM processes as they are executed in the real world. This lack results from the inability to monitor exactly how work is being done in relation to how it was planned. BPM and WfM are discussed to construct a theoretical basis on which process improvement can occur while process mining addresses the lack of being able to monitor real world activity. This thesis continues to develop a methodology that aims to construct a framework by which process mining can first be applied to a practical environment, followed by principles inherent in BPM and WfM. As BPM, WfM and process mining entail numerous different analytical tools, the most applicable of these techniques are chosen for the PAM environment.

The developed methodology is applied to a case study where a maintenance process is investigated to determine the validity of the methodology. By applying the process mining methodology, important performance attributes of PAM process are identified and highlighted. Ultimately, application of the methodology is deemed a viable option to incorporate within a PAM environment to improve supporting processes. Future iterations based on this research can investigate improvements to this thesis and build on the results to improve PAM processes and incorporate on-going process mining within the PAM strategy.

Uittreksel

Toepassing van Proses Ontginning om Besigheids Proses Prestasie in die Fisiese Batebestuur Omgewing te Ontleed

("Applying Process Mining to Analyse Business Process Performance in the Physical Asset Management Environment")

BT. Greyling

Departement Meganiese en Megatroniese Ingenieurswese,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MScIng (IngBes)

Desember 2015

Fisiese batebestuur (FB) het 'n belangrike deel van bate intensiewe organisasies in afgelope jare geraak. Die opkoms van standaarde wat verband hou met batebestuur en die mededingende aard van groot organisasies het gelei tot 'n aanvraag vir prestasie verbeteringe in terme van waarde geskep deur fisiese bates. Hierdie tesis spreek hierdie behoefte aan deur 'n beroep te maak op 'n interdissiplinêre benadering met betrekking tot optimalisering van prosesse wat fisiese bates in die batebestuur strategie omgewing ondersteun. Optimalisering word vergemaklik deur gebruik van Besigheids Proses Bestuur (BPB) en Werkvloei Bestuur (WvB) beginsels in samewerking met proses ontginning. Terwyl BPB en WvB die gereedskap verskaf vir optimalisering, genereer proses ontginning statistieke en modelle vir optimalisering.

Hierdie tesis begin deur die literatuur van FB te ondersoek en hoe prosesse deel vorm van uitvoering funksies van 'n batebestuur strategie. Die verband tussen hierdie ondersteunende prosesse en waardeskepping van fisiese bates

word dan ondersoek. Literatuur hersiening bevind dat daar 'n gebrek aan gestruktureerde en wiskundige analise betrokke is by prosesse soos dit uitgevoer word in die werklike wêreld. Hierdie gebrek is 'n oorsaak van 'n onvermoë om werkvloei te monitor presies hoe dit gedoen word met betrekking tot hoe dit beplan is. BPB en WvB word dan bespreek om 'n teoretiese basis te bou waarop optimalisering kan plaasvind terwyl proses ontginning die gebrek aan die onvermoë om werklike wêreld aktiwiteit te monitor aanspreek. Hierdie tesis gaan dan voort om 'n metode aan mekaar te stel wat daarop gemik is om 'n raamwerk te vorm waarmee proses ontginning eerste toegepas kan word in 'n praktiese omgewing, gevolg deur toepassing van beginsels geaard aan BPB en WvB. BPB, WvB en proses ontginning behels talle verskillende analitiese gereedskap, die mees toepaslike van hierdie tegnieke work dan gekies vir die FB omgewing.

Die ontwikkelde metode word dan toegepas op 'n gevallestudie waar 'n onderhoud proses ondersoek word om die geldigheid van die metodologie te bepaal. Deur die toepassing van die proses ontginning metode is belangrike aspekte van 'n proses geïdentifiseer en uitgelig. Uiteindelik is die toepassing van die metode geag as 'n lewensvatbare opsie om te inkorporeer in 'n FB omgewing om ondersteunende prosesse te optimaliseer. Toekomstige iterasies wat op hierdie navorsing gebaseer is kan verbeteringe aan hierdie tesis ondersoek en bou op die resultate om FB prosesse te verbeter en deurlopende proses ontginning binne die FB strategie te inkorporeer.

Acknowledgements

I would like to express my sincerest appreciation and gratitude to the following people and organisations:

- My supervisor, Dr JL Jooste for his support, input and patience.
- My family for their continuous encouragement throughout the study.
- Everyone in the "penthouse" who made the two years more pleasant by creating a supportive working environment.
- Anglo American for their financial support.

The Author
September, 2015

Dedications

This thesis is dedicated to my parents, Theo and Tersia, for their continuous support and motivation.

Contents

List of Figures

List of Tables

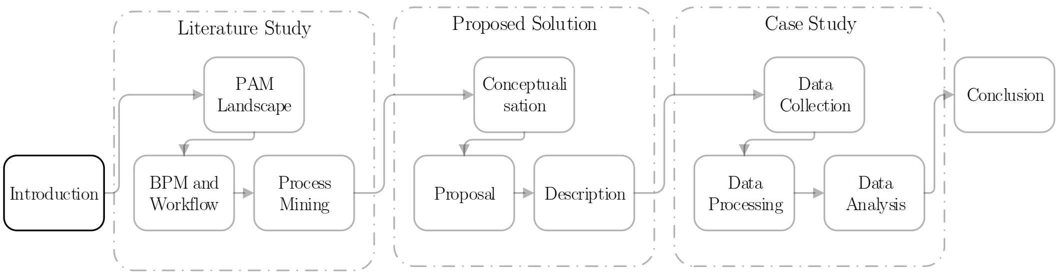
List of Abbreviations

AM	Asset Management
AMS	Asset Management System
BI	Business Intelligence
BM	Business Model
BPI	Business Process Intelligence
BPM	Business Process Management
BP	Business Process
BSI	British Standards Institution
CPN	Coloured Petri-nets
EAM	Enterprise Asset Management
EAMS	Engineering Asset Management System
ERP	Enterprise Resource Planning
ETL	Extract Transform Load
IAM	Institute of Asset Management
ISO	International Standards Organisation
IS	Information System
IT	Information Technology
IvM	Inductive Visual Miner
KPI	Key Performance Indicator
MfMS	Workflow Management System
MXML	Mining Extensible Markup Language
PAM	Physical Asset Management
PAMS	Physical Asset Management Strategy
PI	Performance Indicator
PPA	Process Performance Analysis
ProM	Process Mining
SQL	Structured Query Language
VV&A	Verification, Validation and Accreditation
WfM	Workflow Management
WM	Work Management

Chapter 1

Introduction

This chapter sets out to be the starting point for the reader to understand the basic concepts used within this thesis. It presents Physical Asset Management (PAM) and the area in which improvements can be made. The problems are presented and discussed with the aim of giving a background of the focus area in order to finally present the problem statement. The application methodology is then presented which aims to solve the problem.



Chapter Outcomes
•Develop thesis domain understanding.
•Present problem statement which warrants investigation.
•Present the research design and methodology which aims to achieve the goals presented.

1.1 Physical Asset Management

Asset Management (AM) was a term solely used in the financial realm where it was attributed to investment portfolios. According to Brown and Humphrey (2005), the fundamental idea behind AM is the balancing between risk and return. The context in which AM is used here, is a strategic approach towards operations, management and maintenance to optimise the allocation of the organisation's resources (Flintsch and Bryant, 2006*a*). As AM is now placed in a new domain, the definition of an asset also changes to (Amadi-Echendu, 2004):

> An entity which has the capability to create and sustain value while in current use, or that which appreciates in value because of perceived capability to create value in future use.

Amadi-Echendu (2004) states that the first part of the definition lies more with assets which are physical in nature while the second part ties in intangible assets. As organisation are growing in size and competition rises amongst them, the expected performance of these organisation rise to where AM has become a normal practice as stated by Woodhouse (2014). Amadi-Echendu, Willett, Brown, Hope, Lee, Mathew, Vyas and Yang (2010) moves the discussion in a domain where assets are not only physical in nature, hence the term Physical Asset Management (PAM), but are the base of engineering operations. These "engineering objects" are the supporting and value creating objects in organisations where the value of the entire organisation is dependent on the performance of these physical assets. This then forms the basis for the term Engineering Asset Management (EAM). In some sense, organisations still view PAM as an approach to only maintain physical assets when, as discussed by Amadi-Echendu *et al.* (2010), it is an interdisciplinary and holistic approach to the care and support of physical assets. PAM has moved from only being a maintenance strategy, even though it still encompasses that view to a certain degree (Mobley, 2002), to where it ties in with supporting operations.

As stated by Lingamaneni (2010) the state in which AM research finds itself now can be categorised by the following applications:

1. Current processes applied to current techniques;
2. Current processes applied to new techniques;

3. Innovative processes applied to current techniques;
4. Innovative processes applied to innovative techniques.

By a large degree, research has mainly been done in the field of current processes and current techniques as these have the easiest carry over to industry. Industry in this sense is the main driver in development in these areas. Attention has been given to innovative processes and techniques but as implementation is not always cost effective at first, research and therefore innovation has been slow. This is however not an indication of the importance of this type of research as this is the research which will ultimately bring the biggest contributions to industry (Lingamaneni, 2010).

As PAM gained traction in industry the British Standards Institution (BSI), alongside the Institute of Asset Management (IAM) and industry partners, published a set of specifications as PAS 55 (BSI, 2008). These specifications were published in 2004 and were then revised in 2008 (Woodhouse, 2014). As with many standardisation exercises, these set of specifications helped raise business performance in a strategic and consistent manner.

The PAS 55 literature can be identified to have two parts. PAS 55-1 entails the specifications for the management of physical assets and then part two, PAS 55-2, entails the guidelines for the application thereof. It is important to note that PAS 55 does not only dictate the management for physical assets but also recognises information, financial, intangible and human assets. The synergy of these assets will then, according to BSI (2008), deliver improved physical asset performance. The PAS 55 literature also aims to align an organisation's asset management goals with its own strategic and business plan.

As the implementation of PAS 55 was deemed a success, even on an international scale, there was a need for a set of standards that carries the International Standards Organisation's (ISO) branding. With that, PAS 55 was put up as the basis for an international standard from which the set of ISO 55000 standards were developed and published in January 2014 (BSI, 2014). As stated by Botha (2015), ISO 55000 does however not bring an enormous amount of additional knowledge to the AM body of knowledge. This is mainly because ISO 55000 is a formalisation of PAS 55 on an international scale. Van den Honert, Schoeman and Vlok (2013) does however indicate that ISO 55000 is somewhat an improvement on PAS 55 where information management is concerned as it includes principles of documentation control. It goes on to state that ISO 55000 includes examples of where proactive monitoring is to be included, which PAS 55 does not.

Looking at Figure 1.1 it can be seen that at the foundation of the Asset Management System (AMS), as presented by BSI (2008), lies enablers and controls. IBM (2009) lists these enablers as:

- Strategic planning, development, scheduling and resource allocation;
- Asset maintenance and configuration;
- Spare parts management;
- Maintenance and inspection of systems and equipment; and
- Risk management.

As all of these enablers are the processes by which the physical assets are supported in the organisation, it is necessary to improve not only what is being done to aid assets in creating value but also to improve the way this is being done. Seeing that these processes are at the foundation of the AMS, the improvement of these processes will cause an improvement in the value the physical assets can create. Even though this seems like a fundamental concept to AM, the literature lacks the inclusion of this aspect of AM.

Considering the implementation of PAM initiatives which already introduces challenges alongside the already common resistance to change as highlighted by Mitchell, Hickman and Amadi-Echendu (2007), there is a need to utilise existing resources to aid in performance improvement. This presents itself as an opportunity for the already vast amounts of data currently captured by organisations as part of an AM strategy. This also concurs with the BSI (2014) requirements as data, an intangible asset, is being utilised to extract extra value for the organisation. A readily available data source in asset information systems is the transactional data captured as business processes complete tasks. These business processes include processes that ultimately support

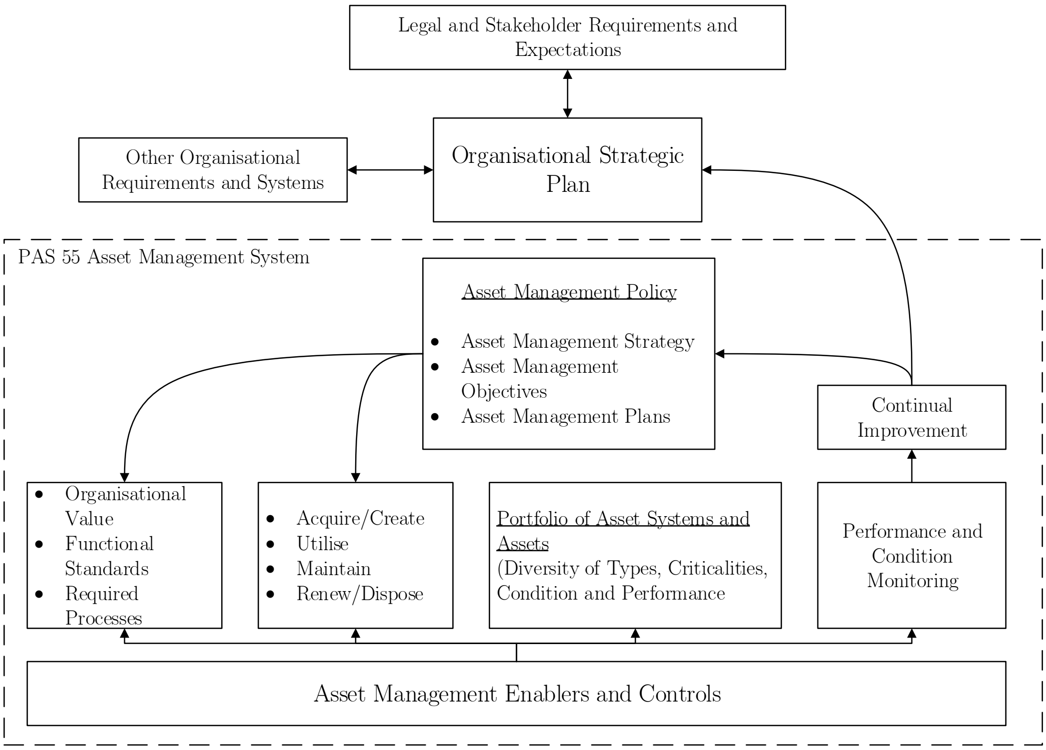


Figure 1.1: PAS 55 asset management system.
Adapted from BSI (2008)

physical asset performance and therefore forms part of the AMS as an enabler.

1.2 Business Process Management and Process mining

As presented by Frolov, Ma, Sun and Bandara (2010), PAM is fundamentally a complex business process. This concept entails that the same principles used to improve business processes within their respective field might be able to aid in the performance of PAM enabling processes in terms of an interdisciplinary approach. The application of this will then focus on processes such as maintenance, spare parts management and asset acquisition where there are numerous interacting tasks which need to be performed to reach a goal or enable the asset to create value.

Within the field of Business Process Management (BPM) there are techniques and methodologies which aim to improve business processes, although these processes have no explicit connection with PAM (Vasconcelos, Caetano, Neves, Sinogas, Mendes and Tribolet, 2001). This includes workflow techniques governing how to do work. These disciplines are often applied to processes with specific characteristics including the following (Georgakopoulos, Hornick and Sheth, 1995):

- Ad-hoc workflows: Where the process revolves around the coordination of people and there is no set pattern. This includes general office processes (i.e. product documentation, sales proposals);
- Administrative workflows: These process are set, predictable and repetitive. They often include simple tasks with little coordination such as ordering parts or reporting;
- Production workflows: Like administrative workflows, these processes are often repetitive and predictable. The difference is however that these processes often require the coordination of multiple information systems and are complex in nature. There is also some decision making logic involved.

As BPM and workflow technology implementation has been successful in these kinds of processes, the possibility exists that the same set of principles can be applied to PAM processes. This would then entail looking at business processes as enablers in the overall PAM strategy, mapping these processes and then evaluating them with the same tool-kit used in BPM. The processes considered can in this case be processes involving the maintenance of physical assets where there is a set of activities and people involved in performing them. This process can then be improved, in principle, to increase the uptime of the physical assets and therefore, increase the amount of value it can create for the organisation.

Koronios, Lin and Gao (2005) discussed the issue of data quality within PAM as a crucial aspect to look at. As the data gathered in a lot of these cases for decision making on higher level which influence outcomes across the entire organisation, the focus should not only lie with gathering the required data but also to obtain data with the necessary quality level. In this case, process mining as discussed in Van der Aalst (2011), allows the implementation of

BPM in the PAM environment by gathering data from already implemented PAM or Enterprise Resource Planning (ERP) system. This not only makes it attractive in enabling the modelling of processes in the organisation with relative ease, but also complies with the requirements of the data being of high quality and not requiring the implementation of new systems. This is important in the sense that organisations already need to deal with resistance to change in terms of applying PAM principles and do not require another system implementation to increase resistance.

Process mining gathers data from the organisation ERP system to build a historical log or transaction log that can be used to rebuild the processes as they happened. This enables the evaluation, not of the planned business process, but of the process as it is happening in reality. Process mining enables the analysis of different parts of the process such as waiting time or tasks completion times which can then be used in the BPM life-cycle where business processes are evaluated and redesigned to meet the organisation needs for the future (Van der Aalst, Reijers, Weijters, van Dongen, De Medeiros, Song and Verbeek, 2007*a*). Another aspect that makes this attractive to processes in the PAM environment is that it does not require previously modelled or planned processes to be effective. It can take the as-is processes and enable improvement within the PAM system. The results of these improvement efforts are then quantifiable in terms of improved asset value creation and not only process completion times.

Process mining alongside BPM forms part of an ever increasingly important PAM strategy where the results surpass that which come directly after initial implementation. This enables a PAM strategy that is progressive in the value that it can return to the organisation. This thesis aims to investigate the idea that process mining within a PAM environment can be used to bridge the current gaps in ISO 55000 by mapping the processes which enable value.

1.3 Problem Statement

With the introduction of ISO 55000 in 2014, there remains a lack of clarity concerning value-enabling processes within the PAM environment. This is especially true if it is considered that ISO 55000 does not add significant insights after the introduction of PAS 55, which was published in 2004. The PAM literature mainly focusses on what needs to be done and not as much how these objectives are to be accomplished. While this enables the use of organisational specific actions to achieve goals, it does exclude a basic understanding of how these actions are to be executed. Enabling processes allow physical assets within the organisation to create value and it is therefore imperative that improvements occur not only to the asset itself but also to the supporting processes. ISO 55000 also does not address best practices within an asset's life cycle especially when looking at intangible assets and the conversion of these assets into deliverables.

The importance of PAM within large organisation calls for a more interdisciplinary approach, drawing from human factors, strategy and in this case, business process management. This is contradictory to what is currently happening in industry where maintenance forms the basis of PAM. The need has thus given rise to a holistic approach that calls for different approaches to improve the PAM foundation.

This thesis focuses on investigating the potential of process mining to map PAM processes in support of BPM. To achieve this, process mining will be investigated as it is currently being used and then applied to the PAM environment to asses the viability within PAM. The formulation of the research question can thus be done as follows:

> *Is it possible to use process mining within the PAM strategy to improve physical asset value creation?*

1.4 Research Objectives

The objectives presented here aim to solve the problem presented within the problem statement. The main problem is broken down into smaller tasks that collectively support the solution. These smaller tasks collectively make the task of answering the research question more manageable. The tasks can be categorised as follows:

1. Gain a full understanding of the relevant topics within PAM, BPM and process mining. Master the application area within PAM while extruding the core influential areas.
2. Gain insight into BPM and process mining to understand the supporting principles and the mathematical models which will be used to realise improvements within PAM.
3. Determine the validity of using BPM and process mining techniques within the PAM environment and assess obtainable outcomes.
4. Select appropriate process mining techniques for use within an application methodology.
5. Align process mining practices with the AM strategy and position the analysis of processes within the strategy to enable improved execution.
6. Implement the application methodology within a PAM environment with the goal to not only test the practical aspects apply also to allow the validation of the application methodology.

1.5 Research Design and Methodology

The following section presents an overview of the research design which governs the research approach used in this thesis and presents the methodology that aims to connect the activities with the objectives.

1.5.1 Research Design

Gregg, Kulkarni and Vinzé (2001) argues that within the Information Systems (IS) discipline, there is a focus on the development and use of software technology to meet the needs set out by businesses. This then places software engineering at the core of any research which takes place within the IS domain.

The research phases and approach in this thesis resembles the paradigms which exist in a strict IS research domain. The paradigms in question here refer to the characterisation which comes with the behavioural sciences and design sciences. Behavioural science aims to set theories in place which propose a model able to explain and predict the behaviour of personnel and the surrounding organisation. The design science aims to introduce innovative means to extend the capabilities of people and organisations. The critical aspect within the design science paradigm is that it requires a deep understanding of the problem at hand to be able to implement a solution to a given problem (Von Alan, March, Park and Ram, 2004).

Hevner (2007) presents a three cycle view of design science research where it is presented as three independent activity cycles. The first of these activities is called the *relevance cycle*. This cycles starts the research endeavour by identifying the requirement in terms of the research problem and then goes on to identify the aspects by which the evaluation of the research results will be judged and deemed as acceptable. The validation in this case needs to be carefully considered as the improvement, if any, needs to be measurable for it to be considered successful. The next cycle for consideration is the *rigor cycle*. With this, the research implements theories and engineering methods as a foundation for application in the domain in question. This cycle ensures that the proposed solution is truly innovative and uses the vast knowledge base available.

At the centre of both of these cycles, the *design cycle* is placed to go back and forth between identifying artefacts (constructs, models, methods and instantiations) as a proposed solution and the research processes. This then forms part of an iterative process to refine the proposed solution. This iterative approach starts by identifying first generation solutions and then evaluates alternatives against each other until the desired design has been achieved.

Creswell (2013) clearly distinguishes between three categories of research design. The first of which, *qualitative*, has a more literary approach while the second category, *quantitative*, incorporates a more empirical approach. These methods exist on a continuum rather than being discrete options Newman and Benz (1998). The third category is thus a mixture between the two, or simply, *mixed methods*.

In this study, a qualitative approach is used with regard to the subject of PAM and the relation of BPM principles to the improvement of PAM. The application methodology of process mining however uses a quantitative approach. This applies to both the gathering of data base entries to build event logs and the analysis thereof. It is however necessary for validation purposes beyond quantitative improvement to include qualitative opinions with regards to the viability of process mining to the current PAM implementation in industry. This thesis is therefore based on a mixed method approach.

Creswell (2013) continues to define four philosophical world views namely post-positivism, constructivism, advocacy, and pragmatism. Considering the research design of this thesis, post-positivism best describes the scientific approach used. This world view contains elements of determination, reductionism, empirical observation and theory verification.

1.5.2 Research Methodology

In order to achieve the goals set out by the research objective in Section 1.4 and answering the research question in Section 1.3 the following research methodology will be implemented. First, a literature study is done in order to understand the problem and all underlying concepts. This will allow the selection of the most viable proposed solution by having the correct perspective and understanding of the problem environment. process mining as a proposed solution will then be discussed in depth by considering an application methodology that aims to answer the research question. This will then be followed by the application of the proposed solution to a case study. The results will then be validated while considering the proposed solution as valid or not.

1.6 Thesis Structure

The flow of this thesis is put together in such a way to enable the reader to follow the research process and thought process. The layout that follows complements the methodology in Section 1.5 above and aims to complete all the research objectives set out in Section 1.4. The manner in which the thesis structure links up with the objectives is shown in Table 1.1.

Table 1.1: Linkage between thesis structure and objectives.

Chapter	Objectives
Chapter 2: Literature Study	1,2,3
Chapter 3: The Process Mining Application Methodology and Requirements	3,4,5
Chapter 4: Case Study	4,5,6
Chapter 5: Closure	6

Chapter 2: Literature Study

The literature study starts with a discussion on the landscape of PAM. The fundamentals are discussed with the identification of where the focus area will be. The discussion then continues with AM Strategy and how information and the AM information system plays a role in different levels of management. BPM is then discussed in conjunction with Workflow management. Here the principles are extracted for possible use in the AM environment after which process mining is discussed as a possible analysis tool. Theory and analysis techniques are researched for use in the application methodology which follows.

Chapter 3: The Process Mining Application Methodology and Requirements

In this chapter, the application methodology is presented with focus on the PAM environment. Standards and general practices in process mining are discussed and used to position process mining to be applicable within PAM. From the techniques discussed in the literature study, the most appropriate ones are selected which will form part of the analysis toolkit with an applicable scope.

Chapter 4: Case Study

The implementation of the application methodology is tested on a case study from the petrochemical industry. The case study is presented in order to understand the background and how it complements the purposes of this thesis. The aim of the case study is to validate the applicability of the process mining application methodology in the PAM environment.

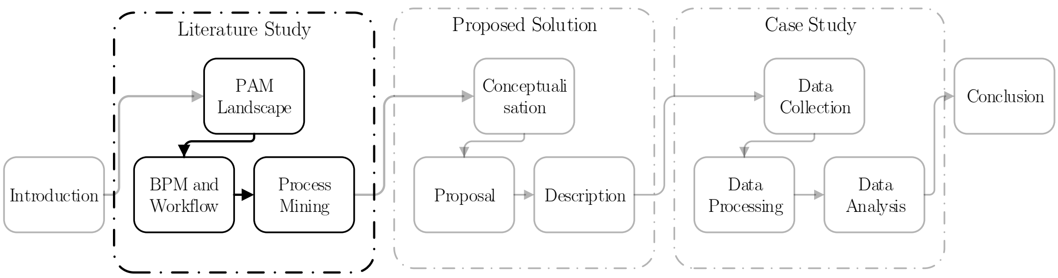
Chapter 5: Closure

In conclusion, the results from the case study are discussed with the focus on answering the research question. Limitations of the study will be discussed as this draws a framework for the discussions to follow. Recommendations based on the results of the case study are then presented in an effort to finally validate the application area of process mining.

Chapter 2

Literature Study

This chapter contextualises how business process are inherent within Physical Asset Management (PAM) and how they play a vital role in improving PAM in practice. A PAM landscape will be presented to give a holistic understanding while focussing on strategic asset management decisions, the processes that support them and then ultimately, the information system from which knowledge is based within the organisation. Business processes and workflow will then be discussed which leads to an overview of how the organisation's information system is used to improve the processes that support the asset management system.



Chapter Outcomes
•Understand the PAM landscape.
•Apprehend the area of influence and how processes exist within PAM.
•Gain insight into how processes support physical asset value creation.
•Comprehend BPM and workflow and their capacity to support PAM.
•Realise the improvement process mining is able to cultivate in its application.

2.1 Physical Asset Management

With the recent increase in the size of "asset-centric" organisations, as referred to by Mohseni (2003), there has been a need for improved efficiency and performance within organisations. The pressure not only comes from competition between organisations but also as investors and stakeholders demand a higher return on their investments. Asset Management (AM) is thus an attempt to understand the risk-reward trade-off to minimise cost and improve the performance of the organisation's assets better. With a contextual focus on the engineering environment, Davis (2007) defines AM as:

> A continuous process-improvement strategy for improving the availability, safety, reliability, and longevity of plant assets, i.e., systems, facilities, equipment and processes.

It is important to note that AM in a more general sense relates strongly to a financial and corporate management environment as demonstrated in Mitchell *et al.* (2007). Each field in which AM operates has a clear definition to make AM its own. It is for this reason that it is important to have a AM definition that relates directly to the engineering environment. Amadi-Echendu *et al.* (2010) conceptualises Engineering Asset Management (EAM) as "the total management of physical, as opposed to financial, assets. "Even though this definition excludes financial assets, the physical assets in this definition have an inherent financial property that warrants management. It is also from this definition where the logic arises to refer to AM in asset-centric organisations as Physical Asset Management (PAM).

Figure 2.1 can be interpreted in such a manner as to say that engineering assets form the foundation of what an organisation is ultimately out to achieve. This foundation consists of equipment, systems, facilities and processes and allow the organisation to create value. Considering the engineering assets as a fundamental part of the organisation's performance, it can be said that the information implementation at the most basic level generates the decision making data that, in less detail, travels up the pyramid and directly influences the decision at a managerial level. It is in the same manner that Amadi-Echendu *et al.* (2010) argues the supporting role of PAM. The organisation's information system thus lies at the core of PAM and enables the decision-making on all levels. The importance of PAM within these large organisation has called for a more interdisciplinary approach, drawing from human factors, strategy and in this case, process management. This argument is supported by Wittwer, Bittner and Switzer (2002). This is contradictory to what is currently happening in industry where maintenance forms the basis of AM. Amadi-Echendu (2004) continues this argument and states that it is thus necessary to expand the practical AM approach beyond maintenance. The need has thus given rise to a holistic approach that calls for different approaches to improve the AM foundation.

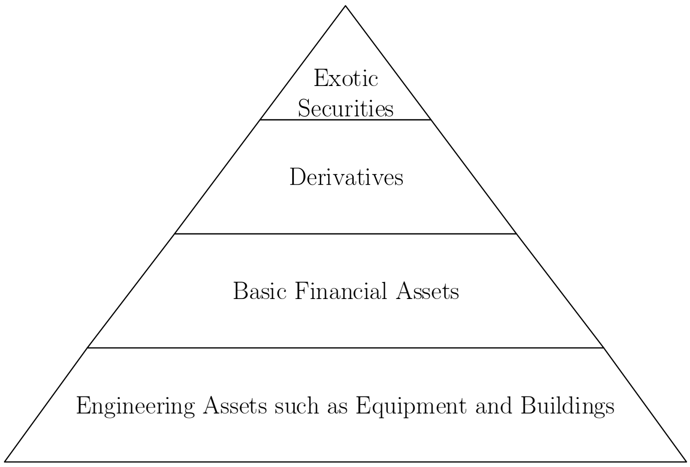


Figure 2.1: The nature and placement of engineering assets.
Adapted from Amadi-Echendu et al. *(2010)*

In today's large organisations, assets have an important part to play in how these organisations achieve success as argued by Woodward (1997). PAM owes its rising importance not only to the supporting role it plays in the care and improvement of the assets within an organisation but also the demanding regulations being set forth by the regulating bodies. The British Standards Institution (BSI) with PAS 55 (BSI, 2008) and the more recent ISO 55000 set of standards by the International Standards Organisation (ISO) (BSI, 2014) are examples of such regulating bodies. These standards give the definition for PAM as:

> systematic and coordinated activities and practices through which an organization optimally and sustainably manages its assets and asset systems, their associated performance, risks and ex-

> penditures over their life cycles for the purpose of achieving its organizational strategic plan.

Madu (2000) supports the argument that a successful and holistic approach to asset management requires the facilitation of IT systems. These IT systems allow the effective use of assets for an organisation's competitive advantage.

An asset in a more general sense applies to an ever-evolving organisation specific register. An "asset" is defined by PAS (2004) as "plant, machinery, property, building, vehicles and other items that have a distinct value to the organisation" . This definition is clearly more inclined to be used within the PAM environment while other definitions for assets have a broader scope to include tangible and intangible objects. Assets in the PAM environment are divided into four generalisations that include: human assets; information assets, financial assets and intangible assets. Within the PAS 55 (BSI, 2008) literature, it is clear that even though PAM does not directly deal with all of these asset types, they do have a direct influence on the performance of the organisation's physical assets. It is therefore necessary to have a holistic approach towards PAM that includes all of these asset types. These dependencies are made clear in Figure 2.2.

When focussing on intangible assets, Chareonsuk and Chansa-ngavej (2010) supports the idea that intangible assets are becoming more important within the value creating process as a supporting mechanism for the direct revenue producing physical assets. Chareonsuk and Chansa-ngavej (2010) also states that this is evident when looking at how modern organisations have integrated intangibles throughout to aid value creation through decision support.

Hastings (2009) goes further and states that AM entails activities that tie in with: the identification of required assets; identifying funding; acquiring the required assets; providing the supporting systems and structures for assets and then the decommissioning and replacement of assets. He also adds that this is how to achieve the desired objectives of the organisation effectively and efficiently. To achieve this Amadi-Echendu *et al.* (2010) and Wittwer *et al.* (2002) argues that AM requires an approach that draws from multiple disciplines. This entails that the management strategy draws from strategy execution, risk assessment, safety management, environmental studies and the

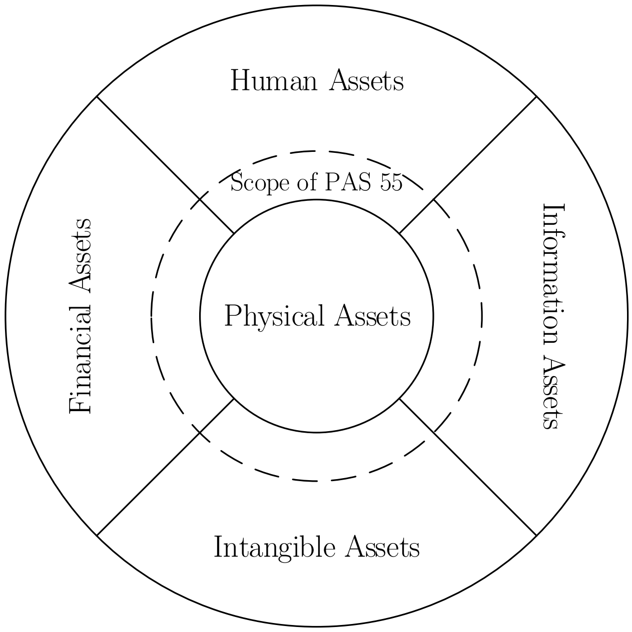


Figure 2.2: PAS 55 context and positioning of asset types.
Adapted from BSI (2008)

humanities. An interdisciplinary approach is used to transcend the limited view of AM as a maintenance plan as argued by Amadi-Echendu (2004).

Wagner, Speranza, Rogers, Bacon, Ispass and Bobchek (2003) claims that AM is fundamentally an alignment strategy between the organisation's corporate goals and its asset spending. This statement stems from a perspective that AM strategies are adopted to reduce costs, manage risks and drive the organisation's objectives and goals. There is an important point here in that alignment is necessary between the organisations goals, decisions made by management and engineering tasks as supported by Brown and Humphrey (2005). The decisions made on all of these levels require reliable asset level data that is able to support any AM strategy. In other words, a fully integrated information system is needed which is able to work in conjunction with the business processes and the people within the organisation.

When considering the AM framework used by UIC (2010), there are three main categories in which AM components play a role. These guideline are also in line with the guidelines set out in PAS 55 (BSI, 2008):

- Organisational decisions and activities: The strategies that enable the organisation to deliver the value intended;
- Enabling mechanisms: The supporting structures of the decisions and activities; and
- Reviewing Mechanisms: These mechanisms form part of the feedback loop that monitors different performance indicators to aid in future decision and ultimately improve performance and efficiency.

The focus of this thesis will lie on the enabling mechanisms and will touch upon reviewing mechanisms where necessary. When looking at the reviewing mechanisms as discussed in UIC (2010), there are six main categories which serve this goal:

1. Asset Information;
2. Risk Management;
3. Life-Cycle Costing Tools;
4. Business Processes;
5. Competencies; and
6. Supply Chain Management.

UIC (2010) goes on to explain that complexity in PAM systems require clear definitions of the business processes which link the components in the framework. These business processes also provide a means by which PAM can integrate itself within the organisation. There is thus a need for a clear process implementation and development strategy.

2.1.1 Physical Asset Management Strategy

The Physical Asset Management Strategy (PAMS) is the organisation's attempt to plan how it intends to achieve its goals (Hastings, 2009). In other words, it is a plan on what has to be done, where responsibilities lie, under which authority they fall and how these activities relate to the PAM activities. The PAMS should be able to coexist with the business strategy of the organisation, as the two will influence each other. According to IAM (2011), the PAMS is done alongside decision-making activities and needs to look at the long-term effects of:

- future demand level;
- levels of service;
- asset deterioration;
- risk change over time;
- new technologies;
- legislation and regulation changes; and
- economic environment changes.

Rajasekar (2014) argues that people and organisations mainly believe it is a good strategy which makes a company successful while it is actually the execution thereof, which should be the focus. With this, he states that between 50% and 80% of strategy executions fail. This argument is supported by Kaplan and Norton (2001) who argues that the focus should not only be on the planning of the strategy but also on the execution. In general terms, Kaplan and Norton (2008) identifies four aspects which make implementation of strategies difficult: (1) The lack of strategic targets and the lack of understanding by employees; (2) The failure to link financial planning with strategic execution; (3) Not enough attention is given to it by managerial levels; (4) Strategy execution does not call for organisation wide cooperation. Rajasekar (2014) supports this view in that he states that strategy execution cannot be done solely by management. He goes on to state that while strategy planning is a top-down managerial activity, execution is a hybrid of top-down, bottom-up and networked efforts. Davis (2007) continues this argument but also adds the aspects in which strategy supports the organisation:

- Supports the documentation of all assets and who is responsible for that asset;
- Enables the capturing of the location of all assets;

- Allows the organisation to determine the condition of all its assets;
- Helps the understanding of an asset's operations;
- Enables the development of an asset care plan to support its assets; and
- Allows effective and reliable operation of its assets over their entire life-cycle.

Hrebiniak (2005) emphasises that execution is the key to a successful strategy. It is therefore necessary for managers to not only focus on the planning and the formulation of a strategy but also the execution thereof. Marakon Associates conducted a study in which gaps in strategy execution were investigated. The survey, as reported by Brannen (2015), showed that respondents rated: (1) better communication strategies; (2) clear identification of tasks to be done; (3) performance monitoring and progress tracking; (4) holding personnel more accountable for actions; (5) giving people freedom and authority; (6) getting better people behind strategy execution from the starting phases and (7) attaching consequences to both success and failure, as powerful focus points. The survey also showed that larger organisations rated themselves lower in terms of strategy execution as compared to smaller organisations. This is possibly due to the inherent complexity which rises as organisation grow larger. With this complexity comes an increasing demand on the coordination required to implement strategies.

2.1.2 Asset Management Decision Levels and Processes

McLaughlin (1995) voices the argument that:

> Decisions are the core transaction of organisations. Successful organisations outdecide their competitors in at least three ways: they make better decisions; they make decisions faster; and they implement decisions more.

In this sense, decisions are at the core of every organisation. When looking at strategic decision-making governing objectives, policies, resources and then management control as a supporting mechanism, it is clear that there are different levels of decision making within an organisation (Straub and Welke, 1998; Otley, 1999). This is also true for AM where decision levels apply to different levels of management but also govern the perspective of the system and the detail of data required for the decision to take place. With the increase of the scope of the managerial decision, the detail of the data required is reduced and vice versa, as shown in Figure 2.3. Because of the information flow within the organisation that is necessary for decision making on all levels, these levels are inherently connected and with that, the boundaries become somewhat distorted. The separators between these levels are mainly determined by the scope and data required for that level. Flintsch and Bryant (2006*a*) argues that even though the decision levels in this case are clearly defined, there is still some overlapping taking place in terms of management. This makes the task of setting boundaries for the required data hard to define and thus a substantial task on its own. According to Haas, Hudson and Zaniewski (1994) and Hudson, Haas and Uddin (1997) these levels are:

- **Strategic**: Encompassing the entire organisation. This includes all assets and systems;
- **Network**: This level has similarities with the Strategic level but with a narrower scope which focusses on departments and the network with which it connects;
- **Program**: The decisions on this level involve policies that govern inter network allocations and actions;
- **Project Selection**: On this level, the monetary allocations are done towards desired projects. Some of the scope is shared with low level managerial actions but requires more detailed data than program or network levels.
- **Project Level**: This levels entails decisions that focus on a specific project application. This includes the work plan to achieve the performance required.

With the application of AM in industry, there is no doubt that performance enhancement has been achieved (Blache, 2010). Brown and Humphrey (2005) argues that there are still untapped opportunities in that organisations still do not map out aspects of their AM processes. With AM becoming more prevalent within industry there is a direct need to know what AM processes are

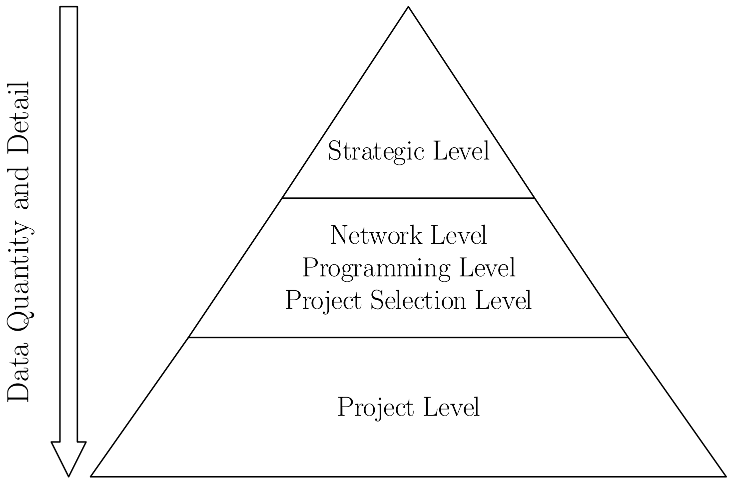


Figure 2.3: Decision making levels and the detail of required data.
*Adapted from Flintsch and Bryant (2006*a*)*

taking place within an organisation. This not only aids in decision making but also allows the improvement of the processes that support the value driving assets (Woodward, 1997).

Lingamaneni (2010) recognises that there are two models that can be used to describe AM Processes. The first of which is the functional model which has been proposed by CIGRE (2002), organises and categorises functions within the organisation by a accountability factor. Secondly, Quak, Smit and Gulski (2006) proposes an approach which is based on the structure of the organisation.

When taking the definition of AM given by AAMCoG (2012):

> Asset management is the process of organising, planning, designing and controlling the acquisition, care, refurbishment, and disposal of infrastructure and engineering assets to support the delivery of services. It is a systematic, structured process covering the whole life of physical assets.

It can be seen that the core of this definition is that AM is fundamentally a process. These processes are activities that ultimately allow the organisation to realise the value of its assets. It is thus necessary to extrude the core supporting processes to improve AM.

2.1.3 Asset Management Information Systems and the Data Driven Business

With the rise of the vast amounts of data being captured by organisations (often referred to as "big data") and the so called "information age", the boundaries of what is regarded as an asset is expanding. When referring to AM in industry, it used to refer only to physical assets but now also include intangible assets. This is evident as the main assets of an increasing number of companies are intellectual and not physical (Tapscott, 1999). It is important to note that fundamentally, the information age forces and inherent shift in how we think about how value in a company is generated. Companies that operate within a value creation model that utilises data and information often do not have this value creation potential within the balance sheets. It is thus highly evident that valuation of intangible assets will become an important aspect of the information age. The changing nature of value can better be described by the following (Tapscott, 1999):

> The Net (information network) provides a new, function-rich, high-capacity, and nearly ubiquitous infrastructure for business. It enables firms to enrich products with information, knowledge, and services for unique competitive advantage.

Big data has not only caused an explosion in the amount of data that can be captured and processed but also, indirectly, the way we communicate. There has been a decline in communication costs, ease and speed. This does not only transform communication but also the way information is distributed. The distribution of information in turn has a major shift in the strategic position of a company, especially if that company relies on that information, proprietary or not, as discussed by Evans and Wurster (1999). It is seen as inevitable that this will cause changes in the structure of companies and the way that they compete.

Evans and Wurster (1999) continues to argue that even though a company might not see itself as an information business, there is a large portion of economic resources that go into the information system that supports its operations. Irrespective of the exact operations of the value chain, the information within the system is what keeps that value chain from disintegrating. This is especially true when the value chain includes proprietary networked information between the organisation, its suppliers, distributors and the customer. Evans and Wurster (1999) goes on to state that the organisation's relationships and brand loyalty are all included in the value chain and are all based on the information network created within the organisation.

With this considered, it is evident that successful organisations are those in which there is a solid foundation for the knowledge base that supports the functions and actions of that organisation. It is thus important for survival in the information age that there exists an interconnected set of minds between which information and knowledge can diffuse to enable value creation and efficient managerial actions. Managing collaboration also ties in with the arguments raised by Hrebiniak (2005) which show that the main problem with strategy is the execution thereof.

In the instance where an organisation's value creation is supported by its knowledge base, the value chain creates a virtual feedback loop. This feedback loop relies on the continual learning and data processing during the operations. Real-time feedback from the environment is crucial for an organisation to survive in the information age as a stationary knowledge base can quickly become obsolete, and with that, the potential value that can be created by an organisation.

The idea behind having an organisation driven by data has only emerged quite recently. When this is coupled with the difficulty and resilience to change established business models, as stated by Teece (2010), the task of integrating data awareness into established organisations becomes a challenging task. In a study conducted by Criscuolo, Nicolaou and Salter (2012) which compared start-up companies to established organisations, it was shown that start-ups offer better innovations in the service industry. The innovation offered by these start-ups is largely due to their ability to build a pure business model which is driven by data and information that supports their services (Hartmann, Zaki, Feldmann and Neely, 2014).

While all the above mentioned aspects are integrated into a conventional business model, Hartmann *et al.* (2014) goes on to argue that there are six dimensions to a business framework that is data driven:

1) **Key Resource**: This is what enables the organisation to create value. When looked at from a data driven business model (DDBM) perspective, the data in drives the primary value creation potential. This resource can be structured, such as Enterprise Resource Planning (ERP) data, or semi-structure data, such as strategic or instructional information.

2) **Key Activities**: These activities support the downstream usage of data. Where in a traditional sense this would translate to the information to support the product value chain, here it is better suited towards enabling the use of data as a resource.

3) **Offering**: As any business framework is somehow centralised around the offering towards the customer, this has to be at the core of any business model. Without being able to offer value towards the customer, a business will surely fail. Within a data driven business model, this offering can have three substantial interpretations. The first two offerings are covered by Fayyad, Piatetsky-Shapiro and Smyth (1996). Either it can form part of the core offering in the form of raw data, or the results after an analytics process can be offered in the form of information or knowledge. The third offering is when the data supports the company's physical product or service.

4) **Customer Segment**: These aspects deal with the target audience of the product. This can be directed either towards a corporate entity or towards people.

5) **Revenue Model**: This dictates the revenue stream/s of a company including either the primary or the secondary forms of cash flow. It is also not exclusively dictated by sales and can come in various forms such as renting, leasing, intellectual property usage as described by Osterwalder and Pigneur (2010).

6) **Cost Structure**: The cost structure points towards the ongoing operating costs. In other words, the other business model dimensions which results in these costs.

2.1.3.1 Information in the Asset Management Environment

According to OpenText (2012), many of industry's main problems are directly related towards AM. These issues include the need to maximise the uptime of assets that generate value for the organisation. Knowledge and information is required about these assets that can be used to ensure that optimal efficiency is achieved. This also requires the enabling functions of AM to be able to support the assets throughout their life-cycle. Work Management (WM) plays an essential role within this supporting structure as described by Schuman and Brent (2005).

When looking at the information that is present within an organisation that supports its function, the majority is unstructured (OpenText, 2012). This entails information that is not physically stored on a central information bank but flows within the organisation. This information is usually found in the form of emails, social interactions, faxes, etc. When different individuals handle critical information and that information is critical to the operations of that organisation, information management becomes a complicated managerial problem. This is especially true when the information in question is the supporting mechanism that aligns business processes and the organisation's value goals and activities as argued in Solaimani and Bouwman (2012).

To address the rising complexity, enterprise resource planning (ERP) has been implemented throughout industry. With the ERP systems, the organisation is able to automate many of the common managerial actions related to different departments and services. The ERP system handles most of the data capturing tasks related to the processes and assets within the organisation. In many cases, the ERP system forms the foundation for a data-driven environment for many organisations from a managerial perspective focused on in house operations.

The data captured within an information system forms part of the intangible assets that will have a life cycle like any other asset. As shown in Yu and Wen (2010), this life cycle is generally in the form of a linear progression which involves the *creation*, *storage*, *usage*, *sharing*, *archiving* and then finally the *destruction* of the data. The proposed data life cycle is shown in Figure 2.4. The *storage* component duration is highly dependent on the type of organisation. Some organisations are very dependent on real-time data where month old data can be discarded whereas some organisations require historical patterns and might store data for years. Independent of the exact situation, the destruction phase is entered as soon as the organisation can no longer extract value from the data.

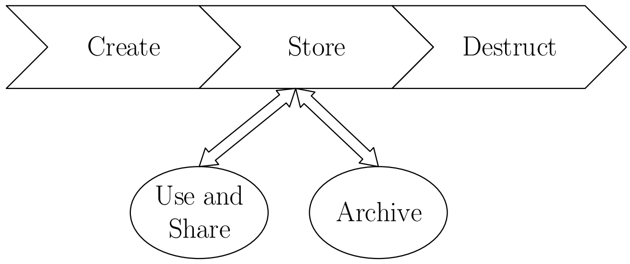


Figure 2.4: Data life cycle.
Adapted from Yu and Wen (2010)

2.1.3.2 Data Management for Asset Management

As mentioned previously, organisations have grown in complexity in recent years and tend to implement Information Technology (IT) systems to handle and store all the data that the organisation deems important. Information in this case is mainly to help drive its business practices. The data in most cases is stored based on the availability and not need, causing the storage of unnecessary data points. These data sets then become a managerial problem as it now becomes a strategic effort to extract value from it. Data managing becomes strategic especially in cases where different departments, applications and AM functions store their data on different systems as stated by Gupta (2009).

AM addresses two types of structured data namely: (1) master data and (2) transactional data. The former entails data relating to the organisation, the customers, the product etc. while transactional data refers to orders, claims and reservations (Gupta, 2009). Troester and Haxholdt (2012) mentions that one of the reasons there is a shift in how organisations manage data and why it has become so important, is the emergence of big data and how it ties in with business analytics. Big data has resulted in a increase of dataset size that allows algorithms and statistical models to reveal correlations on a scale not possible previously. As stated in Akerkar (2014), industry can now analyse enormous clusters of data without hypotheses and look at what correlations the algorithms find. These type of correlations grant great insight into what is happening, it does however not indicate why. It is here where the data analyst steps in.

Given the importance and possibilities that can result from implementing big data, it is important to identify the properties that contribute to this concept. Big data consists of three main attributes O'Reilly (2012):

- **Volume**: At the core, big data draws its value from the sheer size of the data sets. When more data is available, analytics can reveal more insights. This immense volume of data also poses the biggest problems on the technical side as it calls for new ways to store, process and effectively manage datasets.
- **Velocity**: The velocity of the data more commonly refers to the rate at which it enters the organisation. This velocity can be attributed to general data gathering scenarios where it will be later used for batch processing or it can enter a feedback loop that influences decision making. This velocity of data flow is not only limited to data flowing from the environment into the organisation, but can also flow back the consumer. Ultimately, those who are able to quickly utilise new information will gain the a competitive advantage.
- **Variety**: When an organisation gathers data from the environment, it is almost never in a uniform, ordered and process ready state. Data gathering involves collecting data from a variety of sources, in different formats, then determining what is useful, and what is not. Big data in reality involves taking unstructured raw data, applying filtering and restructuring in order for processing to take place. These pre-processing endeavours can involve as much as 80% of the effort (O'Reilly, 2012). It is however common practice to, when possible, keep as much of the raw data as there might be something useful.

Following the discussed attributes of data currently seen in organisations today, it is clear that data management has become more important. It is however not enough for an organisation to capture any data it can and expect results to surface. There needs to be an understanding of what the organisation wants to accomplish and the intention of doing so. It is therefore also important to note how data is being collected for the purposes of AM and what the characteristics is thereof.

Data Collection and Characteristics

Data collection over the last few decades has moved from a manual and time intensive process to where, today, it has become overwhelmingly an automated and easy process. This has mainly been possible because of automation and computerisation as argued by Flintsch and Bryant (2006*a*). Within AM there are three data collection techniques as shown in Pantelias (2005). These include:

- Manual: This would entail a person having to go and observe a variable and then logging a value at a point either in time or at certain time intervals.
- Automated: Here the process is automated, including capturing the data and then the input into the information system;
- Semi-automated: In this case, the capturing process is automated but still requires physical retrieval as it is either not connected to the information system or is capturing data in a different location.

In general, numerous characteristics are important when collecting data. In the case of AM however, there are four characteristics that are deemed to be crucial in decision making as is discussed in Pantelias (2005):

- Data Integrity: In the case when two different observers were to measure the same variable, their value should be the same;
- Data Accuracy: The value measured should be as close as possible to the observed value;
- Data Validity: The value measured should be a true representation of reality; and
- Data Security: Confidentiality and sensitivity should be considered when handling data that includes personal information or is representative of intellectual property.

Within the AM framework, data forms the foundation on which objective decisions are based. It is therefore imperative that the data used in these decisions are there when needed and that they have the required accuracy and reliability for the decisions at hand. With this, certain requirements are placed on data collection. Flintsch and Bryant (2006*b*) identifies three categories that must apply to the data collected:

- Location: This attribute represents the physical location of the asset;
- Physical attributes: This would entail any physical description applicable to the asset that includes visual characteristics or type; and
- Condition: The condition of the asset forms part of this description and can include qualitative or quantitative descriptions.

The detail required for each the above mentioned categories will vary with each decision level. This is assuming there is a need for it at all. The detail and the availability that goes along with these decision levels should thus be considered within the AM strategy, as this will influence the collection procedures.

A Value Chain for Data

Data in its purest form is simply a raw numerical representation of what has been measured. It can be seen as the basis for all objective decisions. It is however important to note that data on its own inherently has little value. The volume of data being gathered does not fundamentally increase value, it merely gives rise to the potential of value creation. Organisations often forget this concept as they are enhancing their data gathering efforts without a need or strategy to enable the implied value.

With the increase of the volume, velocity and variety with given datasets, the challenge has become to be able to interpret this raw landscape into something specific. Figure 2.5 shows a proposed framework devised by Miller and Mork (2013). This framework gives a flow of actions that support overall management of captured data.

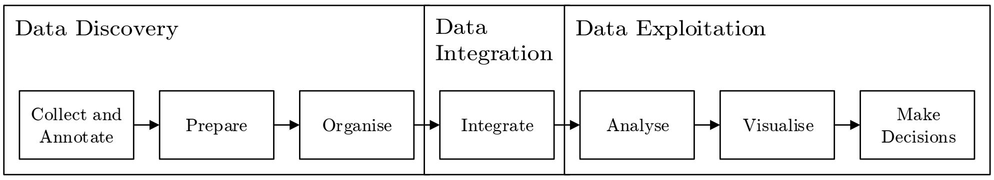


Figure 2.5: Data value chain.
Adapted from Miller and Mork (2013)

The value chain aims to manage and coordinate the data from acquisition to a decision making platform. It also aims to form a collaborative system to improve data acquisition and improve the quality of outcomes that result from decisions based on the extracted information. Most importantly however, it aims to establish a portfolio to maximise the value gained from data acquisition (Miller and Mork, 2013).

Before any decision making can begin, the organisation needs to identify the required data. Data identification also entails preparing the required inventory and then preparing and organising the data assets. The following briefly describes the required steps in the data discovery process:

1. **Collect and Annotate**:First it is important to identify the possible sources of data. Thereafter these sources should be placed within an inventory where the completeness, validity, consistency, timeliness, and accuracy can be described.

2. **Prepare**: Preparing the data involves creating a shared system for universal access by those that require it and creating privacy terms.

3. **Organise**: This step includes sorting the data and establishing the required syntax and format for stored data. This is usually unique to the organisation but it should be noted that standardisation is an important part of knowledge sharing and should be considered to ease communication with the data providers and consumers (Miller and Mork, 2013).

When the data is organised, the integration process starts by setting out a map which constitutes the links between the data and the common representation (Miller and Mork, 2013). This process is mainly to facilitate future analyses.

After the data has been gathered, organised and integrated, decision makers can start exploiting the data to reveal information. This information can be in the form of visualisations or simple correlations depending on the task. This information is then used to enable the person to make informed decisions for the organisation (Miller and Mork, 2013).

Data exploitation can be broken up into three main steps:

1. **Analyse**: The integrated data is analysed in such a way as to preserve all inputs and follow a clear methodology so results can be recreated by others. This helps support results and validity.

2. **Visualise**: This step involves presenting the results from the previous step in such a way as to support the results. The aim of this should be to help stakeholders digest the critical information for them to make decisions.

3. **Make Decisions**: The final step is to use the information at hand to make the best possible decisions and to plan the actions to be taken. Actions should always be traceable to the information which supports that decision (Miller and Mork, 2013).

In short, data is the untouched observations that require processing to extract useful information. The value chain shown in 2.5 is an attempt to present a systematic approach to gain value from raw gathered data. Without such an approach, gathering data becomes cumbersome and may result in the organisation not being able to extract value from its efforts.

2.2 Business Process Management and Workflow

This section will introduce the theory and background of business process management (BPM) and Workflow. First BPM is discussed as an overarching discipline which includes workflow as a subset. BPM will then be linked to the process discussion that occurred in the previous section. The principles used in BPM and workflow will be discussed in order to understand the relevance to the PAM environment and how it may be of benefit.

2.2.1 Business Process Management

When referring to a business or an organisation, there are two main attributes used by Solaimani and Bouwman (2012) to describe their function, namely the business model (BM) and the business process (BP). Osterwalder and Pigneur (2010) defines a BM as: "The rationale of how an organisation creates, delivers, and captures value" while Solaimani and Bouwman (2012) adds that it is "the logic of the intended service". An organisation's BPs are the activities that ultimately support the BM. When talking about a BP, it fundamentally refers to the flow of activities performed within that organisation to produce an intended value. The flow of these activities is set up in such a way as to improve the use of organisation's resources and assets. Poulymenopoulou, Malamateniou and Vassilacopoulos (2003) defines BPs as:

> *Sets of partially ordered and coordinated activities, often cutting across functional boundaries, by which organizations accomplish their missions.*

While Weske (2007), more recently, defined BPs by the following:

> *A business process consists of a set of activities that are performed in coordination in an organisational and technical environment. These activities jointly realise a business goal. Each business process is enacted by a single organisation, but it may interact with business processes performed by other organisations.*

In an ideal case, the BPs are constructed in a way that they are able to seamlessly support the BM. More often than not, there is a misalignment between these two concepts within an organisation that ultimately leads to losses in asset performance and thus the value offering (OpenText, 2012). Solaimani and Bouwman (2012) presents a general framework shown in Figure 2.6 which illustrates how information lies between the organisation's processes and the value it aims to achieve. It is thus clear that the information layer bridges the gap between the processes and the available value creators.

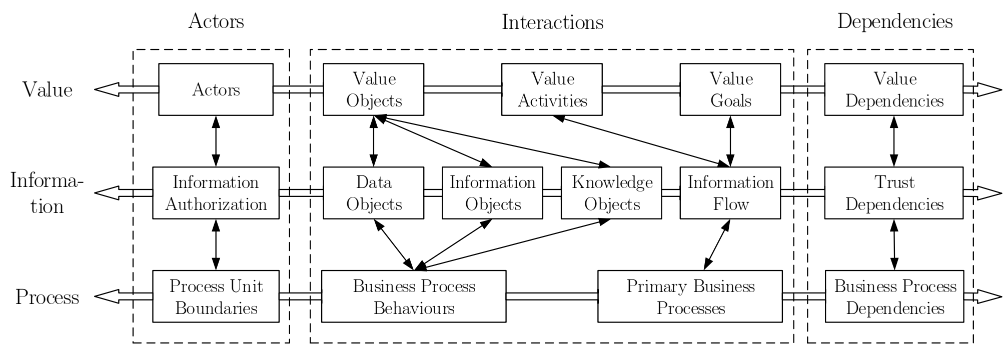


Figure 2.6: Value, information and process (VIP) framework.
Adapted from Solaimani and Bouwman (2012)

With BPs being recognised as being central to any efficient organisation, Business Process Management (BPM) has emerged as a managerial discipline which is primarily used for integration purposes within large organisations where Business Intelligence (BI) techniques are used, including practices such as Online Analytical Processing (OAP) and Data Mining (Chen, Chiang and Storey, 2012). It creates a visualisation platform of the supporting structures that enable value driven services and also allows the improvement of how activities are performed and distributed (Davenport, 2005; AguilarSaven, 2004). Frolov *et al.* (2010) argues that this concept is crucial with the increase in complexity of strategic management.

Van der Aalst, Ter Hofstede and Weske (2003*a*) asserts that BPM was the result of the development that took place in workflow in the nineties. It is because of this that BPM borrows terminology from workflow. Workflow is defined by Coalition (1996) as:

> *The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules.*

With the definition being regarded by some as too restrictive, BPM was brought in as a more flexible area that also encompasses workflow. Van der Aalst *et al.* (2003*a*) goes on to mention that the systems that support BPM need to be process aware in order to achieve desired outcomes. Figure 2.7 illustrates the relationship between BPM and WfM and how they encompass the life cycle of a business process.

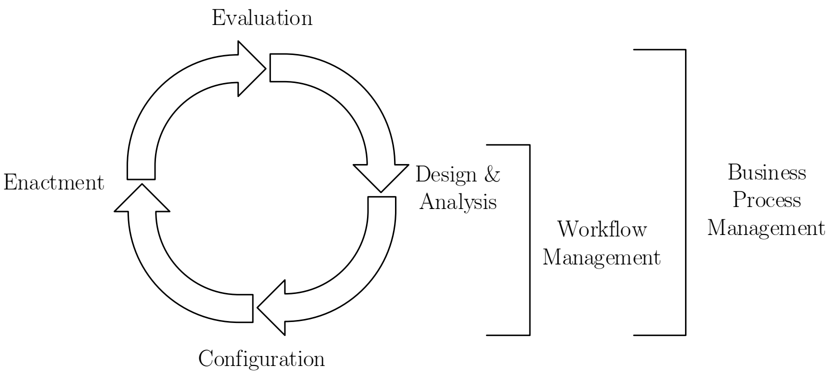


Figure 2.7: BPM life cycle within workflow management and business process management.
Adapted from Van der Aalst et al. *(2003*a*)*

Tinniä (1995) deems it important to note that there are three perspectives from which BPs can be looked at. The first recognises IT as an important part of the BP environment and argues that BPs are ultimately the way to operate more efficiently. This is not limited to using IT as the mapping resource for BPs but also the integration of IT within the BPs, especially in networked processes with a distribution of roles. Secondly, corporate performance can be improved by analysing and mapping: multifunctional business processes, resources requirements within that network and the IT support required. Thirdly, BPs can be looked at as a strategic planning object. This argument places BPs at the core of strategic planning as opposed to products or markets.
When considering BPM, it is important to acknowledge the stages of the BP life cycle. This life cycle as presented by Weske (2007) and in a similar fashion by Van der Aalst (2004) is shown in in Figure 2.8.

During the *design and analysis* stage, Van der Aalst (2004) describes that the existing BPs are modelled and graphically represented to allow analysis to take place. Weske (2007) continues that different BPM techniques are used alongside validation, simulation and verification techniques. This all forms part of the analysis procedure. The exercise is then concluded with the design of the

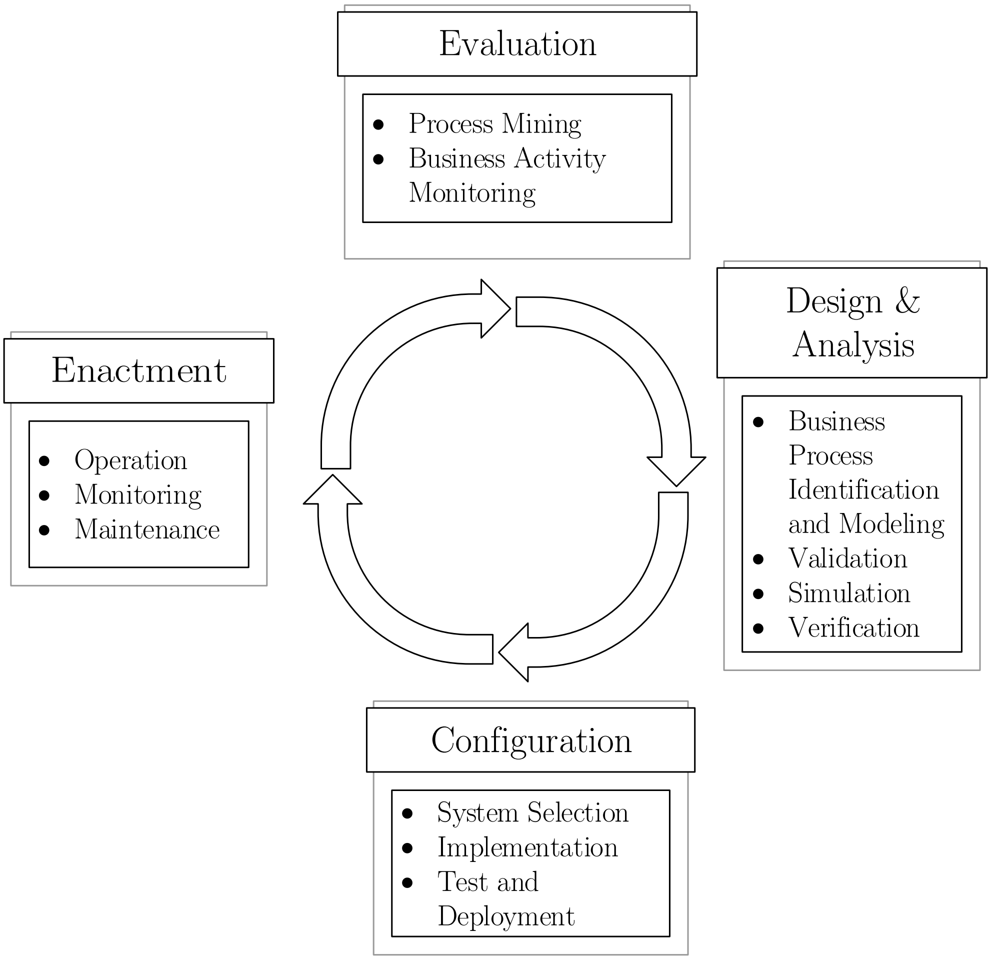


Figure 2.8: Business process lifecycle.
Adapted from Weske (2007)

desired setup.

Configuration involves the implementation of the desired BP. In the simplest cases, it merely suggests a change in policies and procedures. In large modern organisations however, this would involve reconfiguring the information system on which the organisation runs (Weske, 2007).

The *enactment* phase involves the enabling and usage of the newly realised BP. During this phase, new data is generated for later use when process improvement continues. This is usually in the form of historical log files recorded from the operational process (Weske, 2007).

The *evaluation* phase acts the trigger event for BP improvement. This involves the continued monitoring of processes to identify problem areas. This phase is also responsible for the exploitation of new available technologies which can improve performance and aid environmental adaptation within the organisation (Van der Aalst, 2004). *Evaluation* is typically performed off the historical log files generated during the *enactment* process. This thesis will be primarily concerned with this area of the BPM life cycle.

2.2.2 Business Process Intelligence

Business Process Intelligence (BPI) comes into action when the BPM life cycle enters the evaluation phase. BPI is defined by Grigori, Casati, Castellanos, Dayal, Sayal and Shan (2004) as: "a set of integrated tools that support business and IT users in managing process execution quality". When applying BPI within an organisation, there are certain techniques and measuring tools that formally exist to help the organisation achieve its goals. The following techniques are given by Ingvaldsen (2011) as examples:

- Discovery: The observation and analysis of the organisations operations in order to construct process models which represent the true nature of reality;
- Analysis: Analysis of the process models constructed in the previous phase in order to derive useful correlation between the desired process models, people in the process and the resources that are used;
- Prediction: Using the results obtained from analysing the processes, prediction can now be made to how future processes will be executed for use in budgeting or time planning;
- Deviation Correction: When deviations from the original planned process model are detected, tools can now be used to correct the problem area in order to reduce process time or costs;
- Static Improvement: Following the analysis, problem areas can be detected which might require adjustments to the resources assigned or the aspects of the tasks being performed;

- Dynamic Improvement: In contrast to the static improvement of the business process, this would entail altering the way the process flows from start to finish. This could be done in the form of different routing options or how resources are assigned to tasks as the process progresses.

BPI is ultimately the culmination of various disciplines to reach a common goal. The discipline or techniques that are involved include semantics; visualisation, statistical analysis, data mining and process mining. These might all then include their own software foundation which are then brought together to achieve a common goal.

When these approaches are used within the BPM life cycle, there are limitations that often occur and then limit the application and end result of BPI efforts. Casati and Shan (2002) refers to the following aspects:

- Data Quality: The data logged within an organisation is often not suitable for the analysis efforts. Even when the data is suitably presented, it is often filled with noise;
- Performance: The data captured within the information system is not structured for the BPI process. Therefore the data mining efforts are often intensive tasks even though it is small part of the BPI effort;
- Semantics: When data is captured within an organisation, the abstraction level does not suit BPI. This results in most of the data being of little use.

2.2.3 Workflow Management

Workflow Management (WfM) forms the supporting structures that enable business processes where people form the basic executing resource. With WfM, the organisation is able to strategically manage the tasks within business processes which include scheduling work and managing the flow of information that is required to perform tasks as discussed by Georgakopoulos *et al.* (1995). It is clear that WfM forms an integral part of BP improvement as WfM provides methodologies for the modelling of BPs and the re-engineering possibilities to extract maximum value from the current BP framework.

This idea of having an organisation structured around processes started emerging in the 1930's. When looking at the German author Hennig (1934) through to the American author Chapple and Sayles (1961), there is a common thread in that there is a clear potential for using workflow management as an improvement tool. This potential was however only starting to realise in the 1980's where the competitive nature of corporate business became more evident. Workflow management in the sense of representation was described as early as the 1970's where Zisman (1977) used Petri-nets to visually describe the processes within an organisation. Building on his work, Ellis and Nutt (1980) and Mahling, Craven and Croft (1995) continued through the 1980's and 1990's to apply this approach of managing work to office administration procedures.

2.2.3.1 Workflow Management Systems

When talking about Workflow in a more general sense it refers to the process in which documents, tasks or information is transferred from person to person. It also described to collaboration which exists when multiple tasks need to be performed by multiple individuals or entities as discussed in Casati, Ceri, Pernici and Pozzi (1995). Workflow can also be referred to as a manifestation of a business process as a computational model as demonstrated by Palmer (2009). In most cases, organisations use a Workflow Management System (MfMS) to handle the worflow necessary for the completion and execution of processes. The execution of the tasks within the workflow environment will in this case be represented by Petri-net logic, on a computational level, which will be discussed in more detail in Subsection 2.2.5 (Van der Aalst, 2011; Van der Aalst and Van Hee, 2004; De Leoni and Van der Aalst, 2013).

Van der Aalst (2004) categorises four distinct types of WfM systems:

- Pure WfM systems : In this case, organisations use a software package that is designed to be able to handle all of it's WfM requirements;
- WfM add-on components : Organisation that utilise a heavily integrated ERP system can then, in some cases, use the WfM add-on for that ERP system instead of a dedicated package;
- Custom WfM Solutions : Organisation that rely on a secure system or has very specific operation often build their own WfM system; and

- Hard-coded WfM systems : These systems use built in processes by which the organisation has to comply. There is no way to customise or add niche processes.

According to Van der Aalst (2004), most organisations still utilise systems relative to the last two options. There is however a shift towards more powerful commercial products as the flexibility and customisation options towards individual organisations increases.

When looking at the literature, the benefits of a WfMS becomes clear. A WfMS takes a given business process and then translates that into a sequential flow of work. This entails managing and assigning resources, which include the people, data etc. needed to complete the process. Without having improved the BP itself, the WfMS is already able to increase efficiency in how the process is done.

Zur Muehlen and Allen (2000) adds that the WfMS is able to automate much of the needed work within simple BPs and with that, allows quantitative information to be relaid to the managers in real time. This allows the correction of processes and reaction to unforeseen events to happen much faster. Another view that supports the idea is given by Becker, zur Muehlen and Gille (2002) where a WfMS is described as a coordination tool. This coordination occurs between the resources and/or entities that are involved within the business processes. The coordination mainly occurs to address two problems. The first being the dependencies between activities in cases where a process relies on the results or completion of another. Secondly, the management of shared resources on which multiple processes rely. With proper implementation, the WfMS helps realise the organisation's efficiency goals.

2.2.4 Business Process and Workflow Modelling

Modern organisations have evolved from simple in house operation which can be observed rather simply, to large, multi-dimensional and often, collaborative ventures. With the rise in complexity in how these organisations create value by means of a service or product, an inherent increase in the complexity of the processes that drive an organisation has occurred. Van der Aalst (2011) argues that process modelling is an adept method that can be used to handle this growing complexity as it assists in managing and documenting procedures. For this to happen however, proper implementation of an information system is required.

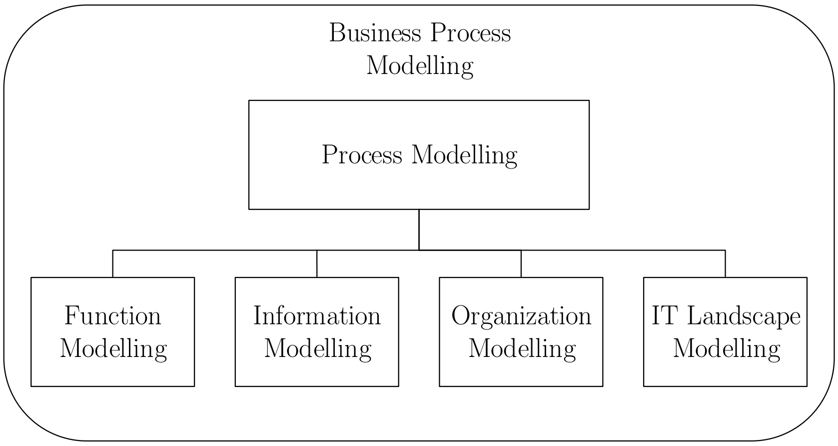


Figure 2.9: Business process lifecycle.
Adapted from Weske (2007)

Weske (2007) argues that BPM is a culmination of different modelling domains as depicted in Figure 2.9. For a complete representation of the business processes, all of the sub domains need to be included. This can however be adjusted when the BPM process requires a specific output which justifies the relaxation of constraints as stated by Weske (2007). With the regards to the model, Van der Aalst (2011) argues that creating a model within the business modelling environment is "an art rather than a science". When creating models, there are three typical errors that are made as discussed by Van der Aalst (2011):

- Models are misinterpreted and it is not realised that they show a simplification of how things aught to be. When a process is designed, it is done as a simplification of reality. This then forms the basis from which deviations in reality take place.

- Mathematical models struggle to accurately capture the workings of human behaviour which then leads to wrong abstractions and conclusions.

- Wrong choice of abstraction level which dictates the process. This is dependant on the use of the model as the desired complexity needs to be chosen as a trade-off between accuracy and complexity.

These errors, when considered, can easily be avoided. This consideration should however include the goal of the modelling outcome. When this is made clear and the situation is understood, the possible errors are seen as compromises to reduce complexity and ease the overall modelling process. The perceived benefits from BPM are discussed by Hlupic (2003); Indulska, Green, Recker and Rosemann (2009) and include the following:

- Allow the improvement of currently implemented BPs;
- Grant greater insight into how BPs are influencing the organisation;
- Ease of cross discipline communication on the workings of BPs;
- Allow implementation of process automation by knowing the supporting structures of that process; and
- Continuous improvement through analysis and simulation of process models.

At the foundation, business processes are described by the workflow that is necessary for it to achieve its goal. This is especially true where workflow is used to describe and understand business process tasks at a conceptual level. By using workflow, further detail is given as it also includes information about process requirements for information system functionality as well as resource usage and resource distribution. The focus in this thesis will lie with workflow patterns being the core of business processes.

2.2.4.1 Workflow Management Concepts and Perspectives

The WfMS is responsible for joining different workflow perspectives to enable the business process that run on top. When talking about workflow modelling, there are different perspectives that can be looked at, all of which form a crucial part of the modelling process. These perspectives are given by Van der Aalst (2004) as:

1. Control flow perspective, which deals with the process definitions and task order;

2. Resource perspective, which gives an organisational view of the structure and roles;

3. Data perspective, dealing with the information generated within the process;

4. Task perspective, which is indicative of the function in the process i.e. the operations;

5. Operational perspective, deals with the application and actions.

To understand the discussions that follow, some concept regarding workflow need to be discussed. The following concepts are described originally and in more detail in Van der Aalst and Van Hee (2004).

- Cases: From an organisation point of view, a WfMS deals with cases. This case refers to the process that deals with a desired output or product. In other words, the process and workflow of dealing with a Condition Monitoring event would more commonly be referred to as a case. Cases can also belong to the same type to such an extent where they follow the same basic route but differ in small details. This makes it easier design new processes for additional cases as templates already exist. Van der Aalst and Van Hee (2004) also states that between the start and completion of a specific case, three attributes exist to describe the current state. First, there are the values for the different attributes referring to that case, second there are the conditions relating to that case and thirdly, the case's contents.

- Task: When breaking down a process to its most fundamental form, it can be seen that it is held together by individual tasks or activities. These tasks refer to the intangible aspect of work. It refers to the thing that a role player needs to finish in order to progress further along the process. Each task needs to be completed in full in order to complete a process. Looking at the progress of a process, a task has boolean properties in the sense that it can only be finished or not.

- Process: Van der Aalst and Van Hee (2004) refers to a process as "a procedure for a particular case type". In this sense, a process is a sequence of tasks and activities that is applied to accomplish the execution of a case. This inherently mean that different cases can use the same process, either the case refers to a standard process that applies exactly or different cases can refer back to a template which forms the basis for the case but is merely adjusted for very specific attributes. As the process builds upon tasks, a condition attribute is added to refer to the order in which different tasks are performed.

- Routing: Within a process filled with tasks, there is a certain path that needs to be followed in order to be able to go from the start of the process to the outcome. Routing refers to this path or branch that lays out the sequential nature of the process. In the simplest form, the routing can exist merely as a single line that goes from task to the next and so on until the process is completed. In most cases however, there are branches within the process that split from on task and then connect to two simultaneous tasks which needs to be completed before moving on. In this case, the routing is referred to as parallel routing.

- Triggering Events: Trigger events exist as the set off point for any process. Trigger events are the events that calls in the need for a case to be satisfied and thus enables the process to start. Either a person or an automated system can perform the triggering. In essence, there are three types of triggering events. First, a person can trigger an event as part of an initiative to accomplish an outcome. Secondly, external or environmental events can enable a trigger as a reaction, and then thirdly, a trigger can have a time attribute and therefore triggered as part of a routine.

Van der Aalst (2004) separates the kinds of workflows into three dimensions. These dimensions encompass the control-flow dimension, the resource dimension and the case dimension. The first dimension, control-flow, handles the type, placement and layout properties of the tasks. This would entail whether the tasks are conditional, parallel, sequential or iterative. The second dimension, resources, covers the appointing of resources (personnel, physical asset) to certain tasks for execution. This dimension goes on to appoint the resource class or role from an organisational entity. The case dimension is attributed towards the different cases that are specified to be executed by a specific process and resource. In general, workflows are processes that were executed to satisfy a case. Van der Aalst (2004) refers to these processes more formally as case-driven processes.

2.2.4.2 Workflow Attributes

Becker *et al.* (2002) distinguishes between five workflow attributes and identifies possible values that can be assigned to those attributes. This classification scheme becomes useful within the modelling environment when the implementation relays back a value for the attribute. The first of these attributes are "participants". Becker *et al.* (2002) does not use the term "role" as is used by numerous other sources to allow for other types of values to be assigned to this attribute. Values such as "machine" or "software" are included as viable values. This allows for a more comprehensive and accurate representation of what occurs in industry as these assets also operate with the restriction of a completion time and thus form part of a business process.

Becker *et al.* (2002) continues and describes a "process structure"attribute. With this, he assigns possible values that consist of a mixture of ad-hoc processes and activities. Van der Aalst (1998) and Koulopoulos (1995) give a better representation of this attribute in the form of "production workflow" which is characterised by a fixed order of tasks and the definition thereof. They also add an "administrative workflow" which entails a well defined template and structure but allows alternate routing where possible. Voorhoeve and Van der Aalst (1997) describes the last process structure value as "ad-hoc workflow" where a generalised template is constructed and with that, various alternate routings are developed for individual cases.

Another attribute that plays an important role is the scope of that process. This scope can be as broad as between numerous organisations, encompassing the entire organisation or simply be within a department or application. Becker *et al.* (2002) also identifies "data granularity" as an important attribute that handles the type of data or information within a specific process. The handover may occur involving documents, objects or attributes.

All of the concepts described above are necessary when considering to model any BP that involves workflow or the handover of documents. While the perspectives in some sense encompass the scope of the modelling project, the concepts and attributes then realise the level of detail that is required. The understanding of these characteristics are thus important when discussing the BP's and workflow within their respective domains. This is especially true here where these characteristics are the relationships between the respective disciplines.

2.2.5 Petri Nets

In 1962, Carl Adam Petri introduced a method of mapping and creating diagrams for concurrent and synchronous distributed systems (Petri, 1962). This would specifically refer to discrete events as shown in DiCesare, Harhalakis, Proth, Silva and Vernadat (1993). These diagrams relied on the system having different states and trigger events which would then cause a shift in this state. The strong mathematical foundation that existed for this modelling tool ensured that it had been widely used and researched as shown in Zuberek (1991). The usefulness of Petri-nets is also shown by the fact that it is still being used today, if not in its original form, in some sort of adaptation to complement a specific scenario.

The power of Petri nets is mostly due to their ability to have the strong mathematical foundation. Their graphical representation also make them accessible and easily understood. This makes them useful within the design process as shown in Zuberek (1991). According to Murata (1989); Peterson, Petri nets have three main components: *places* which are represented by circles; *transitions* are shown as rectangles and then *arcs* which are shown as arrows. Places represent the possible discrete states the system can be in. Transitions refer to the events that cause the state of the system to change. These places and states are then connected by the arcs. It should be noted that two of the same components cannot be connected. A place will always be attached to a state and vice versa. To represent the current state of the system, black dots are placed in the circles (places) which can then be moved to the next when a transition is fired. A transition is only trigger-able when the state is satisfied in terms of tokens.

Table 2.1: Definition of classical Petri net.

Adapted from Murata (1989)

Petri netrs are 5-tuple in the form of $PN = (P, T, F, W, M_0)$ where: $P = (p_1, p_2, ..., p_m)$ represents the places which are finite, $T = (t_1, t_2, ..., t_n)$ represents the finite transitions, $F = (P \times T) \cup (T \times P)$ is the arcs with flow relations, $W : F \rightarrow (1, 2, 3, ...)$ forms the weighted attribute, $M : P \rightarrow (0, 1, 2, 3, ...)$ is the initial marking, $P \cap T = \theta$ and $P \cup T \neq \theta$. When a Petri net without an initial marking exists, it is represented with $N = (P, T, F, W)$. Petri nets with initial markings are denoted by (N, M_0).

Murata (1989) describes that when system is initialised with the starting of M_0, a state change can occur from which the system will be represented by M. With any given value for the marking of $M = \{M(p_1), M(p_2), ...M(p_|P|)\}$, a system is represented with the tokens placed within the given states of the system. For the system to change, a transitioning event needs to take place. This is modelled by $t \in T; t = \{p \in P : (p, t) \in F\}$ which is then the input place for that transition and then $t = \{p \in P : (t, p) \in F\}$ which is the output places for that transition. When considering the transition t, it is enabled within the set M if, and only if $p \in t, M(p) \geq 1$ for all. When this equation is satisfied, the current state M can progress to M'. The transitioning event is expressed by $M[t{>}M'$ alternatively $M \underset{\rightarrow}{t} M'$. It should be noted that with this, when a transition occurs, the token is not moved to the next place, rather it is consumed by the transition which then produces a new token within the next place.

The mathematical foundation above for a classical Petri net simply states that with a given system, tokens are placed within places which represent the current state of the system. The system can then change state by firing a transition. This will cause tokens to be "moved" (by definition they are not moved but destroyed and created). The firing of a transition represents an action has been taken or an event which has occurred. The firing of these events are bound by the availability of the tokens, whether or not there are sufficient tokens within the input places. This will cause the transition to be enabled.

As stated previously, the classical Petri net is discussed above. There have been many adaptations from this for specific cases. Most notably, Jensen (1997) presents Coloured Petri Nets (CPN), where tokens carry different data values which is indicated by different coloured tokens. Marsan, Balbo, Conte, Donatelli and Franceschinis (1998); Coolahan Jr and Roussopoulos (1983); Van der Aalst (1993) and Walter (1983) show how a time attribute can be added to either transitions, places, tokens, or arcs respectively. Temporal systems are analysed using these techniques. Jensen (1997) suggests that by having a Petri net modelled after a given system, two properties of that system can be analysed. The first is behavioural and marking dependant properties that are dependent on the initial state or marking of the system. This entails the analysis of how the initial values for the system influence the outcome of a simulation. The second is the structural properties of the system. Being independent of initial values of the system, the layout can be analysed and then improved.

Looking at works such as Van der Aalst, Desel and Oberweis (2000); Ellis and Nutt (1996), ot can be seen that the majority of applications for Petri nets lie with workflow modelling. The biggest reason for this is that, contrary to other modelling techniques, it can be used for structural modelling as well as qualitative and quantitative analysis. Other reasons include, as discussed above, the formal mathematical background and the ability to show the system graphically. Another important aspect of Petri nets is that they are state-based and not event-based. Van der Aalst (1996*b*) explains that this is important when it needs to be made clear when individual tasks can be fired and that there can be a delay before an event is triggered after the previous task has been completed.

As good a tool that Petri nets are, some authors such as Holt (2000); Han (1998) have stated that they are not ideal for the representation of workflow models. Han (1998) also goes on to argue that it is not possible to represent all business models with Petri nets. This has however not hindered other authors from using various adaptation and variants to apply Petri nets for specific cases. Van der Aalst (1996*a*, 1998) shows that given the three dimension of workflow which entail the process, case and the resource, the process can be mapped when the resource is excluded. The resource can still be modelled within the Petri net environment, but this is dependent on the objectives thereof.

2.2.6 Enterprise Resource Management Systems

To help organisations manage the growing complexities in the large scale economies being faced today, Enterprise Resource Management (ERP) systems have been implemented. These help form the supporting structures that enable semi-automated operations. Brown, Vessey *et al.* (2003) explains that the main objective behind ERP implementation is the integration of the different departments within the organisation onto one central system. The centralisation of these departments enables the business processes, databases and operational data to be available without the fragmentation that usually occurs between departments.

ERP systems have been commercialised in such a manner that they can be applied in different areas of operation without much customisation. This entails that these systems come pre-packaged with modules that can be slightly altered to best suit the application area. Standard departments such as finance, sale, planning, materials management, human resources and decision support come packaged as integrated modules as discussed in Yen, Chou and Chang (2002). While standardised ERP systems come with generic business activities for most organisations, they also have attractive advantages as discussed by Robson:

- Rapid deployment and quick availability;
- Standardised business activities and practices; and
- Backed by continual support and documentation.

As with any standardisation comes the inherent loss of functionality needed for specialised organisations. Standardisation then drives up the cost to customise the system and causes some complexity increase in a system that was implemented to reduce complexity in the first place.

Looking back at the history of ERP system implementation, there have been cases where the implementation has been a costly practice. The implementation puts a lot of strain on the current management and resources, as it does not only require time but also requires the cooperation of the people within the organisation. This implementation faces the resistance to change of the personnel and their willingness to utilise the ERP system the way it was intended. Umble, Haft and Umble (2003); Al-Mashari, Al-Mudimigh and Zairi (2003) and Motwani, Subramanian and Gopalakrishna (2005) have identified the critical factors that should be addressed when successful ERP implementation wants to be achieved. While continual support of the ERP system is critical, Shanks, Seddon and Willcocks (2003) explains that not enough attention has been given to not only the maintenance of the system, but the upkeep and revisions of the ERP version. Ingvaldsen (2011) explains that the management of ERP upgrades and updates are becoming difficult as they have potential compatibility issues. While the reliability and the availability of an organisations central database is of the utmost importance, any downtime caused by these issues can not be afforded. It is here that ERP vendors need to focus as compatibility issues can be crippling to an organisation during an upgrade process. There should be no need to change the current business process layout or aspects because of compatibility conflicts.

With ERP systems such as SAP, a reference model is included that describes the best practices compatible with the ERP system. This reference model includes the entity types and process models from which business process are based. Event-driven Process Chains (EPCs) describe these processes within the SAP environment. These descriptions are visual representations of the business processes in the form of directed graphs as shown in Mendling, Moser, Neumann, Verbeek, van Dongen and Van der Aalst (2006). These business processes, while standardised, have variants added into the ERP system. An organisation can then go and implement the variant most suited to their goals. It is also possible to specify new variations but as this is a costly endeavour, most organisation would rather change their way of operation to suit the ERP system as changing the systems' code would be more difficult and would require development and maintenance costs. Brown *et al.* (2003) shows how the newer generations of ERP have moved away from a vertical business transaction focussed platform to one which is able to support supply chain functions and allow the continuous planning with a decision support framework.

While ERP systems in some cases have shown to not be critical to the organisation's continual operation, it does allow for future growth in terms of performance. In most large organisations, ERP systems are almost unavoidable as the organisation becomes too complex to manage and record transactions otherwise. In most cases then, the benefits outweigh the potential pitfalls.

2.3 Process Mining

With all the information needed buried in clusters of data, it has become more important than ever to be able extract the required information and recognise correlations and patterns. Data Mining is the first step to be taken in cases where the data volumes become too large for conventional information extraction procedures. Data Mining is defined by Larose (2014) as:

> "The process of discovering useful patterns and trends in large data sets."

Van der Aalst (2011) goes on to state that the aim of process mining is thus to take the information that resulted from data mining and use the techniques, methods and tools to make improvements to current business processes. The entire process mining activity then forms part of the business process life cycle and integrates itself as part of the evaluation and design and analysis step.

Within large organisation today, the information systems are able to store numerous types of historic data. Data types include all activities that are trigger based. This is especially true with workflow systems where the start and end of activities are captured as shown in more detail by Van der Aalst and Van Hee (2004). The greater majority of these information systems are some kind of Enterprise Resource Planning (ERP) system or a workflow management system (WFMS), all of which have the ability to store a variation of a *historical log* or *transaction log*. Moreover, with the incredible surge in processing power and integration of these systems, new possibilities of data analysis will enable the utilisation of these historical event logs.

Process mining, or business process mining, is a discipline that seeks to utilise the historical event logs mentioned above and discussed in more detail later in Subsection 2.3.2. The main idea behind process mining is to start gathering information about the current processes taking place instead of conceptually working with workflow design as stated by Van der Aalst and van Dongen (2002). The final goal is to express the analysed log in such a way as to be able to explain the inner workings of the observed processes and visualise them, in this case, by using Petri nets. In general, the processes discovered by process mining can be represented using multiple types of notation languages. This includes Business Process Model and Notation (BPMN), Yet Another Workflow Language (YAWL), Event Driven Process Chains (EPC) and Petri nets. While Petri nets have a strong mathematical foundation, the other languages are often used for their simplified presentation style.

Van der Aalst and van Dongen (2002) stipulates three assumptions that are made with regard to process mining:

1. Each recorded event in the log refers to a *task* within the business process;
2. Each recorded event refers back to a clear *case*; and
3. The recorded events are all ordered.

In some cases, it might be regarded as necessary that an ERP system or WfMS is in place to record events, but it is also possible to apply process mining to logs recorded by other means. The only stipulation is that a log must be present. The form in which this log exists influences the difficulty by which it can be mined and it is therefore preferred that it exists in a digital format and in one location.

2.3.1 Mining Procedure

The process mining procedure in most cases starts from the business intelligence (BI) efforts of the organisation. BI is involved in the organisations approach to capturing, integrating and cleaning enterprise data for decision making (Dayal, Castellanos, Simitsis and Wilkinson, 2009). The first two steps of the process mining procedure coincide with what is more commonly referred to as an ETL (Extract-Transform-Load) process as described by (Vassiliadis and Simitsis, 2009). As most organisations do not have its BI set up in such a fashion as to support process mining efforts yet, the loading in this case is first interrupted by an additional filtering process to make the data more suitable. Care should also be taken when the data in question includes personal information about employees and in some cases, the IP of the organisation. Confidentiality should be recognised and handled according to the policies that are in place.

The overall procedure of process mining can be seen in Figure 2.10. This procedure is also how the case study will be handled later on.

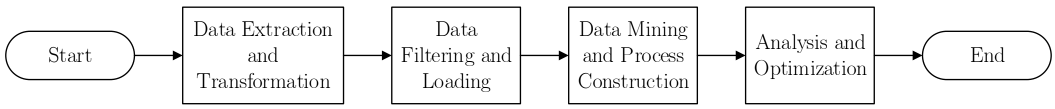


Figure 2.10: Process mining procedure.
Adapted from Van der Aalst (2011)

As with any analytical process, data gathering or data extraction is the start. Even though it might sound simple and trivial, in this case it is far from it as Van der Aalst (2011) discusses in more detail. The situation that is confronted in most organisations is that, with the implementation an ERP system the transactional data may be scattered across different departments. Ideally, it should be stored in a central location often called a data warehouse or data bank (Dayal *et al.*, 2009). When data is scattered it offers somewhat of a challenge, especially when the process being investigated calls for the collaboration between different departments. It is in cases like this where data becomes scattered across different data tables. This will then require a merging effort between the data tables.

Data filtering is part of any analytical procedure and for good reason. With data being recorded in high volumes and in some cases across different departments, there is bound to be some erroneous data. Whether this error is from the software side or some input error from the user. This might also include data that should have been organised under a different process (in this case) or missing property contents that might lead to unreliable results. Van der Aalst (2011) argues that is important to assume that the event log only contains events that truly occurred. Events that are not supposed to be included in the event log are impossible for the algorithms to detect as erroneous. "Noise" in this case does therefore not exclusively point towards data points that should not be included, but also the "outliers" which refer to infrequent events.

Programs such as *Disco* [1] offer filtering capabilities described in Günther and Rozinat (2012), such as a "timeframe filter" which allows sectioning of process data for comparison before and after a certain point in time. The "variation filter" applies to the outlier concept explained above where a certain percentage of variation on the norm can be filtered out. Van der Aalst (2011) also mentions that from a process discovery standpoint, heuristic mining, genetic mining and fuzzy mining can be used to filter out noise. Crude errors are usually corrected before loading the data into the analytics software while more intricate errors or noise can be handled by integrated algorithms.

2.3.2 Process Models and Event Logs

Fundamentally, process mining revolves around the idea of building or "discovering" process models from available workflow log data. These models are a true representation of the idealised model that the organisation set up. This also enables the visualisation of complex processes that are networked within the organisation and would otherwise have been impossible to map. When looking at a process model designed by the WfM system, there is generally no indication as to what extent this planned process has been followed. Conformance towards a planned process is crucial in cases where the organisation relies on the coherence and compliance of business processes as discussed in Becker, Delfmann, Eggert and Schwittay (2012).

By utilising historical log files stored within the information system of the organisation, it is possible to see if the business process has been executed according to what has been planned. Historical log files within the information

[1] https://fluxicon.com/disco/

system are created during the start or the end of a transaction or task. They are in essence the *events* which cause the business process to progress. These historical log files are more commonly referred to as event logs Becker *et al.* (2012); Van der Aalst and Van Hee (2004); Van der Aalst (2011). The data within this event log usually include the following attributes:

- Event ID : This will be a unique number for every activity, usually numbered sequentially as activities are triggered;
- Timestamp : This refers to the exact time and date when the event was triggered;
- Activity : A short description of the event;
- Performed by/Performer : Refers to the person who handled the triggered event; and
- Case ID : Links the activity to a certain process used for a specific case. This attribute is usually present in cases where a single log file handles all the processes instead of process specific log files (Van der Aalst, Reijers and Song, 2005*b*).

The following example is presented in Van der Aalst *et al.* (2007*a*) where an event log is given with four attributes. This event log can be seen in Table 2.2.

A control-flow diagram can then be constructed using this event log. This control-flow diagram is done from a perspective that encompasses the process on a task level. This is also referred to as the *process perspective*. When constructing the control-flow diagram from a given event log, it is important to note which attributes play a key role. For the example presented, it can be seen that there are tasks which coincide with more than one case ie. "activity A" is present for "case 1", "case 2" and "case 3". With a control-flow diagram which is only able to represent one case at a time, the "Case ID" provides a distinguishing field for which *processes* or *cases* are associated with a *task* or in this case, an "activity". The control flow diagram for "case 1" is shown in Figure 2.11.

Table 2.2: Example Event Log.

Adapted from Van der Aalst et al. *(2007*a*)*

Case ID	Activity id	Originator	Timestamp
case 1	activity A	John	9-3-2004:15.01
case 2	activity A	John	9-3-2004:15.12
case 3	activity A	Sue	9-3-2004:16.03
case 3	activity B	Carol	9-3-2004:16.07
case 1	activity B	Mike	9-3-2004:18.25
case 1	activity C	John	10-3-2004:9.23
case 2	activity C	Mike	10-3-2004:10.34
case 4	activity A	Sue	10-3-2004:10.35
case 2	activity B	John	10-3-2004:12.34
case 2	activity D	Pete	10-3-2004:12.50
case 5	activity A	Sue	10-3-2004:13.05
case 4	activity C	Carol	11-3-2004:10.12
case 1	activity D	Pete	11-3-2004:10.14
case 3	activity C	Sue	11-3-2004:10.44
case 3	activity D	Pete	11-3-2004:11.03
case 4	activity B	Sue	14-3-2004:11.18
case 5	activity E	Clare	17-3-2004:12.22
case 5	activity D	Clare	18-3-2004:14.34
case 4	activity D	Pete	19-3-2004:15.56

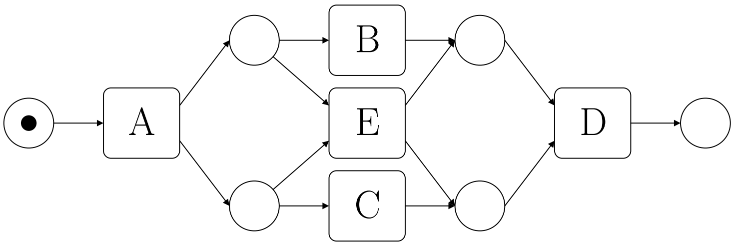


Figure 2.11: The control-flow structure drawn as a petri net.
Adapted from Van der Aalst et al. *(2007*a*)*

2.3.3 Review of Algorithms

Within process mining, the mining algorithm refers to the programming code that handles the input data and generates the process models. There are numerous algorithms that exist today as shown in Tiwari, Turner and Majeed (2008), each with advantages and disadvantages. In a holistic sense, there are three main categories into which mining algorithms fall:

- Deterministic mining algorithms;
- Heuristic mining algorithms; and
- Genetic mining algorithms.

With deterministic models, all data necessary for the outcome is known. One of the most notable properties of deterministic models is that the output is constant for given input variables. In this case, the model created from process mining will always be repeatable. According to Colaço and Dulikravich (2009), deterministic algorithms are generally faster than other methods. An example of a deterministic algorithm is the α-algorithm presented by Van der Aalst, Hirnschall and Verbeek (2002). This algorithm is mainly concerned with the ordering and relational ties of events in an event log.

The main idea behind heuristic algorithms is that in some cases a predefined algorithmic approach is unable to find an optimal solution. In cases like these it is necessary to implement an approach that aims to look for good solution, whether globally optimal or not, by trial and error as discussed in Winston and Goldberg (2004). In the case of process mining, these algorithms still incorporate deterministic algorithms but with that add Pareto principles (frequency indicative of importance) to be able to disregard paths or events that only lead to unnecessary complexity. This unnecessary complexity might be because of a negligibly small percentage of cases having to take an alternate routing that, if included, will only add complexity to the base model. These paths are however not completely disregarded as they form part of the performance metrics later on.

Genetic algorithms were first introduced by Holland (1975). These types of algorithms are premised on the fundamental principles that govern natural selection in evolutionary theory. With this, solutions to a problem are found having a arbitrary starting point and then searching for better solution while disregarding inferior solutions. The search is done by combining attributes used for the previous solutions and also introducing random variations. The procedure is laid out in Konak, Coit and Smith (2006) as a five step process:

1. Randomly generate solutions and evaluate the fitness;
2. Crossover: Generate offspring(new solutions) by using the best solution in previous step; crossover
3. Mutation: Insert random variations in the solution space;
4. Fitness Assignment: Assign a fitness value to solutions based on the outcome; and
5. Selection: Select best solutions based on fitness and use them for the next crossover step until the criterion has been satisfied.

The idea behind this process is that process models are first generated randomly and are then iteratively reduced to more satisfactory solutions by means of the *mutation* and *crossover* steps. It should be noted that the initial process models that are generated are not a representation of the event log. In cases where it is important to note the different path and the number of occurrences of that deviation, a fuzzy miner algorithm can be used to assign a weighting to the arcs based on the occurrence of that path.

With every case, it is important to identify the needs of the study. This will allow the proper selection algorithm and will ensure that the model is not under-fitted. As with all modelling processes, there is a constant trade-off which calls for the balance between the precision, fitness, simplicity and the generalisability of the model. These four modelling criteria can be described by the following (Van der Aalst, 2011):

- Precision: The model does not allow for paths which differ from reality;
- Fitness: The ability to replay exactly what happened in the event log;
- Simplicity: Reduce the complexity of the model by eliminating unwanted paths; and
- Generalisation: The model is not exclusive for a certain event log and can be used to make generalisations about other precesses.

As it is not possible to meet all these criteria at the same time, the importance of every criteria needs to be assessed before analysis. This assessment needs to be based on the scope and desired outcomes of the study set out by the organisation.

2.3.4 Process Structures

In practice, two main process structures convey, in a general sense, how work is being done. Processes in this case are characterised either as a *Lasagne Process* or a *Spaghetti Process* as stated by Van der Aalst (2011). These two names form part of the extremes of how structured a process is. Lasagne processes are characterised by a well structured layout where the tasks are well defined and the constraints form the resulting task structure. Depending on the application area, these are the main types of processes that can be automated. Spaghetti processes on the other hand are mostly unstructured. The processes do not have a preconceived notion of how activities have to be ordered to finish a process. In these cases, there are usually personnel who dictate how the process is done.

In general, applying process mining to Lasagne process does not yield interesting results in terms of process discovery as the process already has a structured foundation. Applying process discovery will only confirm what is already known. In this case, process mining becomes useful when the analysis goes beyond discovery and metric are applied to the given process. Whether the metrics are based on completion and lead times or to which extent the activities comply to the planned process model. Process discovery becomes an integral tool when applied to Spaghetti processes where the unstructured nature makes planning hard. Process discovery will then be able to discover a trend within the process and allow for greater insight into what should be planned for and where a controlled sequence could yield better performance results. From a managerial point of view, turning unstructured processes into structured ones could be highly beneficial for the organisation.

2.3.5 Organisational Mining

Considering process mining as suitable tool to analyse given business processes, it is important to acknowledge a different perspective on the problem. The organisation perspective lends itself as a more holistic view for the given operation of an organisation. It deals with not so much the process as an instance, but rather how this process was done with regards to the work handover in the form of the roles and responsibilities as a network of individuals. The techniques used for this type of analysis as an add-on for process mining owes its development to Van der Aalst *et al.* (2005*b*) and Song and Van der Aalst (2008).

The idea of using interpersonal connections stems from sociometry where interpersonal relationships were presented using graphs or matrices. These concepts are described in more detail in Burt, Minor and Alba (1983); Scott (2012); Wasserman (1994). Since the social network analysis has developed into a well researched field, it has seen numerous application in different fields with the focus in this case being the interaction between the social network and the business process within the organisation (Van der Aalst *et al.*, 2005*b*). The usual procedure of gathering information for the construction of a social network is also circumvented, as all necessary data already exists within the event logs of the ERP system in an implicit manner.

The information in the ERP system considered here contains the detail of who initiated an event. This event refers to either activities or entire cases. Van der Aalst, van Dongen, Herbst, Maruster, Schimm and Weijters (2003*b*) illustrates this in more detail. Van der Aalst and Song (2004) shows how this concept can be implemented within a process mining environment by showing how the handover of work and distribution of roles support the business process. Ferreira and Alves (2012) supports this research and goes on to add the ability to form relational clusters from these social networks.

Considering Table 2.3, Van der Aalst and Song (2004) presents an example from which they first constructs the business process and then goes on to discuss the concept of a sociogram that is based on the *handover of work*. The handover of work in this case is part of two main ideas of graphing social interaction in this context. Ferreira and Alves (2012) states that this *handover of work* entails capturing the occurrence when work by one user is followed by the work of another. This results in a graph where a directed arc shows the direction of the handover and has the weights relating the number of occurrence as can be seen in Figure 2.12. The second aspect that can be studied is when people collaborate. This resulting graph in this case has the same nodes

Table 2.3: Event Log with Personnel.

Adapted from Van der Aalst and Song (2004)

Case Identifier	Activity Identifier	Performer
Case 1	Activity A	John
Case 2	Activity A	John
Case 3	Activity A	Sue
Case 3	Activity B	Carol
Case 1	Activity B	Mike
Case 1	Activity C	John
Case 2	Activity C	Mike
Case 4	Activity A	Sue
Case 2	Activity B	John
Case 2	Activity D	Pete
Case 5	Activity A	Sue
Case 4	Activity C	Carol
Case 1	Activity D	Pete
Case 3	Activity C	Sue
Case 3	Activity D	Pete
Case 4	Activity B	Sue
Case 5	Activity E	Clair
Case 5	Activity D	Clair
Case 4	Activity D	Pete

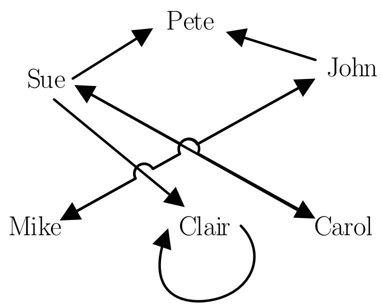


Figure 2.12: Resulting sociogram.

as the previous aspect but has undirected arcs in this case. Ferreira and Alves (2012) argues that in cases where the processes being analysed have a rigid structure, the *handover of work* approach may yield better insight but in cases where the process structure involves a highly networked setup, collaboration may yield more favourable results.

Within a networked setup, the three main structures types are *regular*, *small* *world* and *random* structures respectively (Cowan and Jonard, 2004). The network structures follow a progression from a locally ordered construct, where people (nodes) are closely connected, to a random construct where nodes are randomly interconnected and lack any pattern. This randomness is measured by p where $0 < p < 1$. This structure progression is illustrated in Figure 2.13.

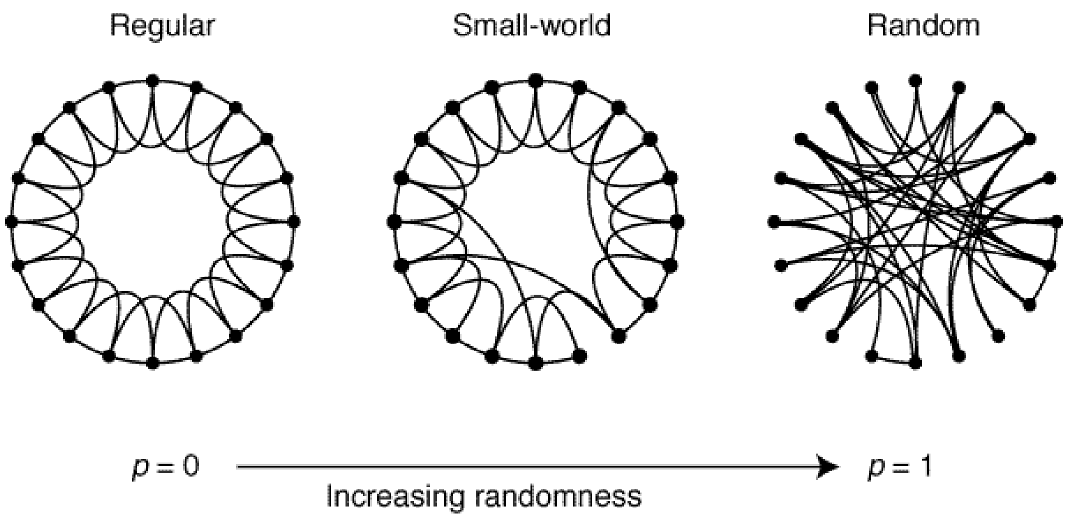


Figure 2.13: Network structures.
Source (Watts and Strogatz, 1998)

The respective network constructs shown above primarily possess two defining attributes namely *cliquishness* and *path length*. Cowan and Jonard (2004) and Watts and Strogatz (1998) describes cliquishness as the local property where nodes share the connections with other nodes while the path length refers to the global property of the average number of steps between nodes. Concerning p, it becomes intuitive that within increase in the randomness of the structure, p, the cliquishness becomes less. This is inversely true for the path length. It can also be seen that when p is very small, the network tends towards a more structured layout with long path lengths while the inverse is true for a large p value. Neither of these situations are optimal as long path lengths cause slow communication but the process continues longer before communication saturation occurs. With short path lengths, there is a rapid diffusion at the beginning but the process saturates very quickly. Cowan and Jonard (2004) showed that there is an optimal solution where the structure resembles a small world configuration ($p \approx 0.09$).

To build these structures from an event log, clustering algorithms are used. These algorithms as discussed by Ferreira and Alves (2012); Van der Aalst *et al.* (2005*b*), where nodes are grouped together based on specified metrics. This aggregation of nodes can of course be based on the two metrics discussed above. The iterative approach ultimately measures the similarity between nodes and then builds a hierarchical cluster where nodes are paired together. Murtagh and Contreras (2011) discusses numerous of these clustering algorithms. Most notably, he discusses the "average linkage" function that builds on the average similarity between nodes given by:

$$D(c_r, c_s) = \frac{1}{n_r n_s} \sum_{i \in c_s} \sum_{i \in c_s} d(i, j) \tag{2.3.1}$$

Where c_r and c_S denote the clusters, n_r and n_s the nodes and then $d(i, j)$ would represent the nodes referred to by node i in c_i and node j in c_s. Also, given an adjacency matrix A_{ij} which contains the weights of the arcs pointing to individual nodes, the equation above simplifies to $D(c_r, c_s) = d(i, j) = A_{ij}$ for the first iteration. Fire, Puzis and Elovici (2013) demonstrates the use of a clustering algorithm that constructed a social network based on data collected online from web based communication. An example of how such a network might look can be seen in Figure 2.14 where the colours denote the different clusters which formed because of the tie strength between the nodes.

There is however a trade-off which should be considered. Coming from a number of nodes with different linkages, one extreme will place all these nodes within a cluster while on the other side, none of them will be joined within the cluster. Fire *et al.* (2013) does point to authors presenting possible solutions in cases like these. Lv, Su, Wang and Zuo (2005); Han and Narayanan (2007) presents a criterion which can be set to stop the algorithm after the criterion has been met while Milligan and Cooper (1985); Jung, Park, Du and Drake (2003) presents a solution which looks at the number of clusters being formed and then terminates the process when a number has been reached. *Modularity* arises as a method from which the cluster can be constructed in such a way as to ensure that the clusters are well described. With this, densely populated groups can be formed while still maintaining the loose connections with other clusters. *Modularity* is calculated using the following equation:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \times \delta(c_i, c_j) \tag{2.3.2}$$

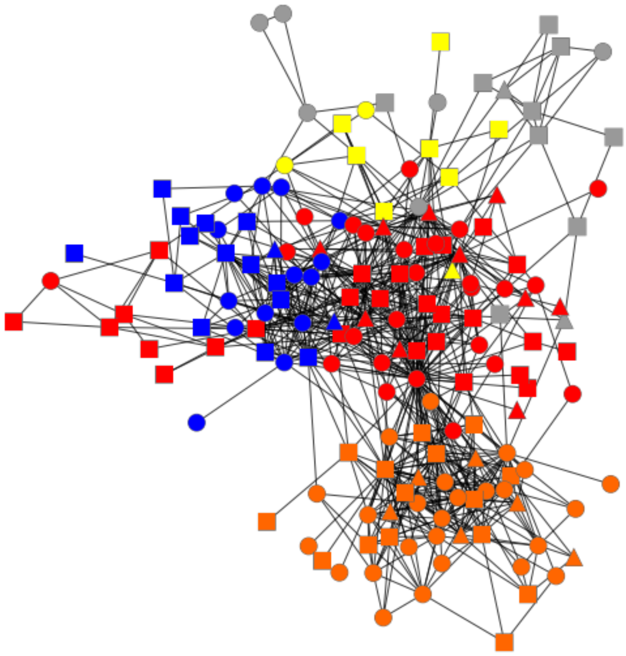


Figure 2.14: Network clusters.
Source Fire et al. *(2013)*

Where the summation is done with regard to each individual arc and m is the total sum of arc weights, k_i, the *degree*, denotes the sum of arcs coming from that node, also referred to as the nodes emanating from it. Also, if nodes i and j are grouped within the same cluster, $\delta(c_i, c_j) = 1$ otherwise zero. Clauset, Newman and Moore (2004); Girvan and Newman (2002) has demonstrated that modularity can be used on numerous different types of network structures of varying sizes. Clauset *et al.* (2004) presents an example where maximum modularity is used to successfully cluster nodes as can be seen in Figure 2.15.

Ferreira and Alves (2012) uses a small network of people demonstrate the iterative process of modularity and visualises this process with the use of a dendogram, shown in Figure 2.16, and a graph showing the value of modularity with each iteration.

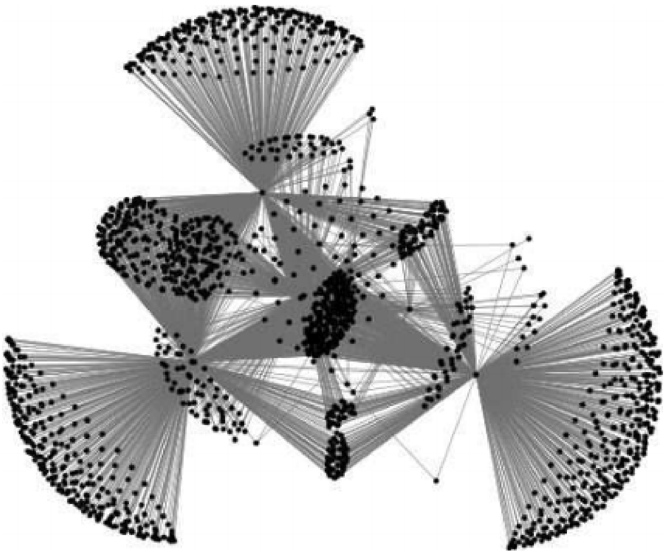


Figure 2.15: Community constructed using maximum modularity.
Adapted from Clauset et al. *(2004)*

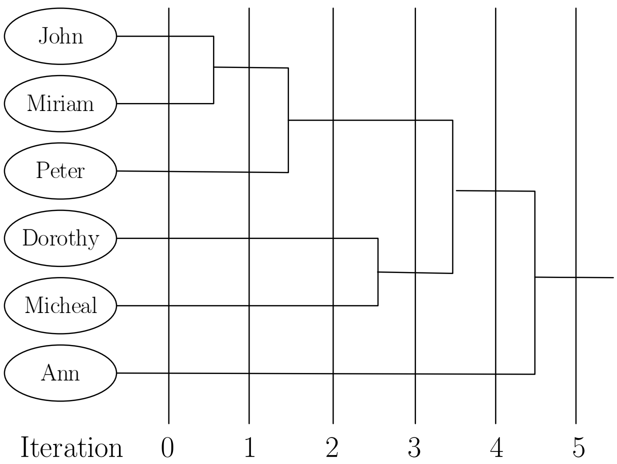


Figure 2.16: Dendogram.

2.3.6 Performance Measurement

Considering the performance aspects of an actual BP, a formal approach called Process Performance Analysis (PPA) is used as an improvement platform. This discipline stems from the roots and foundations laid out by general performance measurement, process monitoring, business activity monitoring and business performance management as discussed by Hornix (2007). There is however a debate over the term performance in this case. While the measurement itself is easy, it is the objective relation to what is actually good and what should be considered a viable metric that is debatable.

2.3.6.1 Quantitative Analysis

Quantitatively, performance is measured by performance indicators (PIs). These PIs are related to the definition of performance which can be defined by, as stated by Douwe P. Flapper, Fortuin and Stoop (1996): "the degree in which an organisation carries its objectives into effect". This definition is re-enforced by other authors such as Lebas (1995). As objectives change with the consideration of different types of organisations, the target measurement of the PIs change. PIs can cover metrics such as time, cost or quality depending on the importance for management and the objectives of the organisation. When an organisation's PIs cover a broad range of metrics, Key Performance Indicators (KPIs) are often identified. These PIs relate back directly to the main goals and strategy of an organisation and are critical for the the success of that organisation as discussed by Reh (2015). Kaplan and Norton (1996) goes on to show how a Balanced Scorecard (BSC) is used to derive KPIs from the BM and strategy of an organisation.

While the BSC can offer the derivation of KPIs relating to different perspectives of the organisation, PPA builds KPIs based on a BP perspective. In essence, these KPIs are only concerned with the contribution that processes have towards the goal of the organisation. These KPIs can highlight problematic processes effectively when goal values have planned or estimated values based on estimation calculated during the modelling phase of the BPM life cycle. These KPIs are not limited to the complete running time of a process as it can also indicate where in the process the deviation are caused. As Hornix (2007) argues, PPA can thus become extremely useful in overall risk analysis, planning scenario outcomes and fixing problematic process areas.

One of the methods to analyse a given event log of a business process is presented by Adriansyah, van Dongen and Van der Aalst (2011). The technique discussed is based on the fitness of conformance which is also supported by Rozinat and Van der Aalst (2008); De Weerdt, De Backer, Vanthienen and Baesens (2011). The fitness in this case is a measure given numerically or otherwise to indicate the level of conformance of processes in reality towards the planned business processes. The conformance in this case primarily points towards the sequential and inclusive nature of the tasks within the process.

Adriansyah *et al.* (2011) goes on to identify that the conformance can be indicative of either skipped or inserted tasks. Therefore, tasks can either be included in the original process and then never have been done or unwanted tasks could have been added. It should also be noted that the severity of either of these cases on any task is not always constant. This thus requires some insight into the process and the organisation to be able to make informed recommendations according to the type of task and its impact. It is also possible that the impact of deviating from the desired process model can be valuable in the sense the model is incompatible with the workings of the real world. In this case, adjustments to the original process model needs to be made.

Another metric that can be used to indirectly measure the performance of a business process is has been proposed by Cardoso (2006). This entails using a complexity metric called the "control-flow complexity" (CFC) metric. This metric focusses on analysing the number of XOR, OR, and AND-splits within a business process. While this is ultimately a very trivial metric, it does offer a useful perspective when considering the redesign of current business processes. The CFC metric is simply given by Equation 2.3.3:

$$CFC(P) = \sum_{a \in P, a is a xor-split} CFC_{XOR}(a) + \sum_{a \in P, a is a or-split} CFC_{OR}(a) + \sum_{a \in P, a is a AND-split} CFC_{AND}(a) \quad (2.3.3)$$

Where a represents the activity and P the given process. It is therefore obvious that the more there are of each attribute, the more complex the process. Even though it can be seen as trivial, the CFC metric does give an easy metric that can be used to objectively measure the complexity within a structural dimension. That being said, it does have the disadvantage of not being able to incorporate the data complexity or the complexity of resources assigned to a specific process.

Cardoso, Mendling, Neumann and Reijers (2006) describes an adaptation of a metric used by Halstead (1977) to measure the complexity of software. This metric is a function of the variable, constants and operators within a certain piece of code that then leads to a quantitative measure of the program's complexity. In its original form, the metric has four variables:

- n_1 = The amount of uniquely identifiable operators;
- n_2 = The amount of uniquely identifiable variable and constants;
- N_1 = The amount of operator occurrences;
- N_2 = The amount of variable and constant occurrences.

Taking this concept and applying it to business processes, the operators now become the activities, splits and general control-flow elements and the variables become a measure relating to data variables handled within the process. The process length, difficulty and volume can now be estimated by the following:

$$N = n_1 \times \log_2(n_1) + n_2 \times \log_2(n_2) \tag{2.3.4}$$

$$D = (\frac{n_1}{2}) \times (\frac{N_2}{n_2}) \tag{2.3.5}$$

$$V = (N_1 + N_2) \times \log_2(n_1 + n_2) \tag{2.3.6}$$

Where Equation 2.3.4 gives the process length, Equation 2.3.5 gives the process difficulty and Equation 2.3.6 gives the process volume. Cardoso *et al.* (2006) also states that the advantage of using these metrics is that they are trivial in a sense that they do not require in depth understanding of the processes. These metrics are able to predict the errors and maintenance needs of a process given past experience of the same metric value. They are also universal across different modelling languages.

Cardoso (2008) shows an example where the CFC metric is not only used to quantify the complexity of a given process, but also how different levels of complexity are translated into a risk profile. The idea behind this can thus be to rank different processes by this risk ranking and then attribute this to the need to constantly monitor that process. This risk can then be combined with a cost metric to determine the criticality of a given process in an organisational sense.

2.3.6.2 Qualitative Analysis

With the occurrence of audits (internal or external) in large organisation to ensure adherence to policies or procedures, Ramezani, Fahland and Van der Aalst (2012) describes how process deviations may be costly towards the organisation. It is therefore necessary to check the compliance towards the guidelines set out by regulatory bodies, managers and stakeholders. Ramezani *et al.* (2012) also mentions that deviation may be indicative of malpractice, fraud, risk indicators and process inefficiencies.

With the existence of either checking the compliance of a process beforehand or after the completion of a process (forward compliance checking and backwards compliance checking Ramezani *et al.* (2012)), the focus here will fall on the latter. This is due to the nature of process mining and the reliance on event logs. Ramezani *et al.* (2012) presents a framework used to test the compliance of a business process. This framework is based on research and work done by Van der Aalst, Van Hee, Van der Werf, Kumar and Verdonk (2011); Elgammal, Turetken, van den Heuvel and Papazoglou (2010); Awad, Decker and Weske (2008); Dwyer, Avrunin and Corbett (1999); Gruhn and Laue (2006); Schumm, Leymann, Ma, Scheibler and Strauch (2010); Giblin, Liu, Müller, Pfitzmann and Zhou (2005); Schleicher, Grohe, Leymann, Schneider, Schumm and Wolf (2011*b*); Schleicher, Fehling, Grohe, Leymann, Nowak, Schneider and Schumm (2011*a*). Within these sources, there are 50 compliance rules that all govern the control-flow of processes. The text then goes on to categorise these 50 rules into 15 formal groups shown in Table 2.4.

Applying compliance checking to BPs can either have a reactive or pro-active effect El Kharbili, de Medeiros, Stein and Van der Aalst (2008). While the former is done after the execution of BPs during an audit stage while the latter is done prior to the completion of BPs. This is usually part of the design stage and during the execution of BPs, aiming to detect non-compliant behaviour before they occur. These two scenario's can be categorised under either forward compliance checking (pro-active) or backward compliance checking.

Forward compliance checking can be broken down into two subgroups where checking is first done in the model design phase and then during the operation of the process. In the design phase, there are approaches that enable the

Table 2.4: Compliance Rules

Category	**Description**
Existence	Overviews the presence or absence of tasks
Bounded Existence	Limits the number of occurrences
Bounded Sequence	Limits a sequence of events
Parallel	Binds the simultaneous occurrence of tasks
Precedence	Limits the precedence of a task over another
Chain Precedence	Limits the precedence of a sequence of tasks over another
Response	Limits the reactant task in response to another
Chain Response	Limits the reactant sequence of tasks in response to another
Between	Limits the inclusion of a task in a sequence
Exclusive	Limits the exclusion of a task in a sequence
Mutual Exclusive	Either/or logic that is applied towards tasks
Inclusive	A tasks is accompanied by another
Prerequisite	If a certain task is absent, another also has to be
Substitute	One tasks replaces another
Co-requisite	Either tasks exist simultaneously or both are absent

designer to build models complying to policies and regulation. There are also techniques that allow the designer to revise models and eliminate problem areas. These techniques are summarised in Table 2.5. During the execution of BPs, run time compliance checking is used to monitor the states of BPs and regulate the compliance in real time. These type of compliance systems usually work by marking compliance states and are triggered as soos as a BP does not conform to that state. Lastly, backwards compliance checking uses a finished processes and then measures the magnitude of compliance towards a given set of rules or policies. Van der Aalst, De Beer and van Dongen (2005*a*) proposes the use of the LTL (Linear Temporal Logic) Checker. This procedure uses a set of LTL rules and matches them against a set of processes. It is then able to separate processes that do not conform to the compliance rules from the ones that do. As mentioned in El Kharbili *et al.* (2008), this approach is less suitable for business analysts as there is no graphical representation of compliance rules.

Table 2.5: Compliance checking techniques

Author	Technique
Ghose and Koliadis (2007)	Presents an approach called compliance patterns where defined BP models have been proven for compliance rules and are used for comparison.
Padmanabhan, Governatori, Sadiq, Colomb and Rotolo (2006)	Presents a framework where processes are seen as social interactions and contractual relationships are governed according compliance rules.
Governatori, Milosevic and Sadiq (2006) Milosevic, Sadiq and Orlowska (2006)	Involves regulation of contract documentation used in BPs and how they relate to business contract compliance.
Namiri and Stojanovic (2008)	Formalises a procedure for creating control patterns which govern compliance rules. These control patterns can then form a generic tool set for ensuring compliance.
Schmidt, Bartsch and Oberhauser (2007)	A semantic approach is taken to compliance checking by designing an ontology which is integrated into the BP models and ensure governance.

2.3.7 The ProM Framework

The ProM framework is an academic tool which is available as an open source platform on which process mining algorithms can be implemented (Van der Aalst, van Dongen, Günther, Mans, De Medeiros, Rozinat, Rubin, Song, Verbeek and Weijters, 2007*b*). This is part of which makes it such an attractive option for academic work. The open source nature of ProM has allowed numerous contributors to create plug-ins for ProM. These plug-ins allow users to add to the ProM framework without decompiling any of the original code and thus add functionality as needed. Currently ProM is in its sixth iteration that brings much needed improvements over previous versions. Verbeek, Buijs, Van Dongen and Van der Aalst (2010) argues that previously plug-in support became a drawback in a more mature process mining environment where commercial products offered a more streamlined experience without the vague separation of what is part of the process mining algorithm and what is part of the original user interface. This has been addressed in version 6.

On the input side of the ProM framework, the event log discussed earlier would

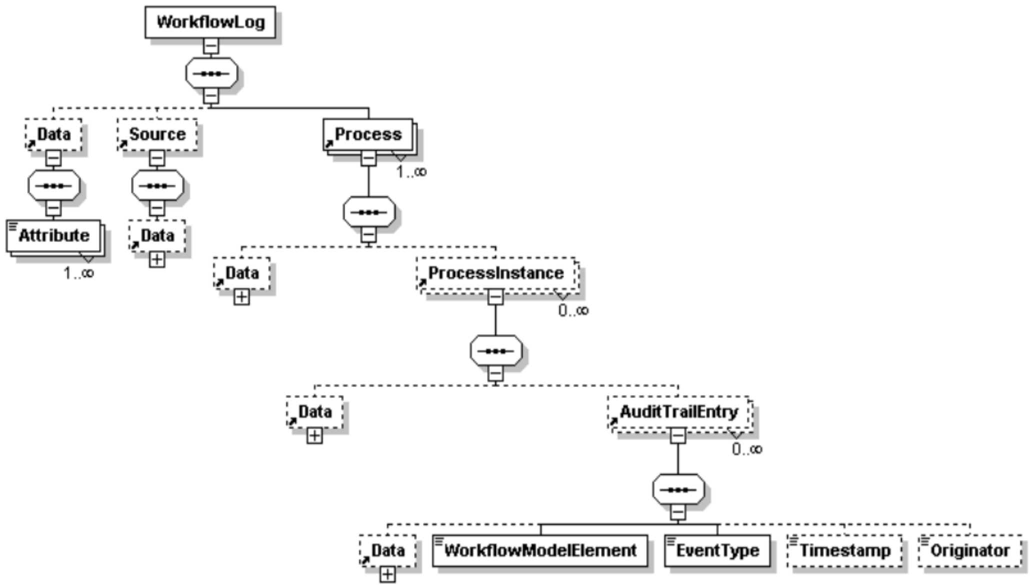


Figure 2.17: The XML schema.

be the data input. ProM can handle numerous software package exports such as Staffware, Oracle, SAP etc. To enable this, the mining extensible markup language (MXML) was developed. This language is based on the XML schema of which the structure is show in Figure 2.17. This format was built with ProM in mind and is thus the preferred input schema. MXML can be generated from any type of historic event log, whether that is transaction logs, audit logs or event tracking in business processes.

One of the most appealing aspects of plug-in support is the ability to add a single needed functionality without it affecting the rest of the framework. For example, the α-algorithm was added as a plugin to enable the construction of petri-nets from event logs as discussed in Van der Aalst *et al.* (2005*b*). There are six categories of plug-ins available as shown in Verbeek *et al.* (2010):

- Mining plug-in: These mining algorithms are mainly used in the discovery process. This includes the α-algorithm and network discovering algorithms;

- Export plug-in: These allow a save function to be implemented. These plug-ins support numerous file formats to export the discovered or analysed process;

- Import plug-in: As data sources often vary, import plug-ins are implemented to allow compatibility between ProM and different sources;
- Analysis plug-in: As follow up on the discovery process, different analysis techniques can be implemented to give insight into the event log in question;
- Conversion plug-in: A small but necessary plug-in where compatibility issues are avoided. With these plug-ins, data formats can be translated into a target format;
- Log filter plug-in: As part of any data mining endeavour, there needs to be data filtering to some extent. This can include cutting out noise (unwanted data points) or simply establishing boundaries for the analysis within the data set.

Within each of these categories, there exists numerous plug-ins. This allows the user to select individual plug-ins for specific cases or a collection of different plug-ins to achieve the desired result. This is ultimately up to the user and is somewhat reliant on their knowledge and goals. While using any platform for process mining, the goal is usually to perform not only process discovery but to involve process analytics.

2.4 Chapter Concluding Remarks

The literature presented above first discusses the application area and presents the theory and background of the problems within the field of PAM. This facilitated a better understanding of the problem stated in Chapter 1 and allowed an informed formulation of the solution. This also allowed the contextualisation of the problem and the application are within PAM for process improvement.

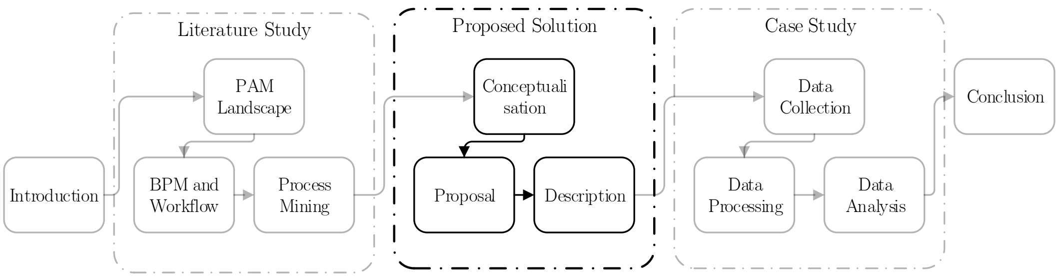
The discussion on PAM highlighted the need for the improvement of the enabling processes which support and allow the value creation of physical assets. This is important for the execution of the AM strategy as the decisions shown on different levels rely on the compliance of planned processes. It also shows that the information necessary for the mapping and improvement of processes in the PAM environment is in most cases already available and doesn't require addition systems to be implemented. With the focus on the processes within PAM, BPM and Workflow are presented as an effort to achieve an interdisciplinary approach. Responsibility of process improvement in this case falls on process mining where it is introduced as a practical analysis tool which can be applied to a process rich environment and offers the tools necessary to analyse and improve given processes. The discussion is then concluded with the performance measures used in the improvement process.

The chapter ends with a discussion on the software framework in which the process mining methodology is applied. This framework forms the centre of collaboration between the information system in which PAM data exists and the process mining methodology which mainly exists within the BPM and Workflow environment.

Chapter 3

The Process Mining Application Methodology and Requirements

This chapter aims to develop the application methodology for process mining within the Physical Asset Management (PAM) environment. This application methodology first aims to position process mining within the PAM environment and from there sets out to improve a supporting process by discovering the real world behaviour and developing metrics which can then be used in improvement.



Chapter Outcomes
•Understanding of the positioning of the application.
•Apprehension of the consideration involved with scoping and boundary setting.
•Familiarity of data collection procedures and data structures.
•An approach for the application of process mining.
•Understanding of analysis techniques.

3.1 Introduction

Chapter 1 presents the area of opportunity where there is a lack of process mapping and improvement involved with the current PAM strategy. It is then theorised that with the addition of principles from the Business Process Management (BPM) and Workflow Management (WfM) domains, this opportunity can be addressed. The opportunity is further examined in the literature study where process mining is proposed as a bridge between mapping the PAM processes and then using BPM and WfM for analysis.

This study aims to develop the application methodology by which process mining can be applied to existing processes within the PAM environment and then analysed. This application methodology should ultimately enable the organisation to make informed decisions based on the results obtained. The goal of improved decision making in the management of processes is to improve the value obtained by supporting the physical assets of the organisation.

The application methodology presented here is conceptualised as a result of the literature reviewed in Chapter 2. It utilises the Process Mining Manifesto discussed in the next chapter as a starting point for best-practices. The application methodology first addresses the target process on which the application is based within the PAM strategy and the considerations which should be addressed when considering implementation. Numerous analytical techniques are reviewed in Section 2.3 out of which the most appropriate ones selected for use in the application methodology. This chapter gives an overview of the developed application methodology by first compiling the phases and steps within the application methodology and then discussing each step. The chapter conclusion then addresses how the results of the analysis support the decisions made within the organisation.

3.2 The Process Mining Manifesto

The overall method which will be followed is in accordance with the process mining guidelines set out by the IEEE Task Force on Process Mining (Van der Aalst and Dustdar, 2012). The "Task Force" was established to support the growing need for research and awareness surrounding log based activity analysis. The members consist of representatives from software vendors, consultancy firms and universities and serve the purpose of promoting the development, research and awareness of process mining (IEEE, 2015). These guideline are set out in the "Process Mining Manifesto" (Van der Aalst, Adriansyah, de Medeiros, Arcieri, Baier, Blickle, Bose, van den Brand, Brandtjen, Buijs *et al.*, 2012).

The Process Mining Manifesto describes a possible project life-cycle for a typical process mining case study which is shown in Figure 3.1. While two different projects will never be the same, the presented life-cycle does create a comprehensive framework for a typical case. This framework does not only conclude the steps, or "stages", of the project but also includes the activities and dependencies. The Process Mining Manifesto presents General Practices (GPs) which serve as guidelines for the application of process mining. These GPs are given below.

GP1: Event Data Should Be Treated as First-Class Citizens

To satisfy this guideline, it is required that the data used for process mining be trustworthy. With this, it should be reasonable to suppose that the events recorded within the event log are reliable and accurate. The event log in question should also be complete with regard to the predetermined scope. Any privacy and security issues should have been addressed before the process mining process starts.

GP2: Log Extraction Should Be Driven by Questions

As many Enterprise Resource Planning (ERP) systems are not yet able to fully support a process mining endeavour as a standard operation, some data mining needs to be applied to the transactional data created by the ERP system. With ERP systems such as SAP where there are thousands of tables with transaction data, it would be unproductive and unwise to start such a procedure without knowing what the outcome should be. The questions should thus be clear to know what data entries could be of use and thus, what to look for.

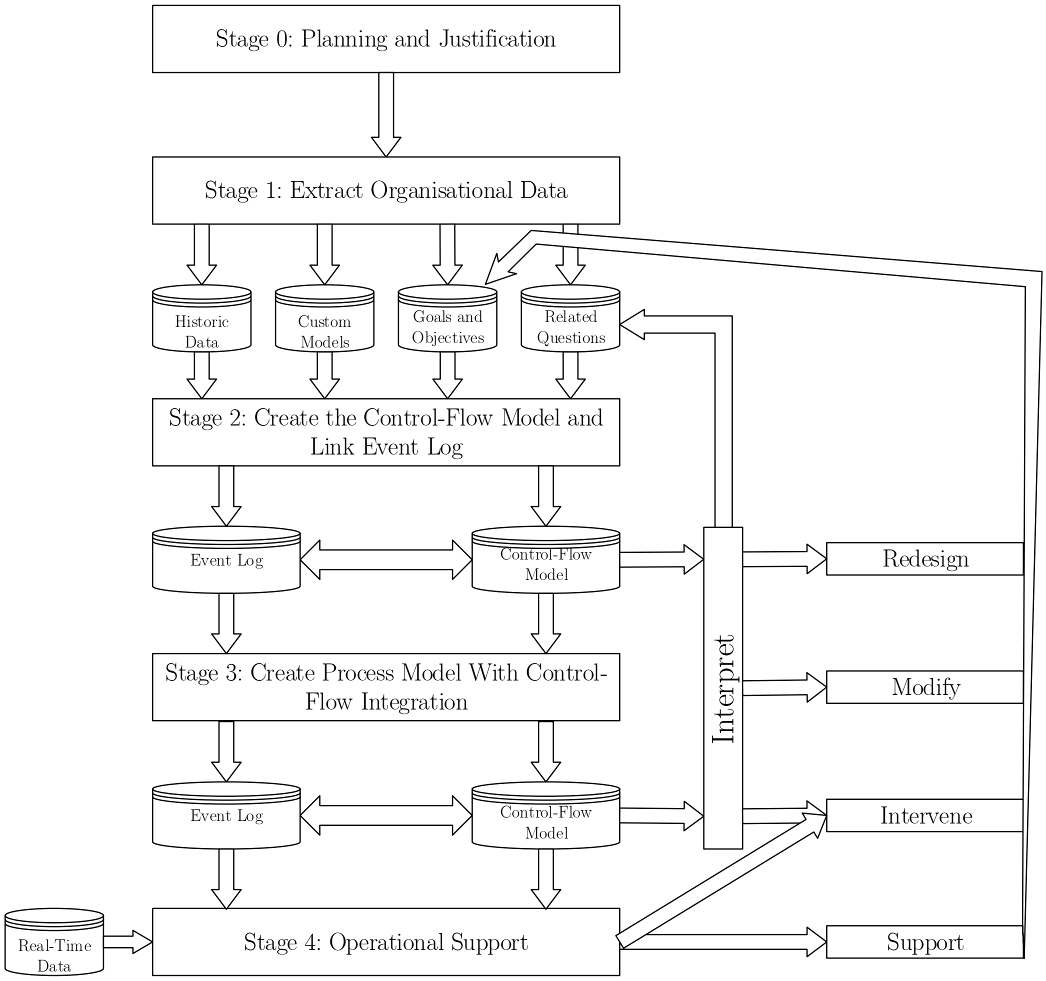


Figure 3.1: Process mining project life.
Adapted from Van der Aalst et al. *(2012)*

GP3: Concurrency, Choice and Other Basic Control-Flow Constructs Should be Supported

As the process mining effort ultimately models the reality of the processes within an organisation, it is important to choose a modelling language that supports and is able to represent the full detail of the process. In this case, Petri nets are used for the control-flow representation and is able to incorporate all fundamental workflow constructs such as sequencing, parallel routing (AND-joints), choice (XOR-joints) and loops. The process mining techniques should be in a position to address these constructs observed within the event log.

GP4: Events Should Be Related to Model Elements

When considering an event log, the tasks in the recorded process should clearly relate back to the original process model. This establishes a relationship between the log and the process to enable conformance and compliance checking and enhancement. The relationship is utilized to set up a "replay" model from the event log which can then be used for analysis. Any deviation from the original model can be highlighted and used for analysis.

GP5: Models Should Be Treated as Purposeful Abstractions of Reality

As with any modelling endeavour, the model constructed from a given event log should be viewed as an abstraction of the real world. The view given by this model should therefore be considered and brought into consideration when lessons are supposed to be learned. There is also no one correct view of the event log as different views convey distinct meanings.

GP6: Process Mining Should Be a Continuous Process

Process mining should be considered not as a once off fix for all organisation problems concerning processes but should rather be a tool within the business process life-cycle. As an organisation is a dynamic operation where variables change with respect to time, process mining should be able to accompany this change with continual support.

3.3 Phases of Application

The process mining application methodology developed in this study aims to address the improvement and shortcomings shown in Sections 1.4 and 1.3 that are concerned with supporting processes within the PAM environment. The idea behind applying process mining, as previously mentioned, is to improve physical asset performance by improving the processes in which the assets achieve value and the processes that support the value creation process. The proposed application methodology is shown in Figure 3.2 as a flow of activities to ultimately obtain the desired results.

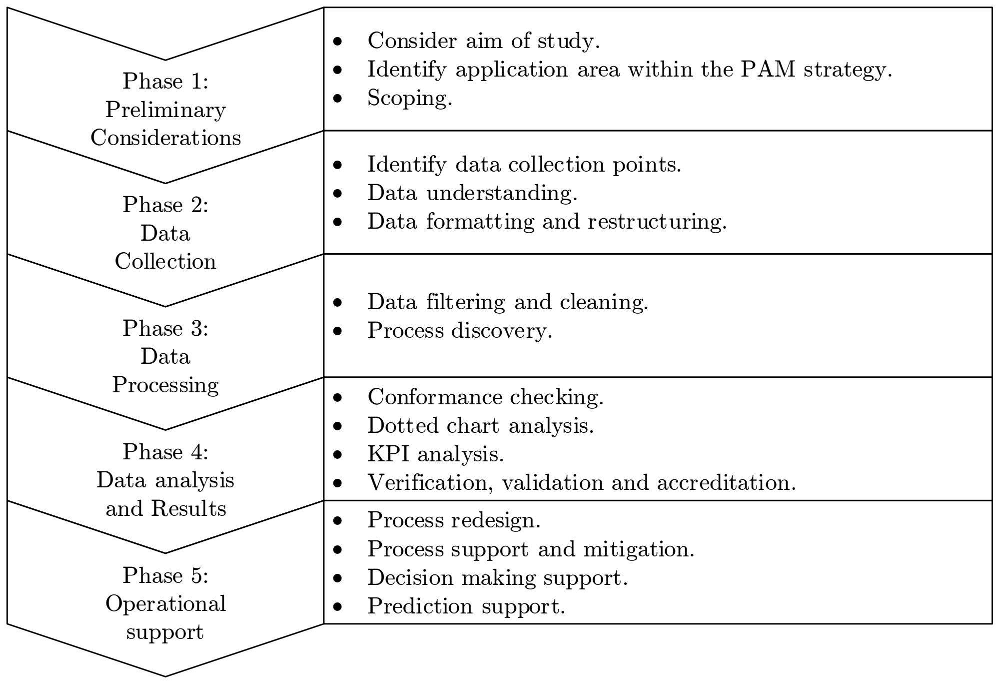


Figure 3.2: Phases of application for the application methodology.

The application methodology consists of five phases, each broken down into discrete steps. The steps within each phase act as the logical flow of activities to accomplish the goal of each phase. The five phases are represented on the left of Figure 3.2 as downward arrows. Each of these phases has the steps within the phase listed on the right. In each phase, all the steps are met before moving on to the next phase. The remainder of this section provides a brief overview of each phase before describing the steps involved.

3.3.1 Preliminary Considerations and Scoping

Phase 1 is primarily concerned with aligning a given problem in such a way as to allow the application of process mining. With this, goals need to be set that are considered realistic and achievable with process mining. The goals should also be considered achievable by the organisation. The scope of the application process has to be carefully considered as it will have a significant influence on what is possible.

3.3.1.1 Study Aim

In the context of this thesis, process mining is applicable to the field of PAM. While there are alternative ideas on the outcomes of process mining, the intended outcomes here are to map and analyse the enabling processes within the PAM strategy. This goal is related to the research question set out in Chapter 1.

The enabling processes mentioned here refer to the processes that support the value creation objective of physical assets as discussed in Section 1.2 and Subsection 2.1.2. The application methodology presented here thus aims to lay forth a methodology by which the goal can be reached.

3.3.1.2 Application Area within PAM

The application area for process mining is an essential component to examine. The focus here should first ensure that the organisation where process mining is being applied is conducting PAM and has a PAM strategy. With this constrain satisfied, the business processes operating within the PAM strategy should be considered as viable application areas. The business processes here refer to the control-flow perspectives where the processes can be broken down to their respective tasks. The nature of process mining also require that the entire business process be recorded in the information system in some form of historical logs. This can include transaction data or indications when the business process has had a state change.

3.3.1.3 Scoping

The focus in this thesis is the implementation of process mining within the organisational boundaries of the targeted organisation. This is mainly due to the information system not recording any other events other than those applicable to the organisation in which the process operates. This however does not imply that other organisations do not have any impact on the business process in question. When considering processes for process mining, it is often beneficial to have general insights regarding the events and people engaged in the process. In some cases, this might be as trivial as acquiring the process details from the planning department. Some cases might require that the analyst put the process details together by gathering information from the people involved across different departments, as an overall view of the process is not available. Details on specific parts of the process are also helpful in cases where waiting times and lead times are dependent on external variables and if the process is being logged manually by personnel or automatically by the information system.

As part of the scoping process, it is often beneficial to determine the goal of the process mining endeavour as part of the process selection exercise. As described by Van der Aalst (2011), there are three types of objective orientations:

- Data-driven: No direct goal but relies on the analyst to discover valuable insights;
- Goal-driven: By targeting specific Key Performance Indicators (KPIs) the process mining application gains a direction on what to investigate and improve; and
- Question-driven: Management in some cases have a desire to answer a question regarding the behaviour of processes or tasks. In these cases, the processes are investigated to gain insight that can then be used to answer the proposed questions.

Data-driven projects are often not preferred as they endorse the exploration of numerous processes with no apparent goal. This might lead to unnecessary expenditures with little gain towards asset performance as the information gained might be of no value. It might however help with decision making on a managerial level. While Van der Aalst (2011) clearly differentiates between goal and question-driven objectives, in a practical sense a combination thereof might offer the most value. From a goal-driven perspective, the goal might be to reduce costs or waiting times on a task/process while a question-driven perspective might ask, where is it possible to reduce costs. It is often the case that KPI related goals starts with some kind of question from a managerial point of view while the goal-driven perspective takes the question and clearly defines the outcome for the analyst.

Scoping should include the boundaries of the analysis which in this study is based on the type of data present and the detail thereof. The data often dictates which types of analysis techniques are suitable to achieve desirable results. Some of the KPI analysis require the presence of timestamps with each record and social network analysis as discussed in Subsection 2.3.5 requires roles and responsibilities to be present within the data. Social network analysis should be avoided in most cases, unless special precautions have been set up. This is not attributable to the limited usefulness but rather the nature in which most ERP and information systems capture information about completed tasks. In most cases, the system captures the person responsible for the transition change of the work item (task). This does then not offer any more information with regards to which personnel were responsible for the given task or even if the person logging the transition had any part to play in the task. For this to be of any value, additional information is thus needed which might be available in supplementary documentation or some specialised information systems.

3.3.2 Data Collection

After the problem has been fully understood and the goals have been aligned with the scope and expectations, phase 2 can be initiated where data collection points are identified. The steps of phase 2 and the current progress of the methodology are shown in Figure 3.3. As data for process mining can come from different sources, the extraction thereof needs to be considered. Data extraction then needs to be followed by transformation that will transform the raw data into a usable format. This needs to be in line with what the process mining application requires concerning formatting. Data collection might require some filtering depending on the situation and the expected outcomes. Loading as described in 2.3.1 is then done which will then create the starting point for the next phase.

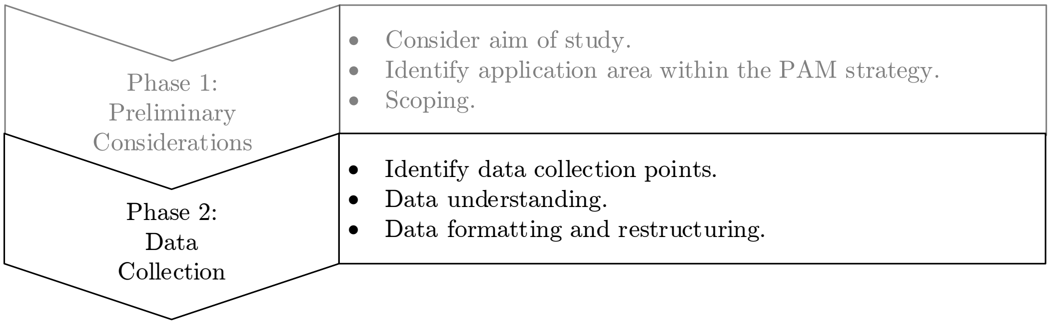


Figure 3.3: Application phase 2.

3.3.2.1 Data Understanding

The complexity and challenges coupled with data collection for process mining varies greatly depending on the organisation. It is not always the case that the data required is stored in one location, even if the scope is only set to include inter-organisational process mining. In some cases, different departments store data in separate locations other than a central data warehouse. This is often the case when an ERP system is not present. The inclusion of an ERP system does however not guarantee that the data is structured within a single location. SAP, for example, includes more than 10000 tables in which the data is fragmented. Methods for extracting data from such sources are discussed in Van Giessel and Jansen-Vullers (2004) and Piessens (2011).

There are further cases where transactional data is not directly stored where it can be extracted in a usable format. In practice, data can be embedded in emails, physical documents and invoices. Often there is also missing meta-data that includes absent timestamps. All this is assuming that the recorded data is accurate and reliable to begin with. This is however not often a problem when the capturing of transactional data has been automated by the organisation's information system.

The problems and challenges faced in applying process mining are mostly unique with regards to each organisation. Whichever the challenges might be, it is the responsibility of the data miner/analyst to take the data available and convert it into an event log with the most accurate representation of reality possible.

With scoping complete, a process deemed suitable for process mining should already have been identified. It is however not preferred to gather all data concerning the targeted process as it might lead to an overflow of data and misleading results. There are three main considerations to keep in mind. First, process mining can either be done on strictly historical data (completed processes), real-time data (ongoing processes) or a combination of both (Van der Aalst, 2011). In general, strictly historical data is utilized to construct models and conduct analysis based thereon while real-time data is often analysed to predict completion time. As the data set for both differ, it should be considered before extracting process data.

The second aspect to consider is the time domain of the analysis. In process mining literature, the time domain is often referred to as the timeframe of the analysis. The timeframe becomes an important aspect for consideration as it is often not necessary and not desirable to simply select the entire log for analysis. This can often lead to too much data which increases the complexity of the analysis and places strain on the analytics software. There are further cases where the process changed over the course of the implementation period. The timeframe should therefore only be inclusive of the desired process path. Fluxicon, a company specialising in process mining, gives Equation 3.3.1 for an estimated timeframe considering a specific process Rozinat (2015):

$$timeframe = Averagecasecompletiontime \times 4 \times 5 \tag{3.3.1}$$

While Equation 3.3.1 is used as a general guideline when selecting the timeframe to be used when extracting processes, there are more approaches to this problem. Van der Heijden (2012) adds that more ways to filter events are based on:

- A fixed time domain;
- Processes which started within a certain time domain;
- Processes which ended within a certain time domain;
- Processes with their start and end events in a time domain; and
- Processes which have intersecting events based on a time domain.

Van der Aalst *et al.* (2007*a*) shows that another important aspect is the level of information embedded within the event data. As this thesis focusses on the control-flow perspective of business processes, the information required about the events include: (1) the type of activity, (2) the case it belongs to and (3) the timestamp. Other information that can be added to events is dependent on the research objective and should be guided by the questions the research aims to answer. Adriansyah *et al.* (2011) shows the addition of a cost metric attached to events to include a financial aspect while Van der Aalst *et al.* (2007*a*) shows the addition of resources. The resources act as the person or asset responsible for the completion of the activity which can then be used to construct sociograms related to the handover of work as shown in Subsection 2.3.5.

3.3.2.2 Event Log Structure

Building event logs from the data available within the organisation is part of the most important steps in process mining. This is primarily because everything that follows in the methodology from here on is based on the event log constructed in this step. It is thus crucial to understand how event logs are structured to be practically useful. Given Figure 3.4 the conceptual view of how an event log is structured with how data points are set up in relation to one another. The conceptual view also includes the hierarchy of information.

When looking at a process from the highest level, it can be seen that the process definition is first identified. The definition specifies which activities belong to the process and the structure of execution. As the definition is unique to a specific process but numerous processes of this type can be executed, a process instance or case is created. As this instance is executed, the trace refers to the actual path that is followed, whether or not that path is as planned. These traces are the focus point of process mining for the reason that they convey reality. Further, these traces consist of events, which in turn have attributes associated with them. A visual representation of the process definition concept is illustrated in Figure 3.5.

While this conceptualisation is that of an ideal case, information systems in general do not capture the information in this manner. Data manipulation is required to format it appropriately. This is primarily due to systems not being designed to monitor processes specifically. They do however monitor transac-

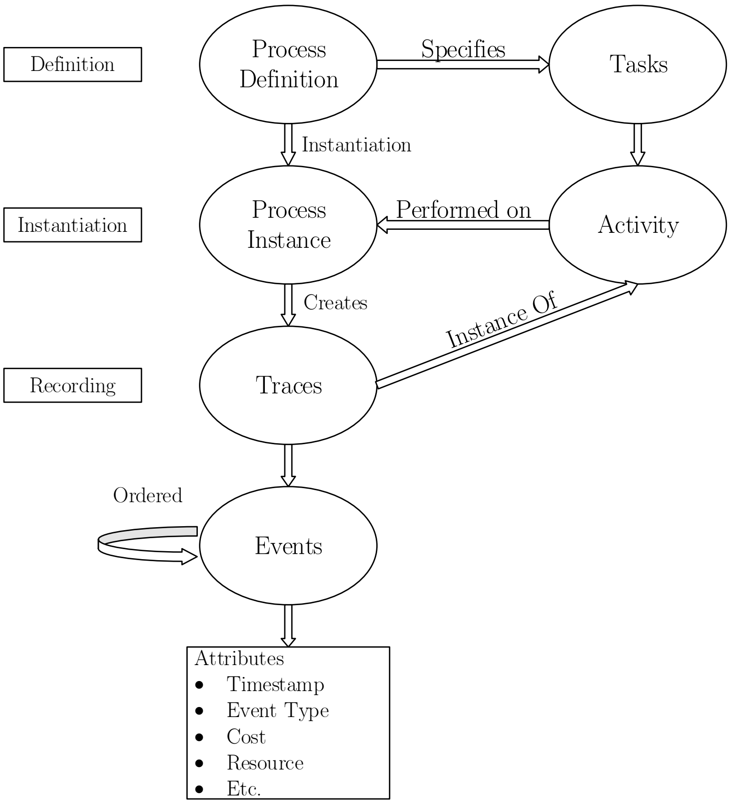


Figure 3.4: Conceptual event log structure.
Adapted from Fliegner (2014)

tions (often financial) which take place and the state changes required to issue new work orders as mentioned by Fliegner (2014). As most of the ERP systems store tables with the previously mentioned data, Structured Query Language (SQL) is used to extract the required information. This data is stored in a temporary database which serves as the data origin from which formatting and filtering can be done. When the table created here is formatted and filtered to the desired extent, it can be loaded into the process mining tool-set.

Fliegner (2014) goes on to identify common challenges concerning the extracted data. Of the challenges mentioned, there are only a few that have not yet been covered. A challenge that is referred to as *snapshots* entails the lifetime of the cases involved in the extracted processes. When looking at the recorded processes, it can be the case that a process was in progress before the recording was started. This is also true for the end of the process as to where the extraction of the event log interrupted a process which is still in progress and thus when being analysed, it seems as though the process is incomplete. These

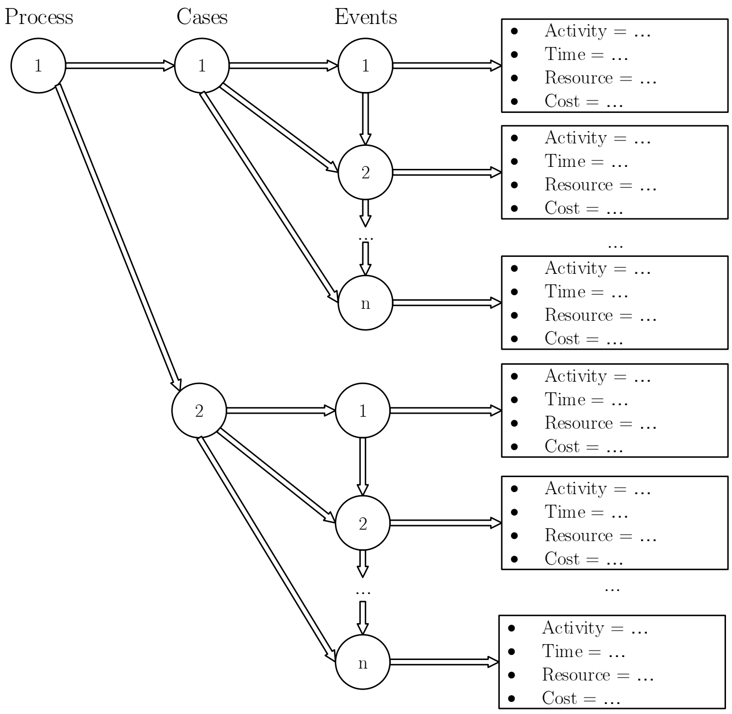


Figure 3.5: Event log structure.
Adapted from Van der Aalst (2011)

kind of processes should be addressed in the context of the filtering stage and assumed to be unwanted (noise), as they give an altered view of reality. Fliegner (2014) also refers to proper *event selection*. The author in this case refers to the level of detail contained in the event databases. As some processes have sub-processes within events, it should be decided if these sub-processes form part of the intended scope and detail level. Preferably, the level of detail should be uniform in the sense that the events in the event log are of the same detail and of the same perspective, as the process mining algorithms will treat all events in the log as equal, distorting trace events.

The next two problems faced are encountered when the event data is distributed over multiple databases. *Event correlation* needs to be ensured. This means that all events are required to belong to a case. This is an easy task when all events are grouped in one database but when the events are distributed, case referrals are often missing. The same concept applies to *timestamps*. When events are grouped within one database, the ordering of the processes already gives enough information to do process mining. However, when events are distributed across multiple databases, timestamps become a necessity in that they are often the only measure by which these events can be ordered.

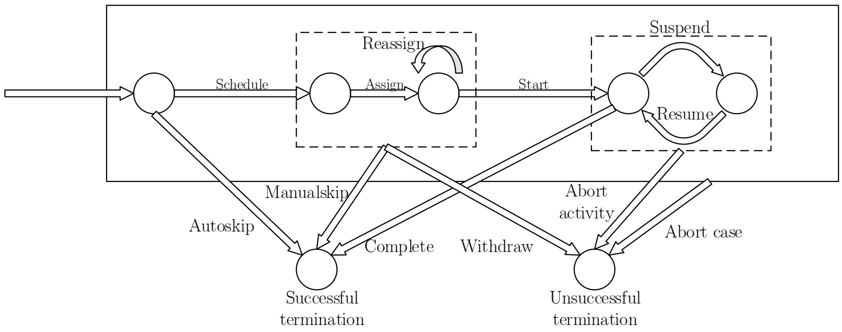


Figure 3.6: Transactional process life cycle model.
Adapted from Van der Aalst (2011)

Figure 3.6 shows a standard model for the life cycle of a *task* as presented by Van der Aalst (2011). It includes not only the preferred route available for the completion of a task but also the variation thereof. Variations include the reassignment of a resource and the termination of a task which can only occur when the task is in a specific state of its progress.

Up until now, an event log had the standard form as shown above, where the tasks were logged as they occurred; by a resource, at a time, belonging to a process instance. There is however another way to represent the same event log in terms of traces. Traces, as presented by Ingvaldsen (2011), are a collection of all the paths followed by the cases belonging to a process instance. This can be of use where a summary is needed whether it is necessary to understand the path followed by an individual case or the distribution of possible paths followed by numerous cases. An example of this is shown in Table 3.1.

Table 3.1: Event log traces

Notice for Maintenance Process	
Number of Instances	Log Traces
5010	ABDEA
460	ACDEHFA
29	ACGHFA

It should be noted that the event logs constructed from whichever records available, are the basis for process mining. In that, it should be recognized that the availability of accurate and reliable event logs are crucial for a successful process mining application. This is especially true when the results obtained from a process mining analysis will later form part of the BPM life-cycle in that they will be used in redesigning of evaluating the current business processes as shown in 2.2.1.

3.3.3 Data Processing

Phase 3 is given as data processing which is performed by first choosing the most applicable and suitable algorithms for the given case. The problem should be used as a reference as there are numerous metrics that can be calculated, most of which requires a slightly different approach. Most of the data processing can be done by available process mining software and should therefore only require the proper setup in most cases. The steps of phase 3 can be seen in Figure 3.7 where the progression of the methodology is also shown.

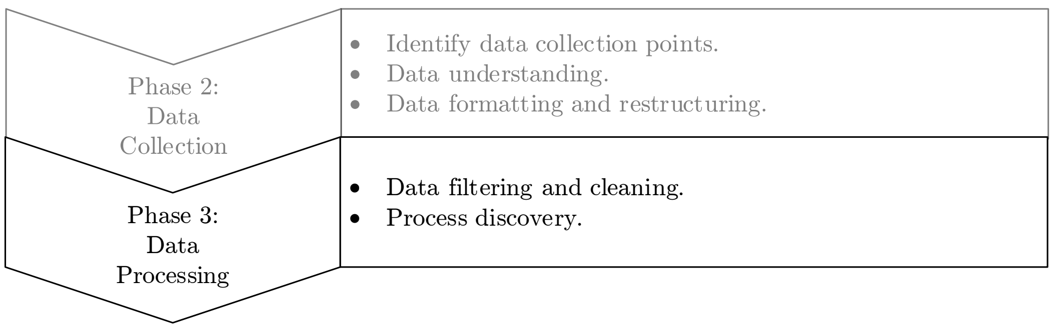


Figure 3.7: Application phase 3.

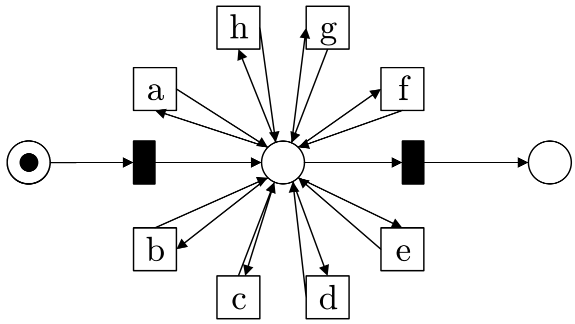
3.3.3.1 Process Discovery

The first step of phase 3 occurs directly after the event log has successfully been loaded into the process mining software. Step 1 of phase 3 entails the discovery of the process within the event log. This can also be seen in the process mining procedure shown in Subsection 2.3.1, Figure 2.10. The first step in process discovery is the selection of the algorithm responsible for mining the event log from which process discovery can be done. The selection of the mining algorithm does have an impact on the processes being constructed and therefore the results might be unreliable or not suited to the initial goals.

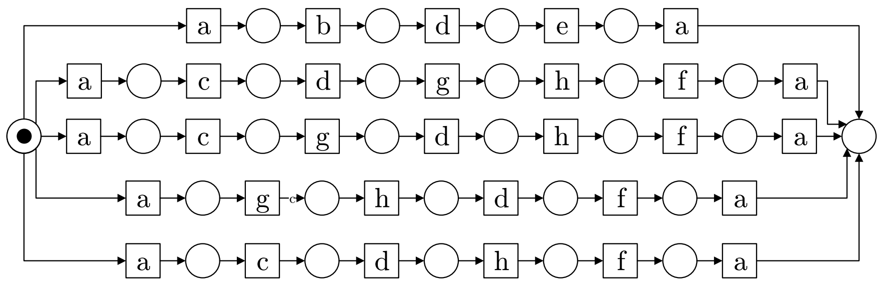
From a practical point of view, mining algorithms balance four quality criteria when performing process discovery as discussed in Van der Aalst (2011). The first consideration relates to the *Fitness* or the ability of the model to represent the actual behaviour in the process log. The second, which closely relates to fitness, is *precision*. Precision should constrain the discovered model not to allow behaviour not part of the original event log. In other words, the model should not be *underfitted* when referring to the desired process and therefore not be too generic to allow any process behaviour. An example of a underfitted model is shown in Figure 3.8. Thirdly, *generalisation* refers to the ability of the algorithm to generalise the behaviour and build a suitable model from that generalisation. This should however be done without *overfitting* the event log (Van der Aalst, 2011). Overfitting can be seen as the attempt of an algorithm to create a model that is too precise in showing event log behaviour. Overfitting then leads to a model where no clear patterns can be observed. An overfitted model can be seen in Figure 3.8. Lastly it is desirable that the model is as simple as possible, therefore *simplicity*.

Different algorithms used in process mining perform each of these criteria differently and should be considered depending on the situation. In Subsection 2.3.3 there is a discussion on the different types of algorithms used in process mining for the discovery of models. The first of which is the α-algorithm. In general, this is not a favourable algorithm as it has difficulty addressing noise, unsuitable behaviour and overly complex routing of cases as discussed in Subsection 2.3.1. It should also be assumed when using the α-algorithm that the event log is entirely complete, although when this is true there are still some complications discussed in full detail in Van der Aalst (2011). For these reasons, heuristic and genetic mining alternatives should be considered, as they perform more favourably with regards to the four quality criteria discussed above and can deal with infrequent paths and internal loops.

A good example is presented in Van der Aalst *et al.* (2011) where a comparison is made between different models with all being constructed from the



(a) An underfitted model.

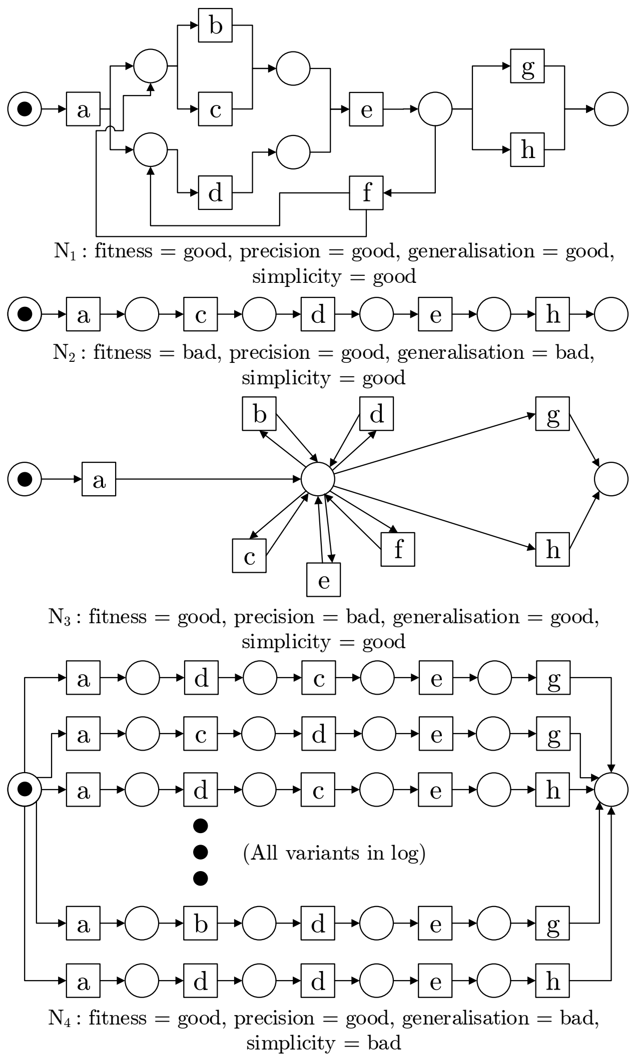


(b) An overfitted model.

Figure 3.8: Example of fitted models for the same event log.
Adapted from Van der Aalst (2011)

same event log. These models clearly show the influence different algorithms can have on the event log when there is a focus on one or more of the four quality criteria mentioned above. These models are shown in Figure 3.9.

Taking Figure 3.9 and considering N_1 to N_4 as a function of the quality criterion, it can be seen that N_4 places an emphasis on fitness and precision. This causes a decline in the generalisation of the model, as well as the models' simplicity. In other words, the model tries to cater too strongly towards every trace instance and results in being a complete visual representation of every trace. N_3 continues to reduce the precision to a point where the model is so generalised as to not bind the process to any constraints. N_2 attempts to make the model as simple as possible but at the cost of it being generalised to a point where the fitness suffers. N_1 finds an optimal solution where a balance is found between all the criteria and results in a usable model.



#	Trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	abdeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdb
1	eg

Figure 3.9: Model alternatives based on algorithm.
Adapted from Van der Aalst (2011)

3.3.3.2 The Inductive Visual Miner

As the selection of appropriate algorithms can be a difficult task, whether it is in the commercial or academic environment, Leemans, Fahland and Van der Aalst (2014*b*) introduces a complete software plug-in for ProM called the Inductive Visual Miner (IvM). This tool aims to package all the steps required for the implementation of process discovery without the conventional iterative process involved with other algorithms when a suitable model needs to be generated. The iterative process usually involved with model creation entails setting parameters and generating a model without the ability to have immediate feedback on changes. The IvM plug-in allows for the animation of the model according to the given event log and for filtering the event log interactively while the changes are observed. The IvM plug-in is available as a free plug-in for ProM and can be installed using the package manager as discussed in Subsection 2.3.7.

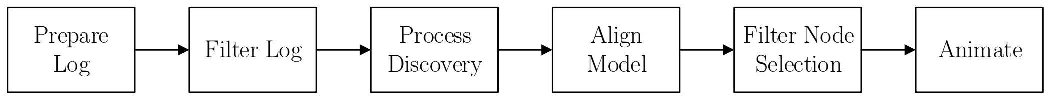


Figure 3.10: IvM tool process.
Adapted from Leemans et al. *(2014*b*)*

Figure 3.10 shows the process implemented within the IvM software plug-in. The process entails the following steps:

1. Prepare Log: Using the perspective classifier setting, the events are classified. The perspective considered here is the control-flow perspective or the resource perspective;

2. Filter Log: Process instances in the log are filtered out according to the frequency by which they occur;

3. Process Discovery: Using the Inductive Miner (Leemans, Fahland and Van der Aalst, 2014*a*), the process model is discovered.

4. Align Model: The log traces are replayed and put on top of the created model for user feedback;

5. Filter Node Selection: Filters are placed on every node (activity) to show the traces which pass through that node; and

6. Animate: Using the timestamps in the event log, traces are animated on the process model. If there are no timestamps, the IvM generates random timestamps for visual feedback.

After the IvM has been run and the filters have been adjusted to the satisfaction of the analyst, the model can be saved and used for further analytic purposed.

The IvM tool also offers to show the deviation of the event log from the generated model. Leemans *et al.* (2014*b*) identifies deviations either as a *log move* or a *model move*. When a log move occurs, it entails a deviation which is not permitted by the model. When a model move occurs, it indicates that there is an event that occurred in the model, but is not present in the log. These deviations are both visualised as dashed red lines which either show a path around an event (model move) or a task which loops into itself (log move).

3.3.4 Data Analysis and Results

Phase 4 is dedicated to giving the necessary insights to allow process improvements. The steps of phase 4 are shown in Figure 3.11. Process mining tools primarily provide calculated results in terms of metrics relating to the given process. It is therefore the responsibility of the analyst to further interpret the results and identify problem areas and applicable results. The metrics calculated are used to give insights into the behaviour of the process, not to implicitly improve the process. That responsibility remains with management and the personnel responsible for process design and redesign.

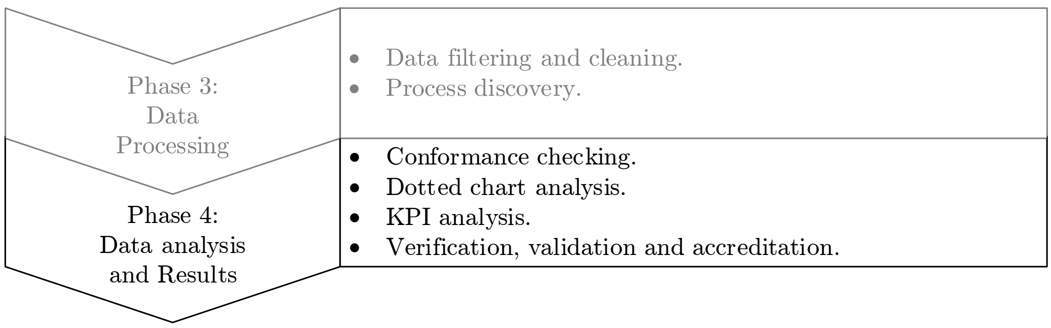


Figure 3.11: Application phase 4.

3.3.4.1 Conformance Checking

Before discussing conformance checking further, it should be noted that even though it is placed after process discovery in this thesis, it does not necessarily mean that it is always the order in which they occur. The "as-is" order is based on a situation where no model is present or where a more realistic model is required. In some cases, this need is not present and then the process discovery step is skipped.

Conformance checking is a process that entails taking an event log and comparing it against a model, whether it is discovered or planned. This allows the conformance checking algorithm to find any misalignment between real and modelled behaviour. Conformance checking is useful in cases where there is a misalignment between the business processes towards the creation of value and when auditing of procedures is required. These auditing exercises might be done to give feedback to the BPM redesigning activity for improved process performance or to repair process models that are out of date.

Rozinat and Van der Aalst (2006) identifies two dimensions of conformance namely fitness and appropriateness. In brief, fitness is a quantitative metric that conveys how closely the model and the event log traces coincide with one other. This alone is however not enough, as two generic models are able to allow for any trace in the event log while it is not a good model to represent the process. The idea is then to balance fitness with another metric called appropriateness. Appropriateness ensures that the model should be clear in the processes it allows but should be compact and simple and therefore minimalistic as described in Subsection 3.3.3.1.

Model Fitness

The principle behind the measure of fitness as presented by Rozinat and Van der Aalst (2006) is to replay the event log over the model and then measure the mismatch. As every trace is replayed, one task at a time, the number of states or tasks that had to be altered or created in the model to fit the replayed log is counted towards a mismatch.

Fitness, f, is calculated by using Equation 3.3.2 where k is the number of unique traces in the event log and n the number of processes as a combined trace. Furthermore, m is then defined as the number of missing states or tasks, r the number of remaining states or tasks and c the number of matching states or tasks. Finally, p refers to the number of created states or tasks. As $m_i \leq c_i$ and $r_i \leq r_i$ for all i, the fitness measure f will always be between 1 and 0.

$$f = \frac{1}{2}(1 - \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i c_i}) + \frac{1}{2}(1 - \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i p_i}) \tag{3.3.2}$$

This should however not be the only aspect of the process instances being taken into account. The areas within the model where problems occur could offer great insight into why and how processes perform sub-optimally. Rozinat and Van der Aalst (2006) also emphasise that the when the fitness of a given trace is calculated and that trace follows a certain path in the model, all the transitional points in the model are compared to the corresponding events in the log. This happens even if the resulting path of the model is not followed. Thus, in terms of the fitness metric, a chain of missing events is just as punishable as a single event.

Model Appropriateness

In the context through which this thesis addresses model appropriateness, it is seen as appropriate in terms of the structure it presents and being able to dictate the behaviour seen in the event log. Assuming the structural appropriateness of the model, it is favourable to make the model as compact and small as possible. The first metric to evaluate the structural appropriateness is simply a metric which, taking into account all processes having a start and end, then takes the number of tasks and divides it by the number of nodes as can be seen in Equation 3.3.3 where T is the number of tasks and n represents the number of nodes.

$$a_S = \frac{T + 2}{n} \tag{3.3.3}$$

The other aspect of appropriateness is how the model addresses, not only all the cases in the event log, but also cases not present in the event log while they are still possible. The danger here is that the model should fit the event log appropriately without being too general to where any case can be replayed. Highly general models become uninformative in describing process behaviour and patterns. The approach to achieve an approximation of suitable appropriateness is discussed in Rozinat and Van der Aalst (2006) where they use the mean of the tasks able to fit in the model as the event log is replayed. This approach ensures that the model is able to address and reflect what has been observed in the event log. This concept is demonstrated in Equation 3.3.4 where k is the number of unique traces, n the number of combined process instances, m the number of tasks and then x would represent the mean number of passed over tasks.

$$a_B = 1 - \frac{\sum_{i=1}^{k} n_i(x_i - 1)}{(m-1) \times \sum_{i=1}^{k} n_i} \tag{3.3.4}$$

When fitness and both metrics of appropriateness are applied to a given study, it is not generally a trivial task to combine and apply them as they correlate in some sense. Because of this, conformance should be tested and adjusted in phases. First, fitness has to be established and any alterations to the model have to be made until a desired fitness has been achieved. The appropriateness is then calculated to finally find the appropriate model for the given process. Going through this entire process is usually only done in a BPM redesigning environment where the outcome is to improve the current workflow. When process analysis is done, the metrics are simply calculated to give an indication of the current process state.

Dealing with conformance in general cases, it should be noted that either the concept can be seen as, the constructed model is not adequate or the behaviour in the log does not follow what was desired. Both of these views are dependent on the application. When the model is constructed from an event log and then tested, the former applies and when a model has been planned and is now used to audit the recorded behaviour, the latter applies.

3.3.4.2 Dotted Chart Analysis

While conformance checking is an analysis technique based on comparing a model to a given event log, a dotted chart analysis closely relates to the discovery process in that it visualises a perspective of the event log. The perspective shown by a dotted chart enables the analyst to discover patterns and trends not possible by looking at the event log. This type of analysis gives an overview of the captured event log that emphasises the timely occurrence of events as described by Song and Van der Aalst (2007) and Van der Aalst (2011).

The dotted chart is a type of historical Gantt chart as it shows when events happened in the event log. A dotted chart can deal with two variables on the axes with time being placed on the x-axis while another component can be chosen for the y-axis. The y-axis can be set up to show the grouping of events according to the resource, process instance or task type as time passed. This can then be used to clearly show the spread of events and therefore the time passed between tasks. The visual spread of dots is able to show numerous performance measures. The x-axis can be set to show time as true historical reflection or as a relative measure where all process instances start at the same point. This can then clearly show process life-cycles and highlight performance indicators where instances are compared to one another.

Song and Van der Aalst (2007) also shows how the dotted chart can show performance metrics in terms of:

- The start of the process in the event log;
- The end position of the process in the event log;
- The average spread of the events in the event log;
- The minimum spread of the events in the event log;
- The maximum spread of the events in the event log.

All of which does have slightly different interpretations depending on the component type selected for the y-axis.

3.3.4.3 Process KPI Analysis

The most common method of obtaining KPIs of a given event log is the log replay method. When considering this kind of analysis, an event log is required for a process in the form of a Petri-net. The Petri-net in can either be obtained by first mining a process from the event log or it can be a planned process where the Petri-net is already available. In either case, when both these elements are loaded, relationships need to be established between the Petri-net and the event log. These relationships are necessary for the system to know which events in the event log belong to which transitions in the Petri-net.

Hornix (2007) describes the following KPIs that are related to business processes that can be used in PPA as discussed in Subsection 2.3.6. First, the overall completion or throughput times of the processes are looked at. Taking L as the output given by the log replay and c as the case or process instance, the total completion time $T(c)$ of any case c is given by Equation 3.3.5.

$$T(c) = MAX_{(x \in L \wedge case(x)=c)} time(x) - MIN_{(x \in L \wedge case(x)=c)} time(x) \tag{3.3.5}$$

Also, taking $\#c$ as the number of cases, the following definitions also apply:

- Average throughput time: $AVG_{c \in C} T(c)$
- Minimum throughput time: $MIN_{c \in C} T(c)$
- Maximum throughput time: $MAX_{c \in C} T(c)$
- Standard deviation of throughput time: $STDEV_{c \in C} T(c)$
- Percentage of lowest throughput time: $AVG_{low}(x, T(c)) = AVG_{c \in C_{low(x)}} T(c)$ Where $c \in C_{low(x)} \leftrightarrow \frac{\#d \in C | T(d) \leq T(c)}{\#c} \leq \frac{x}{100}$
- Percentage of lowest throughput time: $AVG_{high}(y, T(c)) = AVG_{c \in C_{high(y)}} T(c)$ Where $c \in C_{high(y)} \leftrightarrow \frac{\#d \in C | T(d) \geq T(c)}{\#c} \geq \frac{y}{100}$
- Average Case length: $AVG_{normal}(x, y, T(c)) = AVG_{c \in C_normal(x,y)} T(c)$ Where $c \in C_{normal}(x, y) = C \backslash (C_{low(x)} \not C_{high(y)})$

Another aspect worthy of investigation is the arrival rate of the cases or process instances. This can be calculated using any time unit also taking L as the replay output and the case c while the arrival time of every case is now given by $T'(c)$ and thus:

$$T'(c) = MIN_{(x \in L \wedge case(x)=c)} time(x) \tag{3.3.6}$$

The average arrival rate is then given by:

$$\frac{\#c}{MAX_{c \in C} T'(c) - MIN_{c \in C} T'(c)} \tag{3.3.7}$$

The conformance of a given case can also be calculated by means of using the number of cases that fit into the Petri-net as the log is replayed. This however only gives an idea of the real conformance and a dedicated conformance checker should rather be used.

The KPIs described above are only applicable to case instances. The same principles applied to the cases can be applied to the activities within these cases. As these activities are tied to the places within a Petri-net, KPIs can be derived according to the transition of tokens and the firing of atomic transitions as discussed in more detail in Subsection 2.2.5. Transitional times for the tokens are based on the timestamps within the event log. Every transition time is however not tied to a specific token which causes problems when multiple tokens are at the input of a transition. In situations like these, a random function is used to assign a token to a transition time. As the token might belong to the wrong case, it might affect the KPI values calculated for individual cases.

When a token is created within a place, the time it spends there before it is consumed is referred to as the *sojourn time*. Sojourn times are calculated in the same manner as the throughput time of a case shown above. The *synchronisation time* of the token is seen as the time it takes for the transition to be enabled since it consumed a token. When a transition is enabled, and it takes a certain time before the token moves through and is consumed, it is referred to as the *waiting time* of a place. Another useful and critical KPI is the arrival rate of the tokens. As this is strongly related to bottlenecks which from within process instances as in some cases the resource is not able to handle the arrival rate of the tokens. In cases where multiple tokens can occupy the same place, the frequency of which the tokens arrive to the place can be calculated to obtain metrics based on the waiting time associated with the place and the severity of the bottleneck. As part of the predictive capabilities offered by KPI analysis, predictions can be made where splits occur within the process flow. The probability that a token will follow a certain arc can be given as Equation 3.3.8:

$$P(p,t) = \frac{\#x\ell L|trans(x) = t}{\#x\ell L|trans(x)\ell t'|(p,t')\ell A} \tag{3.3.8}$$

Where A represents the arcs in a Petri-net. The time between the triggering of two transitions can also be calculated with Equation 3.3.9, where L is the output of a log replay and the case c:

$$T(c,u,t) = |MIN_{(x\ell L \wedge case(x)=c \wedge trans(x)=u)} time(x) - MIN_{(x\ell L \wedge case(x)=c \wedge trans(x)=t)} time(x)| \quad (3.3.9)$$

In Equation 3.3.9 the two transitions in question have time attribute u and t. To be able to calculate the time between the transitions, each transition needs to fire at least once during the log replay.

The final set of KPIs describe the behaviour of the activities within the traces. The first of which is the waiting time of an activity. This is the time an activity has to idle since it was scheduled to when it can start. Waiting time is then simply calculated by using Equation 3.3.10 where x and $x' \epsilon L$ and represent the schedule time and the starting time respectively.

$$Waitingtime = time(x') - time(x) \quad (3.3.10)$$

The second KPI of value is the execution time of activities. This is simply defined as the difference in time between the starting of the event and the completion thereof. To calculate the execution time, Equation 3.3.10 can be used again. The value of x and x' now change to represent the start time of the event and the end of the event respectively. The sojourn time can also be calculated with Equation 3.3.10 where x and x' represent the scheduling time of the activity and the completion time of that activity.

3.3.4.4 ProM Plug-Ins

The analysis techniques described in this section are implemented in some form within a ProM plug-in. This enables the analyst to calculate all the required values within the same software environment. There are however two cases where the older version of ProM (version 5.2) should rather be used as it is more stable. The same analysis can however still be done in ProM 6.5. The analytical actions and the plug-ins to be used are summarised in the Table 3.2.

Table 3.2: ProM plug-ins for analytic actions

Action	ProM Version	Plug-in	Author
Csv to mxml conversion	Prom Import Framework	-	-
View Summary	Prom 6.5	-	-
Filter Log	Prom 6.5	Filter log using simple heuristics	HMW Verbreek
Filter out log event	Prom 6.5	Filter events	SJJ Leemans
Mine Process Model	Prom 6.5	Mine with inductive visual miner	SJJ Leemans
Check Process model fitness and appropriateness	Prom 5.2	Conformance checker	-
Mine dotted chart	Prom 6.5	Analyse using dotted chart	MS Song
KPI activity analysis	Prom 5.1	Basic performance analysis	-
Log replay for performance	Prom 5.2	Performance analysis with Petri-net	-

3.3.4.5 Verification, Validation and Accreditation

Verification, Validation and Accreditation (VV&A) forms part of any modelling project where the results need to be deemed reliable and credible Petty (2010). Verification addresses the requirements regarding the constructed model. DoD (2003) defines verification as the determination of whether the model is compliant towards the intended specifications. Validation, on the other hand, deals with the testing procedure to check the coherence of the model towards the real world situation (DoD, 2003). In this procedure, tests are performed where results are measured for accuracy based on the intended purpose of the model. Accreditation is described by Petty (2010) as being more qualitative in nature than verification and validation while being similar in the intended outcome. It is a decision process that involves an appointed person for acceptance. The acceptance condition relates to the intended use of the model and is reliant on the desired outcomes. This ensures that while the model may pass according to the verification and validation procedures, it may not be suited for its intended use and, therefore, would fail VV&A. It can be seen that each step, while somewhat similar in nature, addresses important areas of the model and its application.

Van der Heijden (2012) assigns different roles to each stage of the VV&A process. The verification process is ideally the responsibility of the analyst involved in the process mining effort. The responsibility for validation then goes to the project leader who applies context to the model and validates it in accordance to the intended outcomes while utilising his knowledge of the applicable environment. The process manager assesses the qualitative nature of the process model by performing the necessary accreditation checks which then conclude the VV&A procedure.

3.3.5 Operational support

The last phase, phase 5, entails the value realisation of the entire process mining application. An overview of phase 5 is shown in Figure 3.12. Phase 5 uses the results from previous phases and applies the knowledge to the organisation. The primary feedback in phase 5 entails the detection of the real world behaviour, the prediction of how operations should continues and how processes will behave given historical knowledge. Phase 5 ultimately entails the recommended mitigations towards business processes to achieve desired outcomes. It is important that the outcomes of the process mining application link back to the outcomes and objectives that resulted in the using of process mining in phase 1.

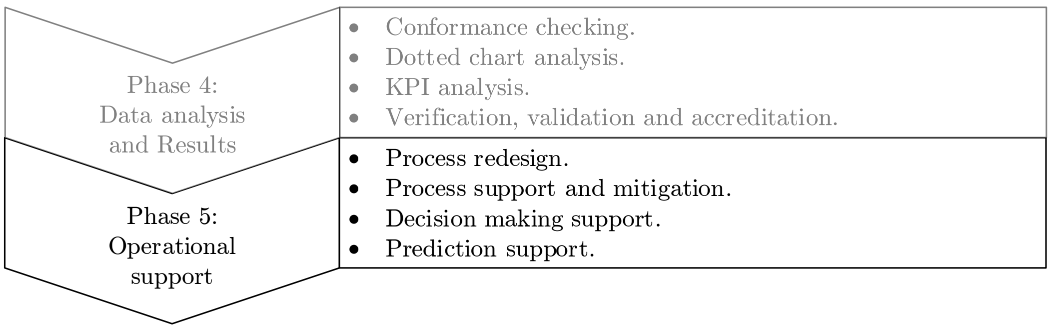


Figure 3.12: Application phase 5.

The most direct way in which process mining contributes to process mitigation is with the use of BPM (discussed in Subsection 2.2). In these cases, process mining is involved in the redesigning of processes within the BPM life-cycle. Adjustments can also be made to the current process by simply changing resource allocations and the way tasks are done. This is a less obstructive and might be all that is necessary to obtain desired results.

On a managerial level, process mining can indicate policy changes which need to occur or which personnel require training or disciplinary action based on their performance or non adherence. Management can also directly intervene in certain processes when monitoring indicates malpractice or that a process is not performing as desired. Results obtained by process modelling within process mining can also offer predictive knowledge based on the past performance of processes. Predictive knowledge can enable improved decision making to take place on higher levels in order to support the timely completion of processes and better resource allocation.

As environments in which organisations operate change and the PAM strategy to succeed within these environment change, it is not recommended that process improvement only occur once (Van der Aalst *et al.*, 2012). Management should thus be aware of the long term plan in which process mining should be included as an ongoing initiative to increase value generated by assets supported by its strategy. Continued support can either include applying process mining to certain aspects at set intervals or a continued monitoring implementation. This is dependent on the case at hand and depending on the volatility of the process. Whichever the case might be, Van der Aalst (2011) states that the process mining application should support operations by detecting deviation, aiding in prediction for decision making and recommending actions.

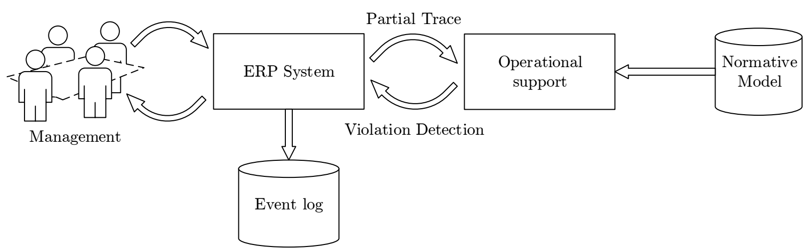


Figure 3.13: Process violation detection workflow.

As part of the organisation's ongoing implementation of process mining, it can aid in the detection of process violations as described in Van der Aalst (2011). Figure 3.13 shows how this process would typically look. A normative model is constructed as part of the planning process which specifies how processes should be conducted. With a normative model in place, process mining can offer real time feedback on how processes are performing with relation to the normative model. Once violations or undesirable events or paths take place, management can step in to mitigate or terminate the current process.

Van der Aalst (2011) also goes on to explain that by using a historical event log captured from a given process, a predictive model can be constructed. This concept is shown in Figure 3.14 where a partial process trace is first extracted from the information system. The predictive model is then used to predict behaviour with regards to the partial trace. This can include predictions on the following:

- Completion time;
- Probability of satisfying certain constraints;
- Total cost when this information is available for process events;
- Probabilities for certain events to occur;
- Probabilities for the utilisation of resources;

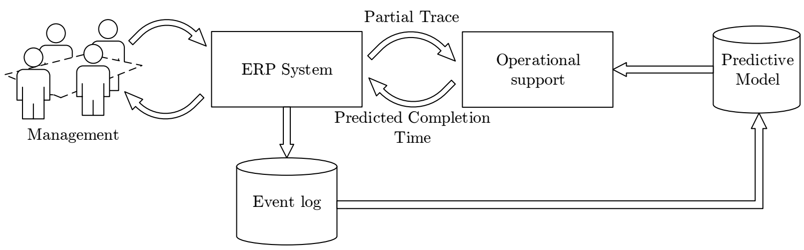


Figure 3.14: Process behavioural prediction workflow.

Van der Aalst (2011) finally states that process mining can support operation in the form of recommendation. The recommendation model is shown in Figure 3.15. While the model might look similar to the predictive model, in this case a prediction isn't sent as a response but rather a recommendation on which actions to take. These recommendation can be based on the predictive model where a process is guided the current state of other processes. For example, if a resource is causing a bottleneck, the recommendation might suggest a path around a resource to improve the completion time. The recommendation made are always coupled with a goal in mind as different goals will determine different routes.

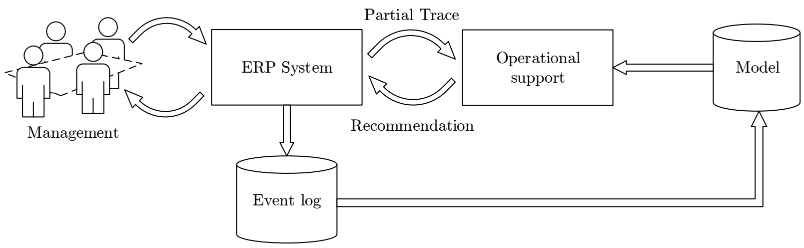


Figure 3.15: Process flow recommendation workflow.

While one suitable operational support model can be chosen for individual cases, the most benefit can be gained when all these models are implemented together. This is not only to maximise the amount of information available for decision making but also as the supporting models offer different perspectives on the same problem area. The models are then able to complement one another and allow increased benefits towards the organisation.

3.4 Chapter Conclusion

This chapter presents the techniques and approaches discussed in Chapter 2 as an application methodology to be applied to the PAM environment. Different techniques were first evaluated and then chosen to best suit the application area. The application phases are presented as steps that can be followed from the conceptualisation of the area of application to the operational support it can provide.

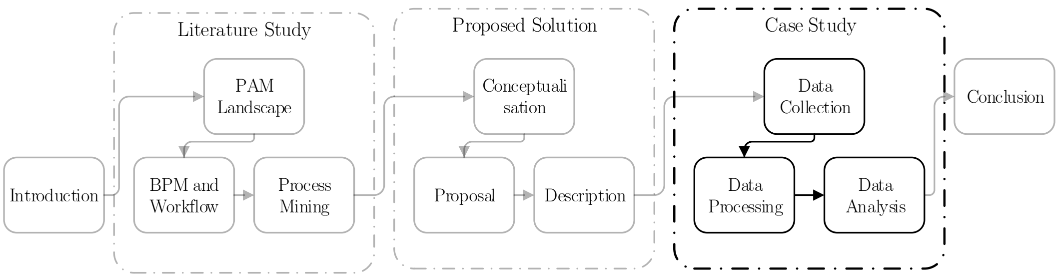
The application methodology presented here, while being focused on the PAM environment, is constructed to be as generic as possible to allow broad application. The techniques selected form the basis of process analysis and can be applied to various situations to ensure that useful insight can be provided independent of the exact situation. Depending on the scope and targeted process, techniques and KPIs can be selected to match the outcomes of the research questions or desired operational support.

To ensure the validity and practicality of the presented application methodology, a real world application should be investigated. Even though the analysis described in Section 3.3.4 is done from first principles, the real world application will be done using the ProM Framework. The use of ProM is not compulsory, but it does ease the handling of the large volume of data usually present in process mining applications. The following chapter shows the validation of the application methodology in terms of an application case study.

Chapter 4

Case Study

The aim of this chapter is to apply the methodology presented in Chapter 3 to a case study. The application aims to investigate the validity of process mining within the PAM environment and hopes to improve a process involved with supporting physical assets. The process model is investigated followed by a discussion on the process behaviour in the real world. The results are discussed in order to understand the process behaviour and determine what actions can be taken to improve the process. Improvements are desired to aid the supporting role within the PAM strategy and the value obtained by the process.



Chapter Outcomes
•Apprehension of the given case study and its role.
•Understanding of how process mining is applied.
•Discovery of process mining insights.
•Understanding of the results obtained.
•Validation of results.

4.1 Case Study Overview

In the following section, a case study is performed as validation for the methodology presented in Chapter 3. Case study data was obtained through a physical asset management (PAM) service provider. The data entails a maintenance process of one of their clients from the petro-chemical industry which has been targeted for the application of process mining. The goal was to use the data provided along with details of the planned process and compare the real world activity to what was planned and gain insights around that.

The nature of the data presented the opportunity to discover a process model from the as-is process activities and then to derive process KPIs from those activities. The model discovered can then be compared to the planned process to show conformance issues. When looking at the driving objectives behind process mining exercises in 3.3.1.3, this case study utilises a data-driven and question-driven approach.

4.2 Design Of Study

As mention in the previous section, the process in question is part of a petro-chemical organisation's maintenance procedure. The process forms part of the organisation's asset management implementation where the information system is able to schedule and manage the work and state changes within the process. From here on, state changes will refer to the completion of one task in the process and the progression to the next. The control-flow perspective of the process can be seen in Figure 4.1 with the preferred path being highlighted in red.

From the case study briefing, it was clear that while the process had a planned route, there was no feedback on what was happening in the real word concerning the actual paths. The planned process shown in Figure 4.1 is also outdated in that the actual process includes more tasks than are shown here. This is as a result of the changing environment in which the process operates and how the organisation updates the process to improve operation. A full list of all the activities and their descriptions are available in Appendix A, Table A.1.

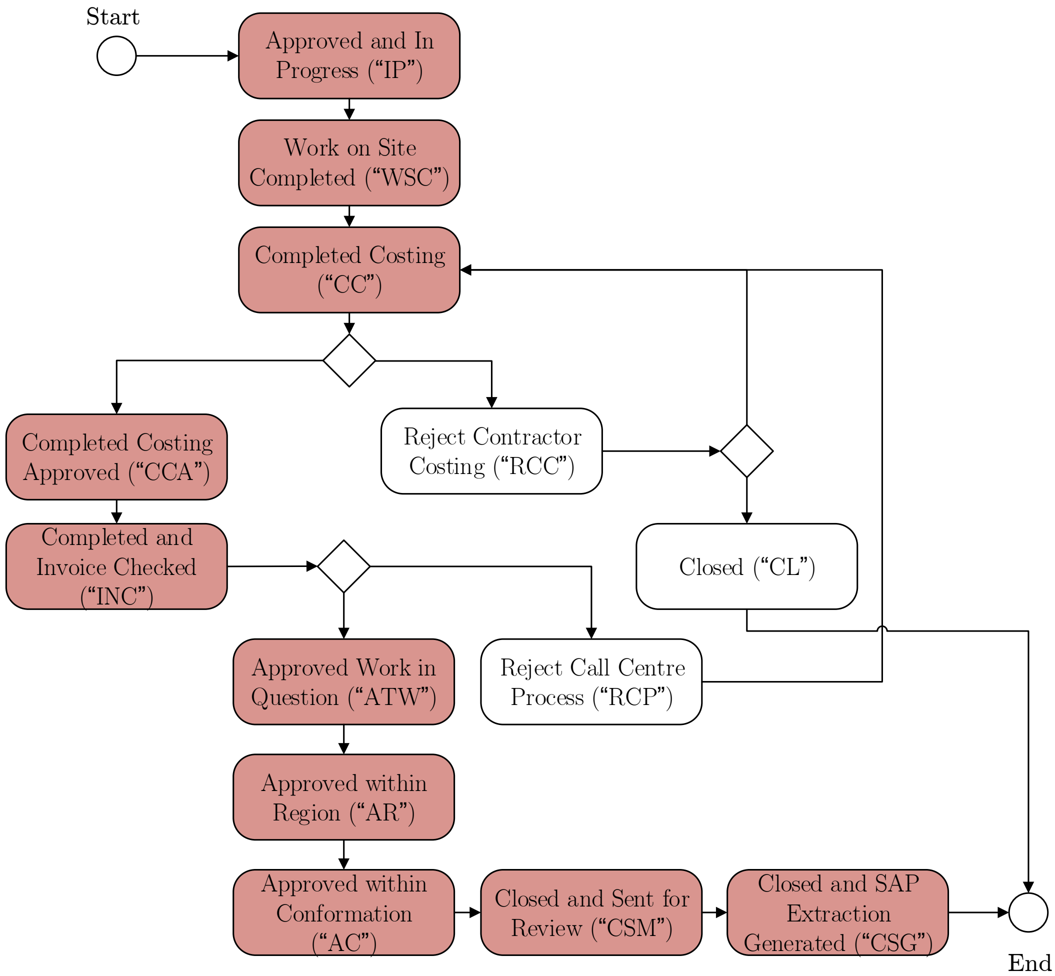


Figure 4.1: Base Maintenance Process (Planned Process).

Because of a non-disclosure agreement, more details regarding the process cannot be unveiled. The case study will thus be missing some much desired detailed analysis. This will also be reflected in the results discussion as specific actions towards process improvement cannot be described, only with regards to the activity descriptions in Table A.1.

4.3 Data Collection

Data collection was done by utilising the organisation's Enterprise Asset Management System (EAMS). Implementation of the EAMS is managed by a PAM service provider who conducts the data analytics with regards to captured asset information. The EAMS has the ability to facilitate an asset register, record the conditions of assets within this register and manage work orders related to the maintenance of physical assets. As work orders captured in this system are subject to the processes which support the physical assets within the organisation, they can be used to reconstruct the supporting PAM processes. It is thus these supporting processes which enable the value creation of the physical assets. As the process can now placed within the PAM strategy of the organisation where value creation occurs, the conditions discussed in Subsection 2.1.2 are met.

Data extraction from the system was done by the technical team leader where he was able to retrieve transactional data via SQL database queries. The EAMS system does not refer to these records as an event log as discussed in this thesis but rather refers to it as a "change log". This entails that every state change of the monitored process is recorded. The fields retrieved for the change log are:

- CaseID;
- Resource;
- Timestamp;
- Transition From; and
- Transition To.

As can be seen, the structure is closely related to the ideal case described in Subsection 3.3.2.2. The only difference results from the *change log* interpretation of the system where status changes are recorded at certain times when the process progresses from one status to another. States referred to here can be interpreted as the progression of tasks or activities within the process. The only issue with this is that there are no separate entries for the start and end of the process states. The EAMS records one entry which describes the transition from one state to the next.

The data is exported to a csv (comma-seperated text-file) which enables universal usage. While it could not be exported to the desired mxml format, it does grant the analyst the ability to open the log with a variety of other applications. The log was reformatted to mxml by the ProM Import Framework software package [1]. The data set contains records from the 4th of January 2014, 5:59 AM to the 8th of June 2015, 11:35 AM. 33,614 records are contained within the data set, which in terms of this case study refers to the number of status changes. A preview of the data is shown in Table 4.1.

Table 4.1: Event log preview

CaseID	Resource	TimeStamp	TransitionFrom	TransitionTo
AH10441#	–	29/4/2015 13:30	CL	RWO
AH10441#	–	29/4/2015 13:31	RWO	WSC
AH10441#	–	29/4/2015 13:33	WSC	CC
AH10441#	–	29/4/2015 14:00	CC	CCA
AH10441#	–	30/4/2015 8:27	CCA	INC
AH10441#	–	7/5/2015 14:37	INC	ATW
AH10441#	–	7/5/2015 14:38	ATW	AR
AH10441#	–	7/5/2015 14:38	AR	AC
AH10441#	–	10/5/2015 3:00	AC	CSM
AH10716#	–	15/1/2015 11:27	CL	RWO
AH10716#	–	15/1/2015 11:28	RWO	WSC
AH10716#	–	15/1/2015 11:33	WSC	CC
AH10716#	–	15/1/2015 12:01	CC	CCA
AH10716#	–	15/1/2015 12:16	CCA	INC
AH10716#	–	16/1/2015 4:51	INC	ATW
AH10716#	–	16/1/2015 13:10	ATW	AR
AH10716#	–	16/1/2015 13:11	AR	AC
AH10716#	–	18/1/2015 3:00	AC	CSM
AH10716#	–	21/1/2015 9:59	CSM	CSG

While utilising manual exploration of the data in Microsoft Excel, some irrelevant data points were observed. The first stage of filtering was simply done by removing all cases that contained only one entry as this does not convey a process and is very likely to be the result of the system or human input error. It was also seen that some cases were opened by the system only to be closed immediately after. As this will also only interfere with the created model and skew results, all cases with only two entries were also removed. A discussion with the data analyst revealed that there was a change in the process model

[1] http://www.promtools.org/promimport/

and the activities within the process from 2014 to 2015. This leads to the discarding of all data for 2014. Even though only 6 months worth of data remained, it still included 198,642 records which are adequate to perform the desired analysis.

4.3.1 Data Transformation

As data is almost never in the exact form preferred for processing, some alteration to the event log had to be made. The first of these alterations was done because of the way the EAMS captures state changes as discussed in the previous subsection. An algorithm was written in R [2], an analytical programming software package, to insert an extra line before every process instance. This extra line copied the "TransitionFrom" field to the "TransitionTo" field. The "TimeStamp" for this entry was simply taken as one minute before the next line to ensure that the event will be placed first during processing. An example is show in Table 4.2.

Table 4.2: Event log row addition example.

CaseID	Resource	TimeStamp	TransitionFrom	TransitionTo
AH10441#		29/4/2015 13:29		CL
AH10441#	ChantaldeS	29/4/2015 13:30	CL	RWO

The second alteration was done with respect to the anonymisation of the data. The event log contains a "resource" field which indicates the person responsible for logging the state change. As part of the agreement with the organisation who supplied the data, this field was anonymised. To do this, the *Anonymize Log* ProM plug-in was used. This plug-in replaced all entries in the resource field with a respective letter.

4.4 Process Mining Analysis

The application of process mining to the event log is done using ProM which was discussed in Subsection 2.3.7. The csv file containing the event log is im-

[2]https://www.r-project.org/

ported and then used to generate a report showing the initial attribute values for the event log. A screen-shot of this can be seen in Appendix B, Figure B.1. The summary can be seen in Table 4.3 with the initial values shown under "Values Before Filtering".

Table 4.3: Summary of event log attributes

Attribute	Value Before Filtering	Value After Filtering
Processes	1	1
Cases	17414	15809
Events	216026	203309
Event Classes	27	26
Events Per Case	Min: 3, Mean: 12, Max: 51	Min: 4, Mean: 13, Max: 45
Events Classes Per Case	Min: 2, Mean: 11, Max: 16	Min: 3, Mean: 11, Max: 16
Log Start Date	Fri Jan 02 6:05 2015	Fri Jan 02 6:05 2015
Log End Date	Mon Jun 08 11:35 2015	Mon Jun 08 11:03 2015

An important part of any application methodology where data is utilised is data filtering. As discussed in Subsection 2.3.1, this is not only to remove unnecessary complexity but also noise (unwanted data points) which is often part of high volume data collection as discussed in Subsection 2.3.1. By using the "Log Filter" plug-in provided with ProM, data filtering was mainly implemented to satisfy time domain conditions discussed in Subsection 3.3.2.1. By implementing the log filter, only cases that started and ended within the considered time frame were included for further analysis.

By using the log filter plug-in in ProM, all instances not starting and ending within the specified time frame were removed. To complement this filtering operation, only the top 95% percentile in reoccurring instance were kept for analysis to ensure that most redundant records are removed. This additional filtering is done to ensure that the model obtained disregards unnecessary complexity but can still convey a useful message as discussed in 3.3.3.1. The values after filtering can be seen in Table 4.3 in the "Value After Filtering" column.

4.4.1 Process Discovery and Understanding

As shown in chapter 3, the first application outcome entails the extraction of the process model based on the activity recorded in the event log. By using the ProM plug-in called the "Inductive Visual Miner" discussed in Sub-subsection 3.3.3.2, the process models are extracted from the event log. Use of the Inductive Visual Miner (IvM) adds the benefit of "knowing" and controlling the conformance and appropriateness mentioned in Subsection 3.3.4.1. Control of conformance can be estimated as the IvM plug-in has the option to set the path and activity percentage which should be included in the model mining process. The former of which, the path setting, indicates the percentage of the trace paths that should be included in the model. The higher this setting is set, the higher the conformance at a loss of appropriateness. The latter, the activity setting, controls the amount of activities that should be included in the model. This can be used to alter the activities to be included from the event log. It should however be noted that the exclusion of activities also causes the decline in the conformance of the model as excluded activities will be seen as missed events.

Using the default path filtering setting of 80%, the process model shown in Figure 4.5 is obtained. The default of 80% is based on the Pareto 80/20 principle where the assumption is that 20% of the event log contains 80% noise and non-useful traces. Figure 4.5 illustrates the complexity of the mined event log. Figure C.1 and C.2 in Appendix C can be referred to for more detail. The model extracted here is clearly not identical to the planned process shown in Figure 4.1. The reason for this is the inclusion of more activities that are not present in the planned process and the lack of compliance towards the planned process.

To extract the planned process from the event log for validation purposes, first the number of activities were reduced to match the number of activities in the planned process. As there are a total of 27 activities described in Appendix A, and the preferred process only includes 13, as shown in Figure 4.1, the activities filter was set to 42%. The process model discovered closely resembles the planned process with the correct amount of tasks but with additional paths. The path filter was decreased until only the desired path was obtained at 47%. The mined process model can be seen in Figure 4.5. This verifies that the event log belongs to the correct process.

ProM has a conformance checker plug-in that can calculate both the conformance and appropriateness of a given model, given as a Petri-net, concerning an event log. Descriptions of the calculations that can be made by this plug-in are shown in Appendix F Figure F.1. Using this plug-in, mined models shown in Figure 4.5 and Figure 4.5 were evaluated. The calculated values can be seen in Table 4.4.

Table 4.4: Conformance checker calculated values.

	IvM Settings		Calculated Values		
Process Model	Activities	Path	Fitness $[f]$	Precision $[aa_B]$	Structural $[aa_S]$
Figure 4.5	100%	80%	0.8348	0.7473	0.919
Figure 4.5	42%	47%	0.8356	1	1

It should be noted that the precision and structural metrics shown in Table 4.4 are more advanced versions of the same metrics discussed in Subsection 3.3.4.1. These more advanced versions correct some errors that are inherent in the replay of event logs over Petri-nets where loops and missed event disproportionately penalise the metrics as discussed by Hornix (2007). Details of these alterations can be seen in Appendix F, Figure F.1. Before proceeding to the mined process for an analysis of the real world situation, a summary of commonly occurring traces was compiled using the ProM process inspector. This summary is shown in Table 4.5.

As can be seen in Table 4.5, the preferred process trace only contributes a total of 11.53% of the entire event log. The next trace, contributing 8.15%, only adds the "RCC" activity which involves rejecting the costing given by the contractor. The rest of the trace is identical to the preferred instance. Considering the traces ranked 3rd, 5th and 7th, it is clear that the order of activities "AR", "ATW" and "AC" are inconsistent. This inconsistency, only considering the top 10 traces, makes up 9.71% of the log.

The mined model can be examined to gain insight into the real world process flow. The full model is divided into three section as shown in Figure 4.5. The

Table 4.5: Discovered log traces

Ranking	# of Instances	Log Trace
1	1822 (11.53% of Log)	AAA;IP;WSC;CC;CCA;INC;ATW;AR;AC;CSM;CSG
2	1289 (8.15% of Log)	AAA;IP;WSC;RCC;CC;CCA;INC;ATW;AR;AC;CSM;CSG
3	756 (4.78% of Log)	AAA;IP;WSC;CC;CCA;INC;ATW;AC;AR;CSM;CSG
4	440 (2.78% of Log)	AAA;IP;WSC;CL
5	415 (2.63% of Log)	AAA;IP;WSC;CC;RCC;CC;CCA;INC;ATW;AC;AR;CSM;CSG
6	385 (2.44% of Log)	AAA;IP;WSC;CC;RCC;CC;CCA;INC;ATW;AR;AC;CSM;RCP;CSM;CSG
7	364 (2.30% of Log)	PAA;IP;WSC;CC;RCC;CC;CCA;INC;AR;ATW;AC;CSM;CSG
8	363 (2.30% of Log)	WSC;CL;RWO;WSC;CC;CCA;RCC;CC;CCA;INC;ATW;AR;AC;CSM;RCP;CSM;CSG
9	334 (2.11% of Log)	AAA;IP;WSC;CC;CCA;RCC;CC;CCA;INC;ATW;AR;AC;CSM;RCP;CSM;CSG
10	323 (2.04% of Log)	PAA;IP;WSC;CC;RCC;CC;CCA;INC;ATW;AR;AC;CSM;CSG

analysis is done from the end to the start of the model as this is the order in which complexity increases. Inconsistencies are clearly demonstrated in the mined process model shown in Figure 4.2 where the model presents the activities as being parallel to one another instead of sequential. The second split in the Petri-net flow shows the parallel occurrence of "CSM", "RCP" and a skip. The Petri-net has a visualisation error in this case. Observing Table 4.5 it can be seen that when "RCP" occurs, it falls into a loop with "CSM" which always precedes "CSG". The algorithm limits itself to a single representation per activity which ultimately caused it not to show "CSM" as a follow up event. It is also in this case where the necessity of a loop should be considered, as including it would mean that more paths are required, leading to a more complex model. As discussed in Subsection 3.3.3.1, it is important to reduce the complexity of the mined model to still be useful and offer valuable results.

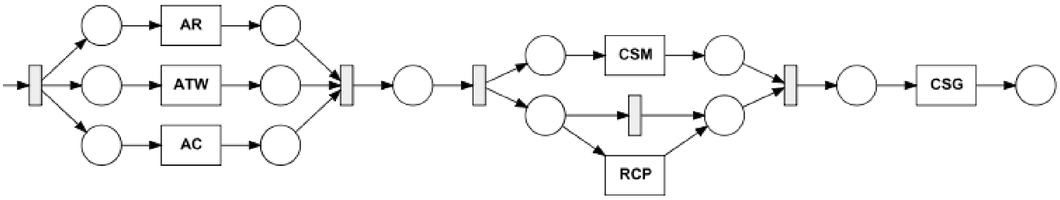


Figure 4.2: Process end examination.

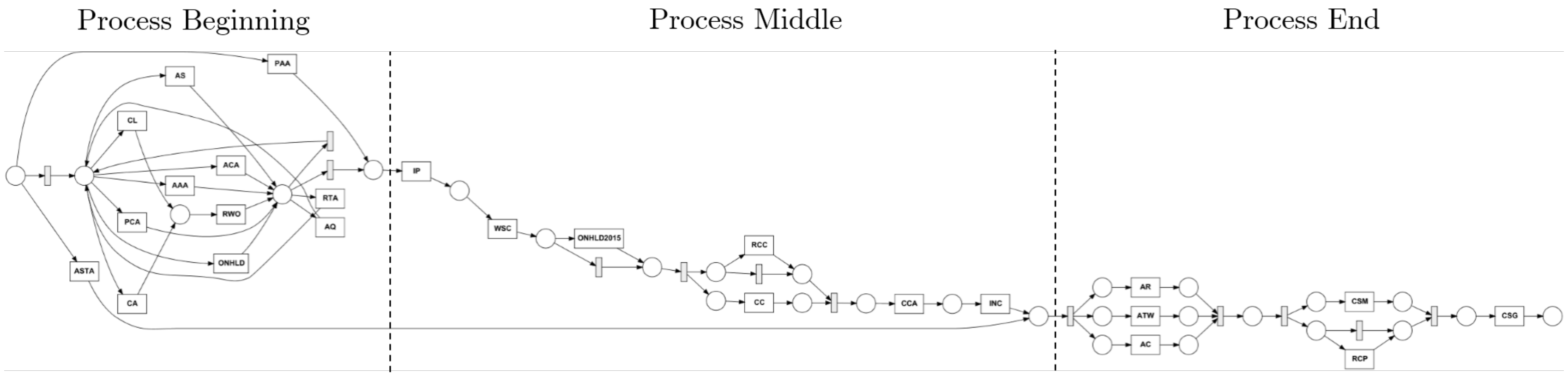


Figure 4.3: Mined process model with 100% activities and 80% paths

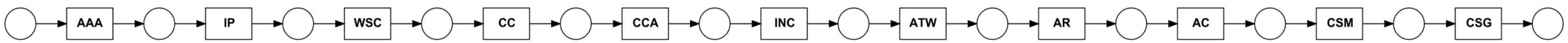


Figure 4.4: Mined process model with 42% activities and 47% paths

An important conformance issue can be seen regarding the "RCP" task. Referring back to the original planned process in Figure 4.1, this task forms part of a joint-split from task "INC" through to "AC". The mined model, Figure 4.5, shows that "RCP" mainly occurs after the "AR","ATW" and "AC" segment and not after "INC".

Another loop in the process model can be identified with the activities "CC" and "RCC" in Figure 4.5. While the "CC" activity is the preferred activity according to the planned model, the "RCC" activity is still part of normal process operations and its presence is not surprising. The "ONHLD2015" task is simply a place holder for instances where the process was stopped in 2014 and was planned to continue in 2015. This shows that a majority of the cases were interrupted after "WSC".

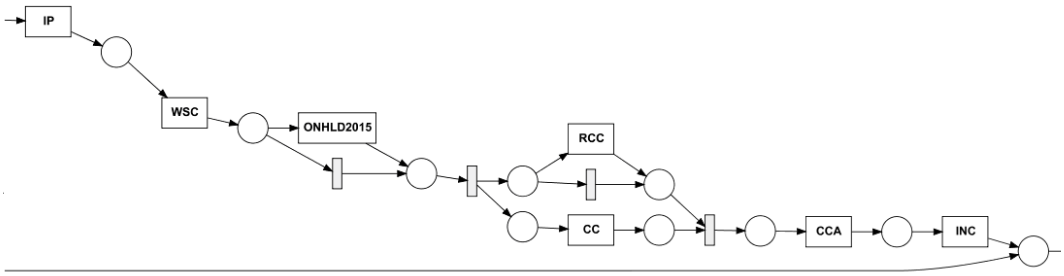


Figure 4.5: Process middle examination.

Figure 4.6 shows the beginning portion of the process model. While the majority of processes start with the "AAA" task, the model in this case does illustrate that there are large inconsistencies when considering lower frequency traces. Loops are also present early on in the processes life cycle which is indicative of processes in some cases struggling to progress. This difficulty is however minimal when finally reaching the "IP" activity. As some processes are cancelled or put on hold, complexity arises in the start as showed in the presence of the "CL", "CA" and "RWO" activities. The majority of waiting activities occur in the beginning of the process, mostly attributed to outside costing approval.

To understand the behaviour of the process model and the real world activities, a summary was compiled of the event and their occurrences. For descriptions of the activities refer to Appendix A. The summary of all activity occurrences

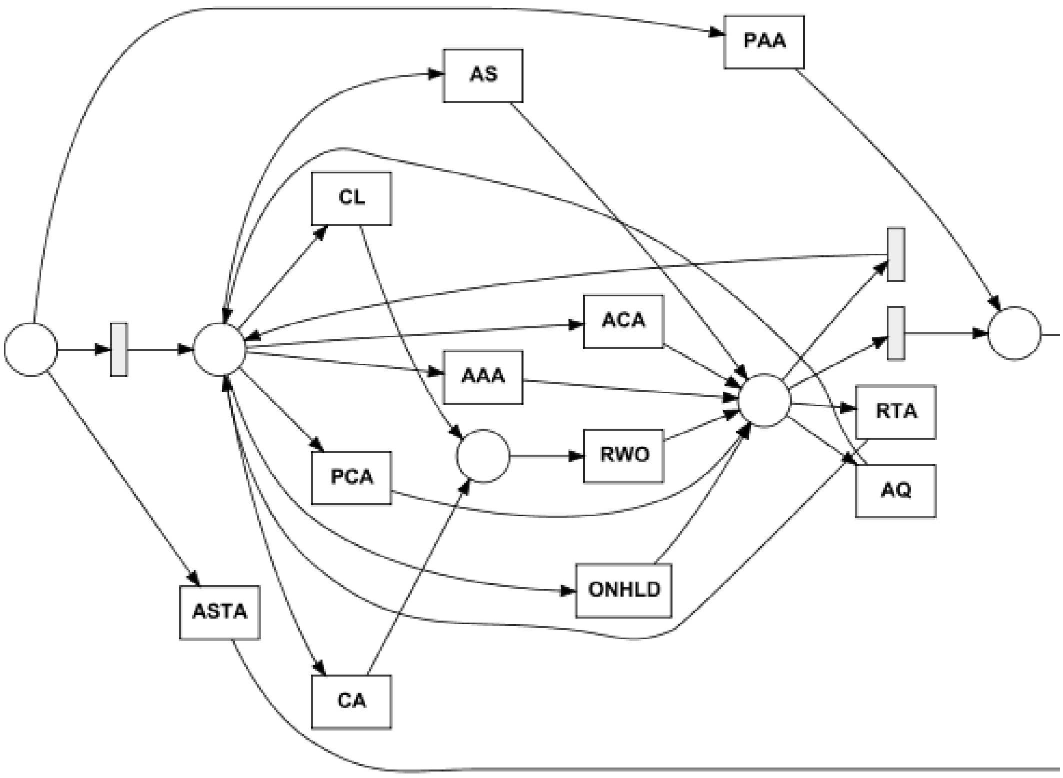


Figure 4.6: Process beginning examination.

is shown in Table 4.6 where it can be seen that the top 11 activities are all activities based on the preferred planned process. It is only from there onwards that undesired events start occurring, at relatively low percentages. The highest of the undesired activities ("RCC", "RCP") both involve the rejection of costs from outside the organisation and therefore count as environmental factors which are outside the scope of control. "RCC" involves the rejection of costing provided by the contractor while "RCP" involves the rejection of a process from a call centre. These types of activities that deal with environmental factors are unavoidable in most processes and their presence should not cause concern. Table 4.6 also shows that 2818 cases were closed by activity "CL". It can then be said that 879 instances were completely shut down by activity "CL" while some of the remaining instances were re-opened by activity "RWO". It should further be noted that activity "WSC" is shown to be a start event in some cases. This is because of projects being approved before 2015. It was decided to not filter out these cases as this would have caused 2381 instances to be lost while there might be valuable information embedded in those instances.

Table 4.6: Log summary of event occurrence.

		Total Occurrences		Occurrence as Start Event		Occurrence as End Event	
Rank	Event Name	Absolute	Relative	Absolute	Relative	Absolute	Relative
1	CC	25015	12.30%	-	-	-	-
2	CSM	18948	9.32%	-	-	-	-
3	CCA	16802	8.26%	-	-	-	-
4	WSC	16745	8.24%	2381	15.06%	-	-
5	ATW	16301	8.02%	-	-	-	-
6	AR	16049	7.89%	-	-	-	-
7	AC	16030	7.89%	-	-	-	-
8	INC	15408	7.58%	-	-	-	-
9	CSG	15053	7.40%	-	-	14930	94.44%
10	IP	14136	6.95%	1326	8.39%	-	-
11	AAA	10143	4.99%	9586	60.64%	-	-
12	RCC	10023	4.93%	-	-	-	-
13	RCP	4284	2.11%	-	-	-	-
14	CL	2818	1.39%	-	-	879	5.56%
15	PAA	2520	1.24%	2516	15.92%	-	-
16	RWO	1920	0.94%	-	-	-	-
17	ASTA	501	0.25%	-	-	-	-
18	AS	246	0.12%	-	-	-	-
19	PCA	218	0.11%	-	-	-	-
20	ACA	72	0.04%	-	-	-	-
21	CA	52	0.03%	-	-	-	-
22	ONHLD	11	0.01%	-	-	-	-
23	ONHLD2015	5	0.00%	-	-	-	-
24	AQ	4	0.00%	-	-	-	-
25	RTA	4	0.00%	-	-	-	-
26	RWC	1	0.00%	-	-	-	-

4.4.2 Dotted Chart Analysis

To visually interpret the behaviour in the event log a dotted chart analysis was done. As discussed in Subsection 3.3.4.2, the dotted chart displays the process instances along the y-axis and time along the x-axis. Every event is displayed as a dot on the chart coinciding with the process instance and the time at which it occurred. This enables the interpretation which was previously not possible when only looking at the log entries.

Due to the limitations of the software package used for analysis, the dotted charts shown in this Subsection are shown as screen-shots. The data could not be exported to be graphed in a different software environment. As a result if this limitation, the axes and their labels are not clear. For all dotted charts in this Subsection, the x and y-axes were included by hand.

Figure 4.7 shows an overview of the entire log with the case instances sorted according to the date and time of the first event from the top. The time frame for this dotted chart ranges from January 2nd to June 8th. As events on the dotted chart are colour coded, a legend is provided. It can be seen in Figure 4.7, that there was a large concentration of events that occurred in the beginning of January. The rate of occurring processes started to decline as the year progressed up until June.

Figure 4.7 also shows attributes of the triggering of certain events. As some of the activities in the process are triggered by personnel and others by an automated system, it should be self evident in the way they are displayed in a dotted chart. Looking at activities such as "CSM", the triggering always occur at set intervals and simultaneously across all instances. This is indicative of an automated state change or activity scheduling.

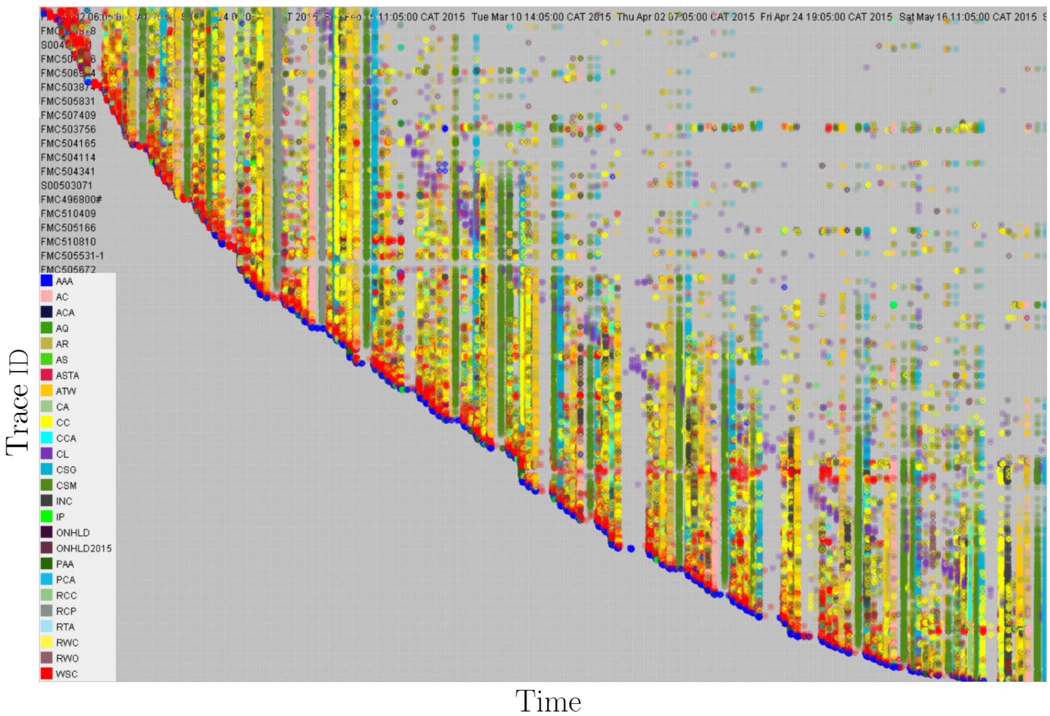


Figure 4.7: Dotted chart with instances sorted by occurrence of first event.

The dotted chart analysis tool allows the manipulation of the time axis in order to view the occurrence of events relative to one another. In other words, as events are usually placed on the exact time instance as they occurred in reality, the starting point of every case is now shifted to where they start at the same time. The time between events remains constant in this time shift. Figure 4.8 shows the same order of cases instances but with the starting points of each case placed relative to one another. The occurrence of events and the length of the instances can now clearly be seen. The main purpose of using a relative time line is to analyse the time it takes for certain events to take place. This argument is illustrated in Figure 4.8 where the activity "CL" normally occurs 1000 hours (41 days) into the process life-cycle. As this activity is bound by a condition in the planned process, it can be seen that this condition is held and is compliant with the planned process.

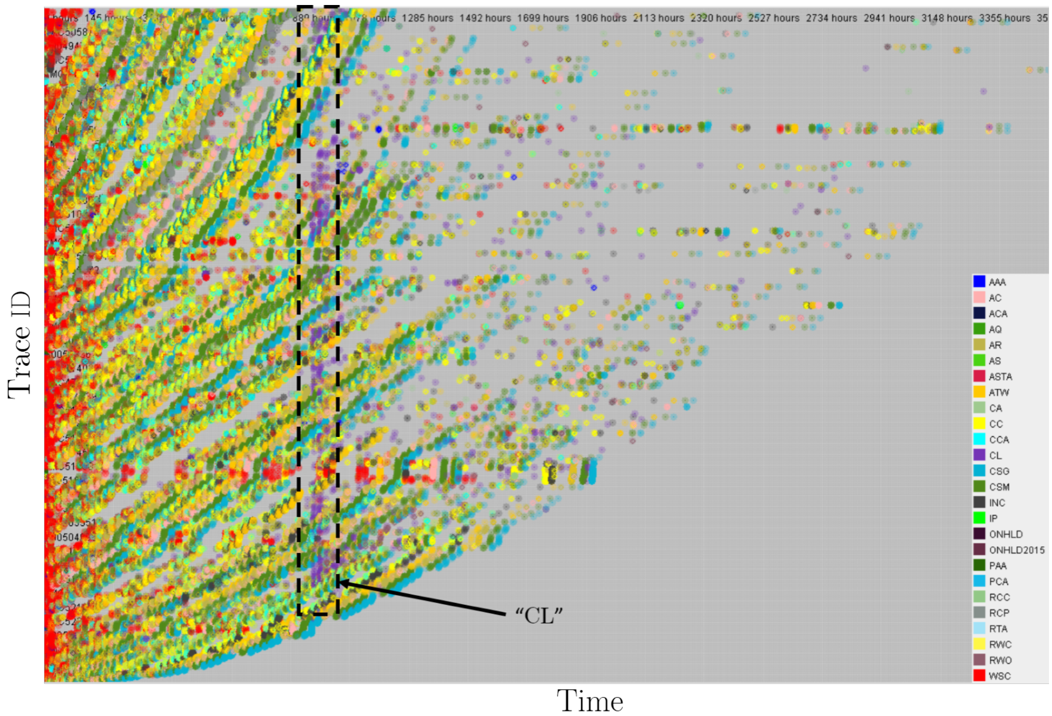


Figure 4.8: Dotted chart with instances viewed relative to one another.

To give an overview of the duration length of the case instances the cases are reordered according to the y-axis. The process instances are still viewed to start relative to one another while the order of instances is changed to order them by their total completion time. The overview of instance durations can be seen in Figure 4.9 where the instances with the shortest duration are shown at the top of the figure while the instances with the longest duration are shown at the bottom. The visual interpretation of Figure 4.9 coincides with the results calculated above and presented in Figure 4.11. When considering that the red dots represent "WSC" and the yellow dots represent "CC", it can be seen that the main cause for long process times is the working time of "WSC".

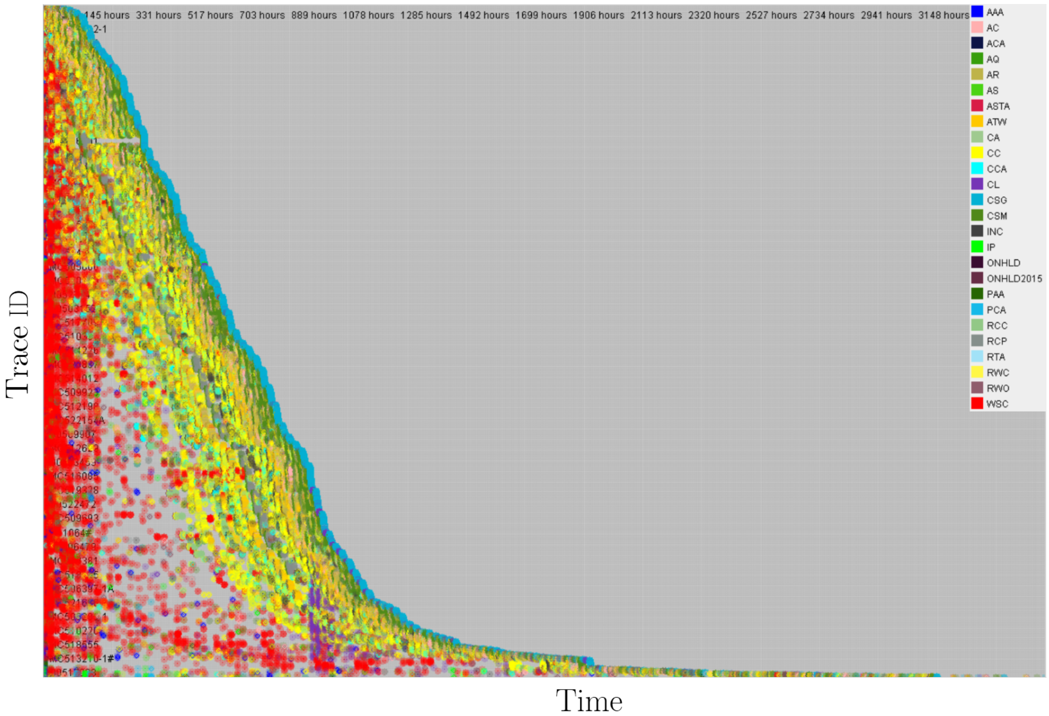


Figure 4.9: Dotted chart with instances sorted by duration.

4.4.3 KPI Analysis

KPI analysis described in 3.3.4.3 shows how the traces within the event log can be analysed to give feedback on certain performance indicators. These indicators can give feedback to management in terms of how the process is performing or how it is changing over time. To calculate the values with regards to the completion times as a monthly KPI, the log replay plug-in in ProM is used as described in Subsection 3.3.4.3. This does not only give the conformance to the process model but also the duration of the cases within the event log. As the event log consists of recordings from January to the end of

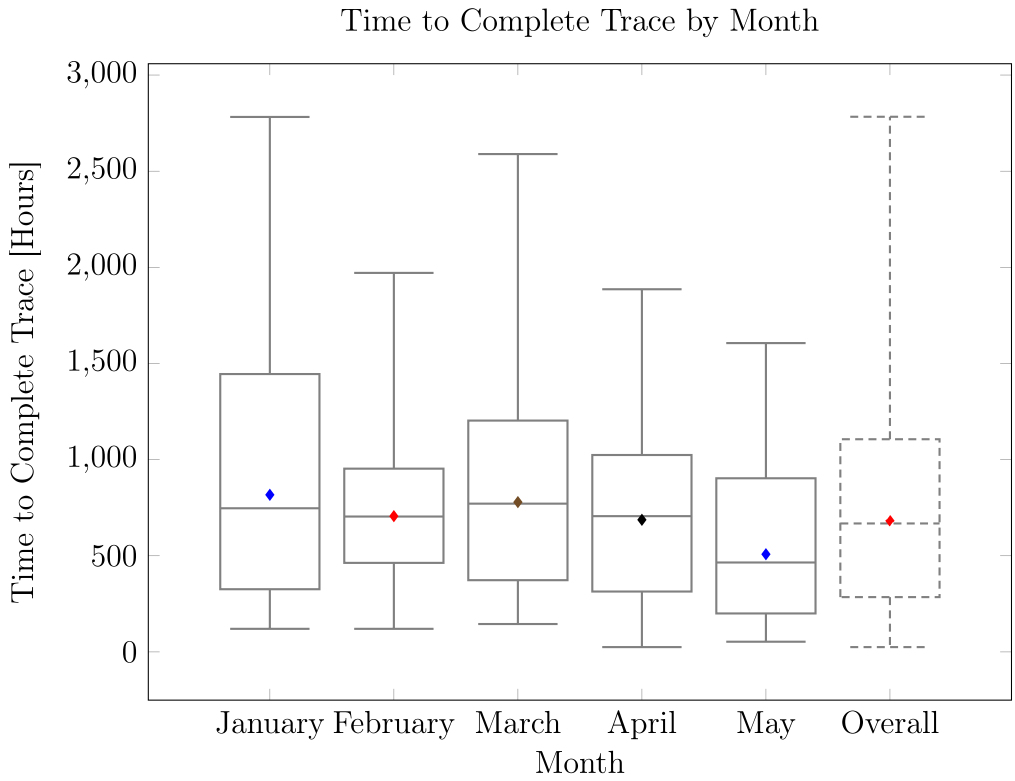


Figure 4.10: Monthly trace times.

May, it is possible to analyse the KPIs with regards to each respective month. The log is divided into separate monthly event logs and analysed separately. The results calculated by the log replay software are presented in Figure 4.10 where it can be seen that process completion times are the worst in January with the highest maximum time, average and spread around the median. The median remains relatively constant throughout while the maximum times follow a downward trend. With averages presented as dots within the box plots, it is seen that they become less over time. It can therefore be said that the completion times of the traces became more densely populated around the median and therefore more reliable with less variation. This downwards trend in the completion times coincides with observations made throughout the dotted chart analysis in Subsection 4.4.2.

After the traces are evaluated, analysis is conducted on the individual tasks within the traces. Analysis on the individual tasks give greater insight into what is happening within the processes and the factors influencing completion times. Using ProM's basic performance metric analysis, the average working

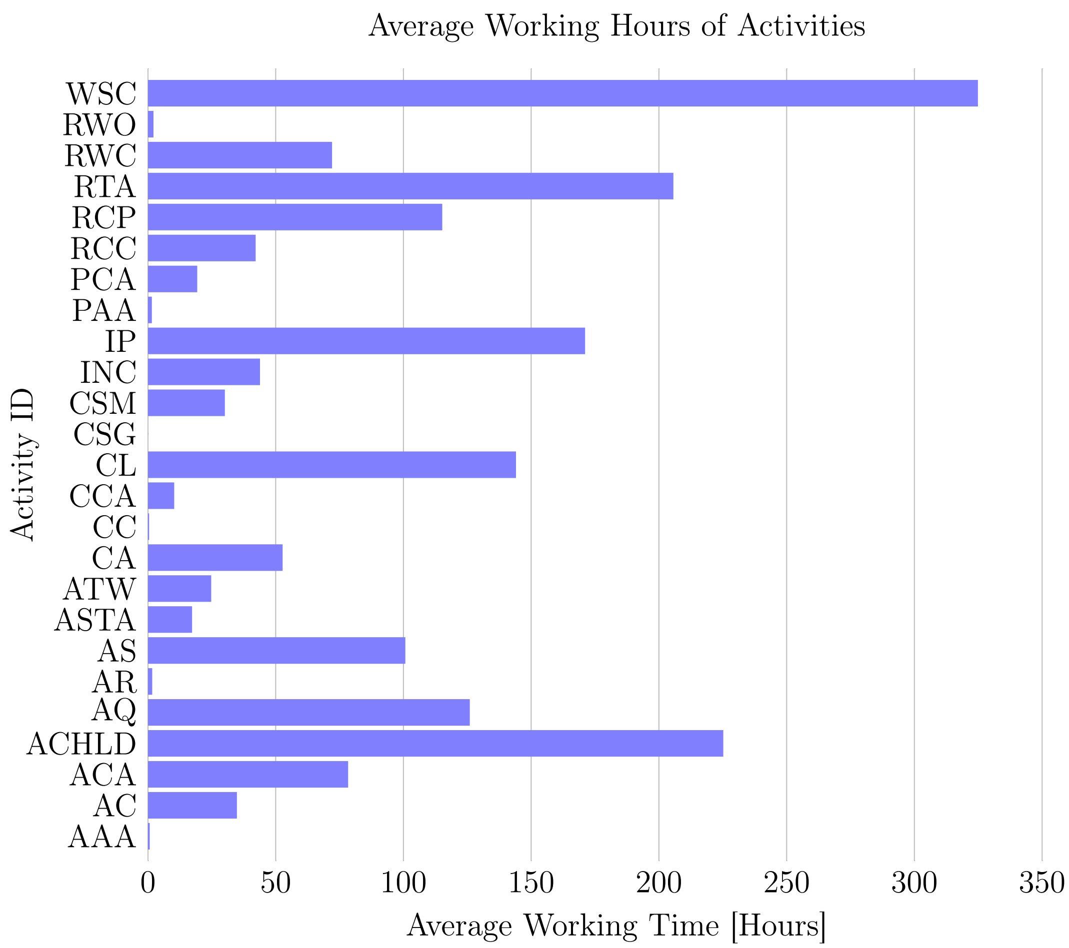


Figure 4.11: Average working times of activities.

time of the activities is calculated. These values can be seen in Appendix D, Table D.1. Figure 4.11 shows a summary of the average working times of different activities involved in the instances.

As can be seen in Figure 4.11, it is clear that "WSC", "RTA", "IP", "CL" and "ACHLD" are the five activities with the longest working times. It is however not the case that the organisation can simply focus on the activity with the longest duration and consider that activity to be the bottleneck within the process. First, the activity might be outside the boundaries of the organisation and therefore not within the scope of what can be improved. Secondly, the criticality or importance of a given activity might be low. As it takes effort and organisational resources to reduce working times in most cases, it is important to assess which activities are most critical to instance completion times. To do this, the average working times of the activities are combined with the frequency by which they occur. As there is no objective mark by which to judge individual activity duration times, the values are normalised to indicate the most critical activity at that stage compared to the others. Equation 4.4.1 shows how the values are first normalised. Equation 4.4.2 then shows how the criticality values are calculated. In these equations, x_i is the tasks' average working time, $min(x)$ is the minimum time within the subset and $max(x)$ the maximum time within the same subset.

This can be seen in Figure 4.12.

$$Normalisation = \frac{x_i - min(x)}{max(x) - min(x)} \tag{4.4.1}$$

$$Activity\ Criticality = Normalised\ Average\ Working\ Time \times Normalised\ Frequency \tag{4.4.2}$$

Figure G.1 in Appendix G shows the calculated values for the normalised average working times and the normalised frequency. The normalised criticality is shown in Figure 4.12 where it is seen that the activities which offer the largest incentive for improvement are "WSC" and "IP". The main consideration here is that the priority of improving the completion time of an activity should be justified by its frequency of occurrence. For example, putting forth effort to improve activities "RTA" or "ACHLD" will not be worth it since they do not occur enough for it to make an impact on overall instance times.

Even though Figure 4.12 shows that activity "WSC" is a bottleneck in most cases and that improving it will improve overall performance, it is outside the boundaries of the organisation. This does not mean that it cannot be improved, but rather that this might lie outside the scope of what can be controlled realistically. "IP" would then be the next most likely candidate to be involved in improvement. The incentive to improve "IP" becomes larger when considered that it is part of every process as a precursor to "WSC".

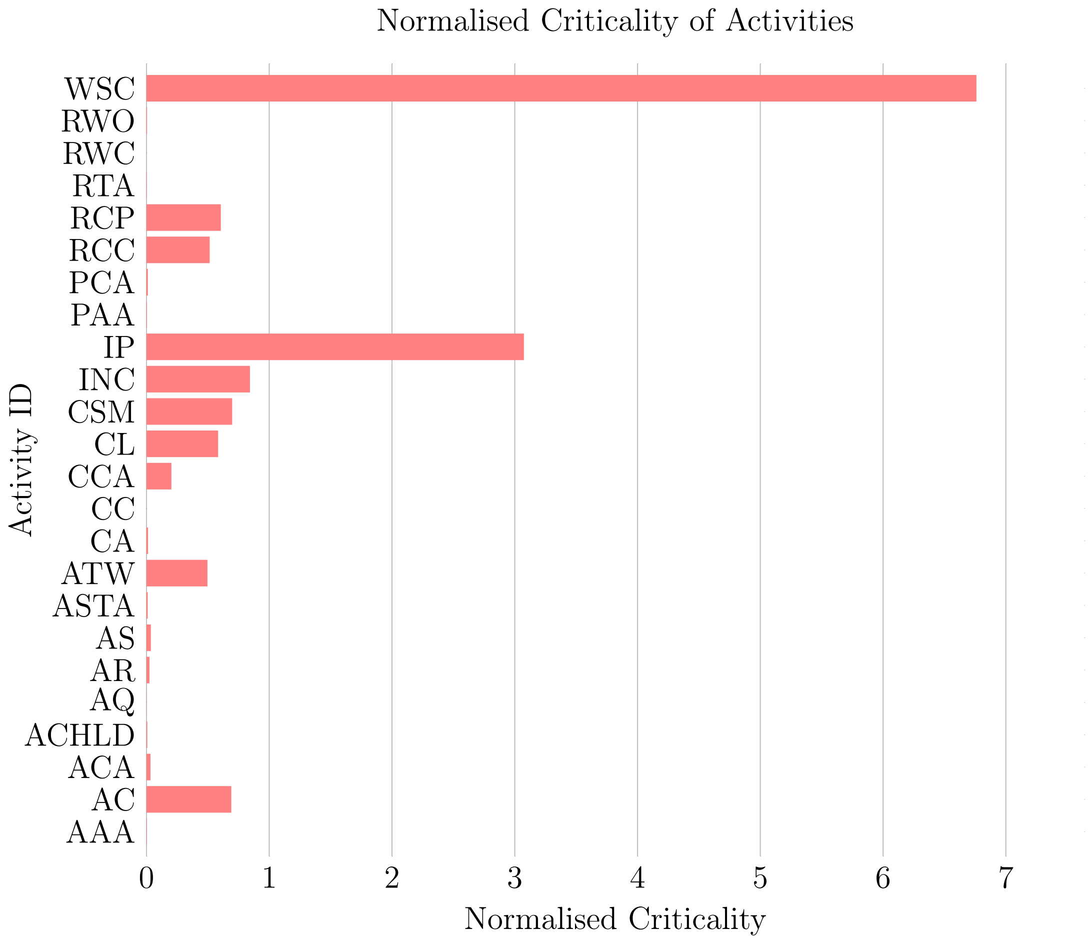


Figure 4.12: Activity impact on process completion time.

4.5 Discussion of Results

As the application of the methodology is complete, the results obtained can be summarised. When considering the mining of a process model from the data provided, it is seen that the there are a small amount of processes which conform to the planned process. The number of recorded instances that coincided with the planned process only forms part of 11.53% of the entire log. The remaining recorded instances ranged from slight deviation to very low compliance instances.

Implementation of the IvM software plug-in provided the ability to construct models based on the historic data provided by the PAM service provider. The use of the IvM software plug-in eased the mining of the process model and provided the ability to construct good models when considering the metrics shown in Table 4.4, Figure and Figure . The model mined by the IvM plug-in offered good reliability and good precision as shown in Table 4.4.

By utilising the dotted chart analysis, along with the KPI analysis, it was determined that the bottleneck for most of the process instances is the "WSC" activity. While this activity is shown to be the most critical considering it's value seen in Figure 4.12 where it has the highest criticality score, it is also be the most difficult to improve.

The KPI analysis further shows that the process performance improved during the five months from the start of 2015. Mean process times are reduced by almost 38%, while the maximum process time slowed down by 42%. The process time also became more constant with less variation around the mean.

4.6 Process Mining Methodology Validation

Chapter 3 presents an application methodology to model and analyse PAM supporting processes. In this methodology, the most applicable analytical techniques are identified from the process mining literature. The aim of only selecting certain techniques is to make the methodology practical and appropriate towards the PAM environment. The techniques should also offer desirable outcomes for aiding decision making in PAM.

Validation encompassed the application of process mining to a case study by using the presented application methodology. The data for analysis is obtained through a PAM service provider where a process is monitored by their EAMS software. A mixed approach is used, when referring to Subsection 3.3.1.3, where the data is explored while attempting to improve a process.

By applying process mining using the application methodology a model could be obtained regarding the flow of activities as they occur in the real world. The first iteration of this model is too complex as a result of the amount of noise within the data. A suitable model is obtained when filtering is used as can be seen in Figure 4.5. Complexity in this model occurs as a result of deviations from the planned process. Filtering the data further reveals the planned process as seen in Figure 4.5. Implementing various analysis techniques discussed in Subsection 3.3.4, problem areas within the real world implementation of the process are identified. The extraction of a model relating the real world activities and the results obtained by the KPI analysis are able to serve as decision support for process improvement.

The application of the methodology shows that the methodology is practical to implement within the PAM environment and that it offers insightful performance metrics based on supporting processes. The methodology is therefore able to aid management to make strategic decisions based on the results presented by applying the methodology presented in this study.

4.7 Chapter Conclusion

This chapter presented a case study where the methodology presented in Chapter 3 was applied to data provided by a PAM service provider. The aim of this case study is to confirm the validity of process mining in the PAM environment.

The chapter opened with a discussion on how data was obtained for the case study and the process on which the data is based. Background information is then given on the process to help gain an understanding of the results presented towards the end of the chapter. The data is anonymised for the most part and no exact details were given with regards to the operation thereof. Anonymisation did not influence the primary application of the methodology but did have some impact on the scoping of the process and constructing realistic objective to achieve.

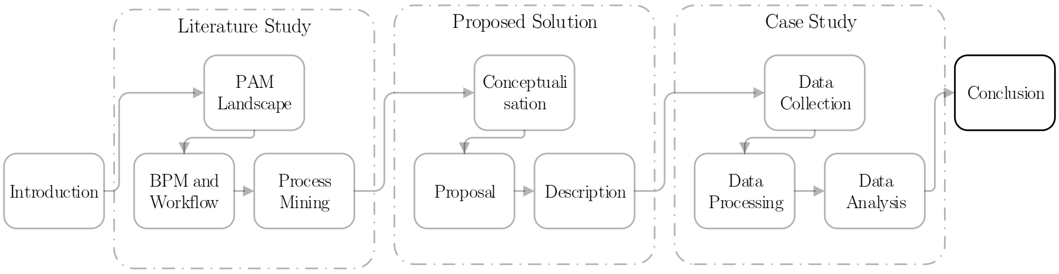
Section 4.4 begins with the importation of the data into ProM. This was then followed by a brief summary report. This summary showed an overview of the contents of the data with regards to the number and types of cases. With insight obtained from the summary, the data is filtered to suit the needs of the study. The case study then utilised different software plug-ins within ProM for the analysis where the different plug-ins coincided with the different steps of the methodology. With the analysis complete and the results obtained, a discussion followed where an overview is given towards the knowledge obtained throughout the analysis and the recommendation for operational support.

The application of process mining resulted in the validation of the application methodology and the answering of the research question. This was achieved as a result of the methodology being able to grant insightful results with regards to the performance of the studied process and applicability within the PAM environment.

Chapter 5

Closure

This chapter aims to summarise the research findings and bring forth a conclusion to the study. The chapter includes a discussion regarding the limitation of the presented methodology and also presents the recommendation for further research.



Chapter Outcomes
•Obtain an understanding of the study as a whole.
•Acknowledge the limitations of the study.
•Answer the research question.
•Establish recommended future research.
•Final discussion.

5.1 Summary and Conclusion

As the popularity of PAM has grown in today's competitive industrial environment, it is necessary to examine PAM strategies and identify opportunities for improvement. Industry has implemented PAM as a tool that enables the organisation to improve maintenance and operations of its physical assets. With the focus primarily on the maintenance activities, certain aspects of value creation have been neglected. This thesis explores the idea of using BPM as a supporting discipline within PAM to assist in improving PAM strategy processes. Supporting PAM strategies with BPM aims to improve the output and ultimately, the value that can be obtained from the organisation's physical assets.

The literature review in Chapter 2 describes the fundamental principles of PAM, BPM, WfM and process mining as an application platform. Within the literature review, limitations were identified as well as the potential opportunities presented by BPM and process mining to address these limitations. Identification of applicable techniques and tools are summarised on which the application methodology is based.

Chapter 3 presents the methodology used to apply process mining to a PAM process. The chapter entails the selection of different techniques and tools from the literature review, which are deemed suitable for the PAM environment. Selected tools are discussed in detail to gain an understanding of the principles and reasons for their selection. The methodology concludes with a discussion on how the results obtained through process mining supports the goals and objectives of the organisation. This is mainly done through operational support as discussed in Subsection 3.3.5.

To validate the aforementioned methodology, a case study is conducted in Chapter 4. Data for the case study is provided by a PAM service provider where a maintenance process is described that operates within the petrochemical industry. Data is analysed and results are obtained which offer insight into the real world operations of the process and show improvement opportunities which can be addressed by management. The aim of the validation was to, not only, acknowledge that the research objectives were met, but also that the research question in Section 1.3 had been answered. The results show that the methodology is applicable within the PAM environment and that, based on the results, improvement can occur. Consequently, the research question is answered, which results in the completion of the final research objective.

5.2 Limitations

As with every research study and the application of a proposed methodology, limitations are present. The application of process mining in the PAM environment unveiled the following limitations:

- The basis for the application of process mining implies that there is a reliable historical event log present on which the analysis can be based. Without a historical log, process mining cannot be performed.
- Some types of analysis techniques are dependent on the level of data availability in the event log. Social network analysis for example requires not only the person responsible for logging the event but also the people who conducted the work.
- The value of performing process improvement by using process mining techniques can be difficult to quantify. The assumption is that process improvement leads to higher performance and thus higher value.
- As a result of the limitations set forth by the organisation that provided the data, the scope of the case study was rather limited. This included the exact details of the operations of the process and its interactions.
- The data provided omitted detailed descriptions of the interactions on the activities within the process. This omission resulted in the lack of an analysis discussing the cause of bad performance and recommendations to mitigate performance issues.

As the study progressed, the above mentioned limitations were encountered. The limitation did however not prevent objectives from being met. The following section provides recommendations for further studies based on the results of the study and the limitations presented above.

5.3 Recommendations for Future Research

Recommendations for future research are put forth in this section. The recommendations listed below are based on the knowledge gained through the case study and the limitations presented in the previous section. Future research may encompass the following:

- Future research might include an investigation into the improvement of event log capture and storage systems for use in the PAM environment.
- In this thesis, the control-flow perspective is considered within the project level of decision making as discussed in Sub-subsection 2.2.4.1 and Subsection 2.1.2 respectively. Future work might focus on different perspectives or decision levels.
- As this study focussed on the processing of process mining data and the presentation of the results, future studies can extend or shift the research scope. This scope can then include working with management to implement and measure the obtained process improvement.
- The improvement of processes involved with the support of physical assets do not covey a direct correlation to added value. Future research might investigate the measure by which physical assets improve their value contribution to the organisation as their supporting processes improve.

The suggestions presented above give recommendations on future work for the improvement of work done in this thesis. It also presents new areas of research that complements this thesis in the improvement of asset performance within the PAM environment.

List of References

AAMCoG (2012). Guide to integrated strategic asset management. Tech. Rep., AAMCoG.

Adriansyah, A., van Dongen, B.F. and Van der Aalst, W.M. (2011). Conformance checking using cost-based fitness analysis. In: *Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International*, pp. 55–64. IEEE.

AguilarSaven, R.S. (2004). Business process modelling: Review and framework. *International Journal of production economics*, vol. 90, no. 2, pp. 129–149.

Akerkar, R. (2014). *Big Data Computing*. CRC Press.

Al-Mashari, M., Al-Mudimigh, A. and Zairi, M. (2003). Enterprise resource planning: A taxonomy of critical factors. *European journal of operational research*, vol. 146, no. 2, pp. 352–364.

Amadi-Echendu, J. (2004). Managing physical assets is a paradigm shift from maintenance. In: *Engineering Management Conference, 2004. Proceedings. 2004 IEEE International*, vol. 3, pp. 1156–1160. IEEE.

Amadi-Echendu, J.E., Willett, R., Brown, K., Hope, T., Lee, J., Mathew, J., Vyas, N. and Yang, B.-S. (2010). What is engineering asset management? In: *Definitions, concepts and scope of engineering asset management*, pp. 3–16. Springer.

Awad, A., Decker, G. and Weske, M. (2008). Efficient compliance checking using bpmn-q and temporal logic. In: *Business Process Management*, pp. 326–341. Springer.

Becker, J., Delfmann, P., Eggert, M. and Schwittay, S. (2012). Generalizability and applicability of model based business process compliance checking approaches a state of the art analysis and research roadmap. *BuR-Business Research*, vol. 5, no. 2, pp. 221–247.

Becker, J., zur Muehlen, M. and Gille, M. (2002). Workflow application architectures: classification and characteristics of workflow-based information systems. *Workflow handbook*, pp. 39–50.

Blache, K.M. (2010). Phd:benchmarking a better understanding. benchmarks shed light on maintenance & reliability perceptions.

Botha, A. (2015). Is iso 55000 an oxymoron, or merely the inner circle of asset management? Tech. Rep., Pragma.

Brannen, L. (2015 Jun). Upfront: The 33 1/3 percent strategy solution. `http://businessfinancemag.com/bpm/upfront-33-13-percent-strategy-solution`.

Brown, C., Vessey, I. *et al.* (2003). Managing the next wave of enterprise systems: leveraging lessons from erp. *MIS Quarterly Executive*, vol. 2, no. 1, pp. 45–57.

Brown, E.R. and Humphrey, G.B. (2005 May/June). *Asset Management for Transmission and Distribution.* IEEE power and energy magazine.

BSI, I. (2014). Bs iso 55000 series asset management. Tech. Rep., ISO.

BSI, P. (2008). 55-2: Asset management. part 2: Guidelines for the application of pas 55-1. *British Standards Institution.*

Burt, R.S., Minor, M.J. and Alba, R.D. (1983). *Applied network analysis: A methodological introduction.* Sage Publications Beverly Hills, CA.

Cardoso, J. (2006). Process control-flow complexity metric: An empirical validation. In: *Services Computing, 2006. SCC'06. IEEE International Conference on*, pp. 167–173. IEEE.

Cardoso, J. (2008). Business process control-flow complexity: Metric, evaluation, and validation. *International Journal of Web Services Research (IJWSR)*, vol. 5, no. 2, pp. 49–76.

Cardoso, J., Mendling, J., Neumann, G. and Reijers, H.A. (2006). A discourse on complexity of process models. In: *Business process management workshops*, pp. 117–128. Springer.

Casati, F., Ceri, S., Pernici, B. and Pozzi, G. (1995). Conceptual modeling of workflows. pp. 341–354. Springer-Verlag.

Casati, F. and Shan, M.-C. (2002). Semantic analysis of business process executions. In: *Advances in Database TechnologyEDBT 2002*, pp. 287–296. Springer.

Chapple, E.D. and Sayles, L.R. (1961). *The measure of management: designing organizations for human effectiveness*. Macmillan.

Chareonsuk, C. and Chansa-ngavej, C. (2010). Intangible asset management framework: an empirical evidence. *Industrial Management & Data Systems*, vol. 110, no. 7, pp. 1094–1112.

Chen, H., Chiang, R.H. and Storey, V.C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, vol. 36, no. 4, pp. 1165–1188.

CIGRE (2002). Guidelines for life management techniques for power transformers. Tech. Rep., CIGRE Working Group.

Clauset, A., Newman, M.E. and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, vol. 70, no. 6, p. 066111.

Coalition, W.M. (1996). Terminology & glossary. Tech. Rep., Technical Report WFMC-TC-1011.

Colaço, M.J. and Dulikravich, G.S. (2009). A survey of basic deterministic, heuristic and hybrid methods for single objective optimization and response surface generation. *Thermal Measurements and Inverse Techniques*, vol. 1.

Coolahan Jr, J.E. and Roussopoulos, N. (1983). Timing requirements for time-driven systems using augmented petri nets. *Software Engineering, IEEE Transactions on*, , no. 5, pp. 603–616.

Cowan, R. and Jonard, N. (2004). Network structure and the di'usion of knowledge. *Journal of Economic Dynamics and Control*, vol. 28.

Creswell, J.W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Criscuolo, P., Nicolaou, N. and Salter, A. (2012). The elixir (or burden) of youth? exploring differences in innovation between start-ups and established firms. *Research Policy*, vol. 41.

Davenport, T.H. (2005). The coming commoditization of processes. *Harvard business review*, vol. 83, no. 6, pp. 100–108.

Davis, J. (2007). What is asset management and where do you start? *Journal American Water Works Association*, vol. 99, no. 10, pp. 26–+.

Dayal, U., Castellanos, M., Simitsis, A. and Wilkinson, K. (2009). Data integration flows for business intelligence. In: *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 1–11. Acm.

De Leoni, M. and Van der Aalst, W.M. (2013). Data-aware process mining: discovering decisions in processes using alignments. In: *Proceedings of the 28th annual ACM symposium on applied computing*, pp. 1454–1461. ACM.

De Weerdt, J., De Backer, M., Vanthienen, J. and Baesens, B. (2011). A critical evaluation study of model-log metrics in process discovery. In: *Business Process Management Workshops*, pp. 158–169. Springer.

DiCesare, F., Harhalakis, G., Proth, J., Silva, M. and Vernadat, F. (1993). *Practice of Petri nets in manufacturing*. Springer.

DoD, U. (2003). Dod modeling and simulation (m&s) verification, validation and accreditation (vv&a). *United States Department of Defense Instruction*, , no. 5000.61.

Douwe P. Flapper, S., Fortuin, L. and Stoop, P.P. (1996). Towards consistent performance management systems. *International Journal of Operations & Production Management*, vol. 16, no. 7, pp. 27–37.

Dwyer, M.B., Avrunin, G.S. and Corbett, J.C. (1999). Patterns in property specifications for finite-state verification. In: *Software Engineering, 1999. Proceedings of the 1999 International Conference on*, pp. 411–420. IEEE.

El Kharbili, M., de Medeiros, A.K.A., Stein, S. and Van der Aalst, W.M. (2008). Business process compliance checking: Current state and future challenges. *MobIS*, vol. 141, pp. 107–113.

Elgammal, A., Turetken, O., van den Heuvel, W.-J. and Papazoglou, M. (2010). Root-cause analysis of design-time compliance violations on the basis of property patterns. In: *Service-Oriented Computing*, pp. 17–31. Springer.

Ellis, C.A. and Nutt, G.J. (1980). Office information systems and computer science. *ACM Computing Surveys (CSUR)*, vol. 12, no. 1, pp. 27–60.

Ellis, C.A. and Nutt, G.J. (1996). Workflow: The process spectrum. In: *Proceedings of the NSF Workshop on workflow and Process Automation in Information Systems*, pp. 140–145. Citeseer.

Evans, P.B. and Wurster, T.S. (1999). Strategy and the new economics of information. In: *Harvard Business Review*.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, vol. 17, no. 3, p. 37.

Ferreira, D.R. and Alves, C. (2012). Discovering user communities in large event logs. In: *Business Process Management Workshops*, pp. 123–134. Springer.

Fire, M., Puzis, R. and Elovici, Y. (2013). Organization mining using online social networks. *arXiv preprint arXiv:1303.3741*.

Fliegner, W. (2014). Extracting process-related information from erp systems for process discovery. *Research in Logistics & Production*, vol. 4.

Flintsch, G. and Bryant, J. (2006*a*). Asset management data collection for supporting decision processes. *US Department of Transport, Federal Highway Administration, Washington, DC*.

Flintsch, W.G. and Bryant, J. (2006*b*). Asset management data collection for supporting decision processes. Tech. Rep., US Department of Transportation.

Frolov, V., Ma, L., Sun, Y. and Bandara, W. (2010). Identifying core functions of asset management. In: *Definitions, Concepts and Scope of Engineering Asset Management*, pp. 19–30. Springer.

Georgakopoulos, D., Hornick, M. and Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and parallel Databases*, vol. 3, no. 2, pp. 119–153.

Ghose, A. and Koliadis, G. (2007). *Auditing business process compliance*. Springer.

Giblin, C., Liu, A.Y., Müller, S., Pfitzmann, B. and Zhou, X. (2005). Regulations expressed as logical models (realm). In: *JURIX*, pp. 37–48.

Girvan, M. and Newman, M.E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826.

Governatori, G., Milosevic, Z. and Sadiq, S. (2006). Compliance checking between business processes and business contracts. In: *Enterprise Distributed Object Computing Conference, 2006. EDOC'06. 10th IEEE International*, pp. 221–232. IEEE.

Gregg, D.G., Kulkarni, U.R. and Vinzé, A.S. (2001). Understanding the philosophical underpinnings of software engineering research in information systems. *Information Systems Frontiers*, vol. 3, no. 2, pp. 169–183.

Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M. and Shan, M.-C. (2004). Business process intelligence. *Computers in Industry*, vol. 53, no. 3, pp. 321–343.

Gruhn, V. and Laue, R. (2006). Patterns for timed property specifications. *Electronic Notes in Theoretical Computer Science*, vol. 153, no. 2, pp. 117–133.

Günther, C.W. and Rozinat, A. (2012). Disco: Discover your processes. In: *BPM (Demos)*, pp. 40–44. Citeseer.

Gupta, B. (2009). Master data management: Managing data as an asset. Tech. Rep., ciber.

Haas, R., Hudson, W.R. and Zaniewski, J.P. (1994). *Modern pavement management*.

Halstead, M.H. (1977). Elements of software science.

Han, K.J. and Narayanan, S.S. (2007). A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system. In: *INTERSPEECH*, pp. 1853–1856. Citeseer.

Han, Y. (1998). *HOON-A Formalism Supporting Adaptive Workflows*. University of Georgia, Department of Computer Science.

Hartmann, P.M., Zaki, M., Feldmann, N. and Neely, A. (2014). Big data for big business? a taxonomy of data-driven business models used by start-up firms. University of Cambridge.

Hastings, N.A.J. (2009). *Physical asset management*. Springer Science & Business Media.

Hennig, I.K.W. (1934). Wesen und aufgabe der betriebswirtschaftlichen organisationslehre. In: *Einführung in die betriebswirtschaftliche Organisationslehre*, pp. 1–8. Springer.

Hevner, A.R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, vol. 19, no. 2, p. 4.

Hlupic, V. (2003). Business process modelling: potential benefits and obstacles for wider use. *International Journal of Simulation: Systems, Science and Technology*, vol. 4, no. 1-2, pp. 62–67.

Holland, J.H. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press.

Holt, A.P. (2000). Management-oriented models of business processes. In: *Business Process Management*, pp. 99–109. Springer.

Hornix, P.T. (2007). Performance analysis of business processes through process mining. *Master's Thesis, Eindhoven University of Technology.*

Hrebiniak, L. (2005). Business strategy: execution is the key. *Financial Times.*

Hudson, W.R., Haas, R. and Uddin, W. (1997). *Infrastructure Management-Design, Construction, Maintenance, Rehabilitation, Renovation.*

IAM (2011). Asset management an anatomy. Tech. Rep., The Instritute of Asset Management.

IBM (2009). Enabling the benefits of pas 55: The new standard for asset management in the industry. Tech. Rep., IBM.

IEEE (2015). Ieee cis task force on process mining. Http://www.win.tue.nl/ieeetfpm/doku.php.

Indulska, M., Green, P., Recker, J. and Rosemann, M. (2009). Business process modeling: Perceived benefits. In: *Conceptual Modeling-ER 2009*, pp. 458–471. Springer.

Ingvaldsen, J.E. (2011). Semantic process mining of enterprise transaction data.

Jensen, K. (1997). *Coloured Petri nets: basic concepts, analysis methods and practical use*, vol. 1. Springer Science & Business Media.

Jung, Y., Park, H., Du, D.-Z. and Drake, B.L. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization*, vol. 25, no. 1, pp. 91–111.

Kaplan, R.S. and Norton, D.P. (1996). Translating strategy into action: The balanced scorecard. *Harvard Business School Press, Boston.*

Kaplan, R.S. and Norton, D.P. (2001). *The strategy-focused organization: How balanced scorecard companies thrive in the new business environment.* Harvard Business Press.

Kaplan, R.S. and Norton, D.P. (2008). *The execution premium: Linking strategy to operations for competitive advantage*. Harvard Business Press.

Konak, A., Coit, D.W. and Smith, A.E. (2006). Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety*, vol. 91, no. 9, pp. 992–1007.

Koronios, A., Lin, S. and Gao, J. (2005). A data quality model for asset management in engineering organisations. In: *IQ*.

Koulopoulos, T.M. (1995). *The Workflow Imperative: Building Real World Business Solutions*. John Wiley & Sons, Inc., New York, NY, USA. ISBN 0442019750.

Larose, D.T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Lebas, M.J. (1995). Performance measurement and performance management. *International journal of production economics*, vol. 41, no. 1, pp. 23–35.

Leemans, S.J., Fahland, D. and Van der Aalst, W.M. (2014*a*). Discovering block-structured process models from event logs containing infrequent behaviour. In: *Business Process Management Workshops*, pp. 66–78. Springer.

Leemans, S.J., Fahland, D. and Van der Aalst, W.M. (2014*b*). Process and deviation exploration with inductive visual miner.

Lingamaneni, S. (2010). Modelling and execution of asset management process via workflow automation.

Lv, T.-y., Su, T.-x., Wang, Z.-x. and Zuo, W.-l. (2005). An auto-stopped hierarchical clustering algorithm integrating outlier detection algorithm. In: *Advances in Web-Age Information Management*, pp. 464–474. Springer.

Madu, C.N. (2000). Competing through maintenance strategies. *International Journal of Quality & Reliability Management*, vol. 17, no. 9, pp. 937–949.

Mahling, D.E., Craven, N. and Croft, W.B. (1995). From office automation to intelligent workflow systems. *IEEE Expert*, vol. 10, no. 3, pp. 41–47.

Marsan, M.A., Balbo, G., Conte, G., Donatelli, S. and Franceschinis, G. (1998). Modelling with generalized stochastic petri nets. *ACM SIGMETRICS Performance Evaluation Review*, vol. 26, no. 2, p. 2.

McLaughlin, D.J. (1995). Strengthening executive decision making. *Human Resource Management*, vol. 34, no. 3, pp. 443–461.

Mendling, J., Moser, M., Neumann, G., Verbeek, H., van Dongen, B.F. and Van der Aalst, W.M. (2006). *Faulty EPCs in the SAP reference model.* Springer.

Miller, H.G. and Mork, P. (2013). From data to decisions: A value chain for big data. In: *Smart It.* IEEE Computer Society.

Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, vol. 50, no. 2, pp. 159–179.

Milosevic, Z., Sadiq, S. and Orlowska, M. (2006). *Towards a methodology for deriving contract-compliant business processes.* Springer.

Mitchell, J., Hickman, J. and Amadi-Echendu, J. (2007). *Physical Asset Management Handbook.* Clarion Technical Publishers.

Mobley, R.K. (2002). *An introduction to predictive maintenance.* Butterworth-Heinemann.

Mohseni, M. (2003). What does asset management mean to you? In: *Transmission and Distribution Conference and Exposition, 2003 IEEE PES*, vol. 3, pp. 962–964. IEEE.

Motwani, J., Subramanian, R. and Gopalakrishna, P. (2005). Critical factors for successful erp implementation: Exploratory findings from four case studies. *Computers in Industry*, vol. 56, no. 6, pp. 529–544.

Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, vol. 77, no. 4, pp. 541–580.

Murtagh, F. and Contreras, P. (2011). Methods of hierarchical clustering. *arXiv preprint arXiv:1105.0121.*

Namiri, K. and Stojanovic, N. (2008). Towards a formal framework for business process compliance. In: *Multikonferenz Wirtschaftsinformatik*, vol. 259, p. 2008. Citeseer.

Newman, I. and Benz, C.R. (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum.* SIU Press.

OpenText (2012). Information drives asset performance. Tech. Rep., OpenText.

O'Reilly (2012). *Big Data Now.* 1st edn. O'ReillyMedia Inc.

Osterwalder, A. and Pigneur, Y. (2010). *Business Model Generation.* Alexander Osterwalder and Yves Pigneur.

Otley, D. (1999). Performance management: a framework for management control systems research. *Management accounting research*, vol. 10, no. 4, pp. 363–382.

Padmanabhan, V., Governatori, G., Sadiq, S., Colomb, R. and Rotolo, A. (2006). Process modelling: the deontic way. In: *Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling-Volume 53*, pp. 75–84. Australian Computer Society, Inc.

Palmer, N. (2009). Workflow management coalition. In: *Encyclopedia of Database Systems*, pp. 3550–3550. Springer US. ISBN 978-0-387-35544-3.

Pantelias, A. (2005). *Asset Management Data Collection for Supporting Decision Processes.* Master's thesis, Virginia Polytechnic Institute and State University.

PAS, E.P.A.S. (2004). 55-1 & pass 55-2.

Peterson, J.L. (). *Petri net theory and the modeling of systems*, vol. 132.

Petri, C.A. (1962). *Kommunikation mit Automaten.* Ph.D. thesis, UniversitÃ¤t Hamburg.

Petty, M.D. (2010). Verification, validation, and accreditation. *Modeling and simulation fundamentals: Theoretical underpinnings and practical domains*, pp. 325–372.

Piessens, D. (2011). Event log extraction from sap ecc 6.0. *Eindhoven University of Technology, Eindhoven, The Netherlands, Master's thesis.*

Poulymenopoulou, M., Malamateniou, F. and Vassilacopoulos, G. (2003). Specifying workflow process requirements for an emergency medical service. *Journal of medical systems*, vol. 27, no. 4, pp. 325–335.

Quak, B., Smit, J.J. and Gulski, E. (2006). Decision support for maintenance management of high voltage components. In: *Properties and applications of Dielectric Materials, 2006. 8th International Conference on*, pp. 262–266. IEEE.

Rajasekar, J. (2014). Factors affecting effective strategy implementation in a service industry: A study of electricity distribution companies in the sultanate of oman. *International Journal of Business and Social Science*, vol. 5, no. 9.

Ramezani, E., Fahland, D. and Van der Aalst, W.M. (2012). Where did i misbehave? diagnostic information in compliance checking. In: *Business Process Management*, pp. 262–278. Springer.

Reh, F. (2015). What you need to know about key performance indicators. Available at: `http://management.about.com/cs/generalmanagement/a/keyperfindic.htm`

Robson, W. (). Strategic management and information systems: an integrated approach.

Rozinat, A. and Van der Aalst, W.M. (2006). Conformance testing: Measuring the fit and appropriateness of event logs and process models. In: *Business Process Management Workshops*, pp. 163–176. Springer.

Rozinat, A. and Van der Aalst, W.M. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems*, vol. 33, no. 1, pp. 64–95.

Rozinat, D.A. (2015). *Disco User's Guide*. Fluxicon. Retreived 24th of July 2015, http://fluxicon.com/disco/files/Disco-User-Guide.pdf.

Schleicher, D., Fehling, C., Grohe, S., Leymann, F., Nowak, A., Schneider, P. and Schumm, D. (2011*a*). Compliance domains: A means to model data-restrictions in cloud environments. In: *Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International*, pp. 257–266. IEEE.

Schleicher, D., Grohe, S., Leymann, F., Schneider, P., Schumm, D. and Wolf, T. (2011*b*). An approach to combine data-related and control-flow-related compliance rules. In: *Service-Oriented Computing and Applications (SOCA), 2011 IEEE International Conference on*, pp. 1–8. IEEE.

Schmidt, R., Bartsch, C. and Oberhauser, R. (2007). Ontology-based representation of compliance requirements for service processes. In: *Proceedings of the Workshop on Semantic Business Process and Product Lifecycle Management (SBPM 2007)*.

Schuman, C.A. and Brent, A.C. (2005). Asset life cycle management: towards improving physical asset performance in the process industry. *International Journal of Operations & Production Management*, vol. 25, no. 6, pp. 566–579.

Schumm, D., Leymann, F., Ma, Z., Scheibler, T. and Strauch, S. (2010). Integrating compliance into business processes. *Multikonferenz Wirtschaftsinformatik 2010*, p. 421.

Scott, J. (2012). *Social network analysis*. Sage.

Shanks, G., Seddon, P.B. and Willcocks, L. (2003). *Second-wave enterprise resource planning systems: Implementing for effectiveness*. Cambridge University Press.

Solaimani, S. and Bouwman, H. (2012). A framework for the alignment of business model and business processes: A generic model for trans-sector innovation. *Business Process Management Journal*, vol. 18, no. 4, pp. 655–679.

Song, M. and Van der Aalst, W.M. (2007). Supporting process mining by showing events at a glance. In: *7th Annual Workshop on Information Technologies and Systems*, pp. 139–145.

Song, M. and Van der Aalst, W.M. (2008). Towards comprehensive support for organizational mining. *Decision Support Systems*, vol. 46, no. 1, pp. 300–317.

Straub, D.W. and Welke, R.J. (1998). Coping with systems risk: security planning models for management decision making. *Mis Quarterly*, pp. 441–469.

Tapscott, D. (1999). Introduction. In: *Creating Value in the Network Economy*. Harvard Business School Publishing.

Teece, D.J. (2010). Business models, business strategy and innovation. *Long Range Planning*.

Tinniä, M. (1995). Strategic perspective to business process redesign. *Management Decision*, vol. 33, no. 3, pp. 25–34.

Tiwari, A., Turner, C.J. and Majeed, B. (2008). A review of business process mining: state-of-the-art and future trends. *Business Process Management Journal*, vol. 14, no. 1, pp. 5–22.

Troester, M. and Haxholdt, M. (2012). How to manage your data as a strategic information asset. Tech. Rep., sas.

UIC (2010). Guidelines for the application of asset management in railway infrastructure organisations. Tech. Rep., UIC.

Umble, E.J., Haft, R.R. and Umble, M.M. (2003). Enterprise resource planning: Implementation procedures and critical success factors. *European journal of operational research*, vol. 146, no. 2, pp. 241–257.

Van den Honert, A., Schoeman, J. and Vlok, P. (2013). Correlating the content and context of pas 55 with the iso 55000 series. *South African Journal of Industrial Engineering*, vol. 24, no. 2, pp. 24–32.

Van der Aalst, W. (1996*a*). Petri-net-based workflow management software. In: *Proceedings of the NFS Workshop on Workflow and Process Automation in Information Systems*, pp. 114–118. Citeseer.

Van der Aalst, W. (2011). *Process mining: discovery, conformance and enhancement of business processes.* Springer Science & Business Media.

Van der Aalst, W., Adriansyah, A., de Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., van den Brand, P., Brandtjen, R., Buijs, J. *et al.* (2012). Process mining manifesto. In: *Business process management workshops*, pp. 169–194. Springer.

Van der Aalst, W., Van Hee, K., Van der Werf, J.M., Kumar, A. and Verdonk, M. (2011). Conceptual model for online auditing. *Decision Support Systems*, vol. 50, no. 3, pp. 636–647.

Van der Aalst, W. and Van Hee, K.M. (2004). *Workflow management: models, methods, and systems.* MIT press.

Van der Aalst, W.M. (1993). *Interval timed coloured Petri nets and their analysis.* Springer.

Van der Aalst, W.M. (1996*b*). Three good reasons for using a petri-net-based workflow management system. In: *Proceedings of the International Working Conference on Information and Process Integration in Enterprises (IPIC96)*, pp. 179–201. Cambridge, Massachusetts.

Van der Aalst, W.M. (1998). The application of petri nets to workflow management. *Journal of circuits, systems, and computers*, vol. 8, no. 01, pp. 21–66.

Van der Aalst, W.M. (2004). Business process management demystified: A tutorial on models, systems and standards for workflow management. In: *Lectures on concurrency and Petri nets*, pp. 1–65. Springer.

Van der Aalst, W.M., De Beer, H. and van Dongen, B.F. (2005*a*). *Process mining and verification of properties: An approach based on temporal logic.* Springer.

Van der Aalst, W.M., Desel, J. and Oberweis, A. (2000). Business process management models, techniques and empirical studies.

Van der Aalst, W.M. and Dustdar, S. (2012). Process mining put into context. *IEEE Internet Computing*, vol. 16, no. 1, pp. 82–86.

Van der Aalst, W.M., Hirnschall, A. and Verbeek, H. (2002). An alternative way to analyze workflow graphs. In: *Advanced Information Systems Engineering*, pp. 535–552. Springer.

Van der Aalst, W.M., Reijers, H.A. and Song, M. (2005*b*). Discovering social networks from event logs. *Computer Supported Cooperative Work (CSCW)*, vol. 14, no. 6, pp. 549–593.

Van der Aalst, W.M., Reijers, H.A., Weijters, A.J., van Dongen, B.F., De Medeiros, A.A., Song, M. and Verbeek, H. (2007*a*). Business process mining: An industrial application. *Information Systems*, vol. 32, no. 5, pp. 713–732.

Van der Aalst, W.M. and Song, M. (2004). Mining social networks: Uncovering interaction patterns in business processes. In: *Business Process Management*, pp. 244–260. Springer.

Van der Aalst, W.M., Ter Hofstede, A.H. and Weske, M. (2003*a*). Business process management: A survey. In: *Business process management*, pp. 1–12. Springer.

Van der Aalst, W.M. and van Dongen, B.F. (2002). Discovering workflow performance models from timed logs. In: *Engineering and Deployment of Cooperative Information Systems*, pp. 45–63. Springer.

Van der Aalst, W.M., van Dongen, B.F., Günther, C.W., Mans, R., De Medeiros, A.A., Rozinat, A., Rubin, V., Song, M., Verbeek, H. and Weijters, A. (2007*b*). Prom 4.0: comprehensive support for real process analysis. In: *Petri Nets and Other Models of Concurrency–ICATPN 2007*, pp. 484–494. Springer.

Van der Aalst, W.M., van Dongen, B.F., Herbst, J., Maruster, L., Schimm, G. and Weijters, A.J. (2003*b*). Workflow mining: A survey of issues and approaches. *Data & knowledge engineering*, vol. 47, no. 2, pp. 237–267.

Van der Heijden, T. (2012). Process mining project methodology: Developing a general approach to apply process mining in practice.

Van Giessel, M. and Jansen-Vullers, M. (2004). *Process mining in SAP R/3*. Ph.D. thesis, Master's thesis, Eindhoven University of Technology, Eindhoven.

Vasconcelos, A., Caetano, A., Neves, J., Sinogas, P., Mendes, R. and Tribolet, J. (2001). A framework for modeling strategy, business processes and information systems. In: *Enterprise Distributed Object Computing Conference, 2001. EDOC'01. Proceedings. Fifth IEEE International*, pp. 69–80. IEEE.

Vassiliadis, P. and Simitsis, A. (2009). Near real time etl. In: *New Trends in Data Warehousing and Data Analysis*, pp. 1–31. Springer.

Verbeek, H., Buijs, J., Van Dongen, B. and Van der Aalst, W.M. (2010). Prom 6: The process mining toolkit. *Proc. of BPM Demonstration Track*, vol. 615, pp. 34–39.

Von Alan, R.H., March, S.T., Park, J. and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, vol. 28, no. 1, pp. 75–105.

Voorhoeve, M. and Van der Aalst, W. (1997). Ad-hoc workflow: problems and solutions. In: *Database and Expert Systems Applications, 1997. Proceedings., Eighth International Workshop on*, pp. 36–40. IEEE.

Wagner, E., Speranza, E., Rogers, J., Bacon, L., Ispass, A. and Bobchek, C. (2003). Research priorities for successful asset management: A workshop.

Walter, B. (1983). Timed petri-nets for modelling and analyzing protocols with real-time characteristics. In: *Protocol Specification, Testing, and Verification*, pp. 149–159.

Wasserman, S. (1994). *Social network analysis: Methods and applications*, vol. 8. Cambridge university press.

Watts, D.J. and Strogatz, S.H. (1998). Collective dynamics of small-world networks. *Nature*, vol. 393.

Weske, M. (2007). *Business Process Management: Concepts, Languages, Architectures.* Springer.

Winston, W.L. and Goldberg, J.B. (2004). *Operations research: applications and algorithms*, vol. 3. Duxbury press Boston.

Wittwer, E., Bittner, J. and Switzer, A. (2002). The fourth national transportation asset management workshop. *International Journal of Transport Management*, vol. 1, no. 2, pp. 87–99.

Woodhouse, J. (2014). Briefing: Standards in asset management: Pas 55 to iso 55000. *Infrastructure Asset Management*, vol. 1, no. 3, pp. 57–59.

Woodward, D.G. (1997). Life cycle costing theory, information acquisition and application. *International Journal of Project Management*, vol. 15, no. 6, pp. 335–344.

Yen, D.C., Chou, D.C. and Chang, J. (2002). A synergic analysis for web-based enterprise resources planning systems. *Computer Standards & Interfaces*, vol. 24, no. 4, pp. 337–346.

Yu, X. and Wen, Q. (2010). A view about cloud data security from data life cycle. In: *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*, pp. 1–4. IEEE.

Zisman, M.D. (1977). Representation, specification and automation of office procedures.

Zuberek, W. (1991). Timed petri nets definitions, properties, and applications. *Microelectronics Reliability*, vol. 31, no. 4, pp. 627–644.

Zur Muehlen, M. and Allen, R. (2000). Embedded vs. autonomous workflow–putting paradigms into perspective. *Excellence in Practice*, vol. 4, pp. 49–58.

Appendices

Appendix A

Activity Descriptions

Table A.1: Activity descriptions.

Task ID	Task Description
AAA	Awaiting Approval
AC	Approved with Conformation
ACA	Awaiting Approval - Await Cost Approval
ACHLD	Approved - Confirmed - On Hold
AQ	Awaiting Approval - Await Quote
AR	Approved within Region
ARSKM	Await SKM Approval
AS	Awaiting Spares
ASTA	Approved - Region - Sent Out
ATW	Approved Work in Question
CA	Cancelled
CC	Completed Costing
CCA	Completed Costing Approved
CL	Closed
CSG	Closed SAP Extraction Generated
CSM	Closed Sent for Review
INC	Completed and Invoice Checked
IP	Approved and In Progress
ONHLD	Awaiting Approval - On Hold: Next Financial Year
ONHLD2015	Approved - On Hold 2015
PAA	Project Awaiting Approval
PCA	Approved - Project Cost Approved
RCC	Reject Contractor Cost
RCP	Reject Call Centre Process
RIV	Re-invoice
RTA	Awaiting Approval - Retailer to Approve
RWC	Reject - Work Order Costing
RWO	Re-Opened Work Order
WINV	Wrong invoices submitted by contractor
WIP	Work In Progress
WO	On Key Work Order
WSC	Work On Site Completed

Appendix B

Event Log Summary Details

Figure B.1: Event log summary screen-shot from ProM.

Appendix C

Mined Process Inspection

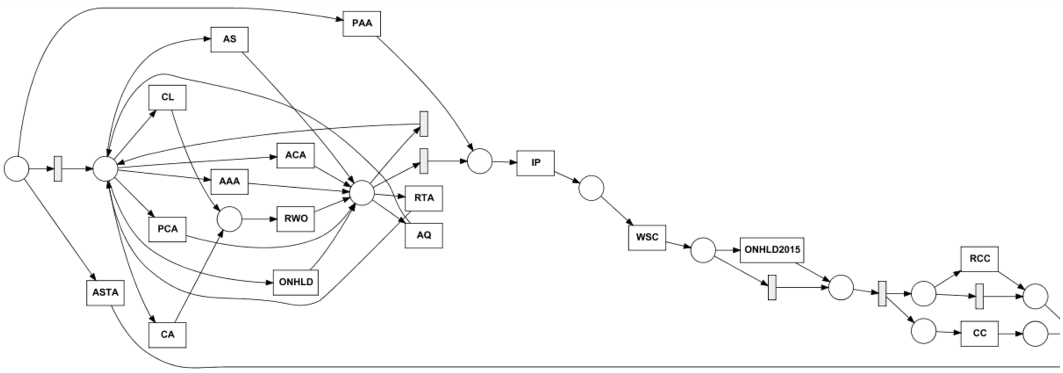


Figure C.1: Initial mined process with 100% activities and 80% available paths (first half).

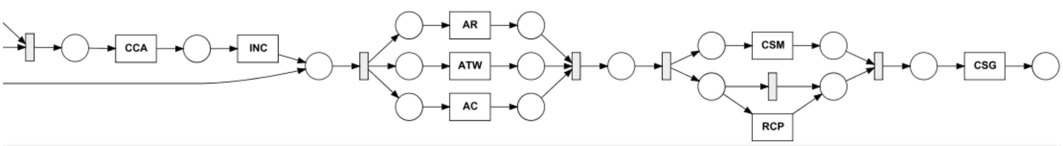


Figure C.2: Initial mined process with 100% activities and 80% available paths (second half).

Appendix D

Activity Performance Metrics

Table D.1: Performance metrics of all activities.

Activity	Average	Frequency	Stadev	Min	Median	Max
AAA	0.703	10742	21.739	0	0.017	1170.433
AC	34.822	16961	18.691	0	36.733	117.317
ACA	78.359	357	226.545	0	0.017	2136.5
ACHLD	225.198	29	241.279	0.017	122.483	503.383
AQ	125.979	11	163.847	0	22.633	431.95
AR	1.676	16980	18.059	0	0.017	1081.85
AS	100.749	305	184.187	0	26.917	1373.1
ASTA	17.273	524	29.751	0	2.1	343.333
ATW	24.763	17241	51.153	0	0.117	846.617
CA	52.711	214	160.653	0	0	1169.067
CC	0.458	26003	1.929	0	0.283	68.3
CCA	10.281	17452	21.231	0	1.067	163.767
CL	144.063	3431	218.052	0	38.75	2850.317
CSM	30.105	19871	24.797	0	28.767	560.55
INC	43.859	16384	52.589	0	27.208	1073.95
IP	171.109	15199	324.386	0	39.333	3239.2
ONHLD	1204.684	46	1164.072	0	1956.283	2669.55
ONHLD2015	299.818	11	423.094	0	0.017	1123.017
PAA	1.528	2891	47.955	0	0.017	2105.817
PCA	19.292	522	110.103	0	0.017	1464.033
RCC	42.16	10419	62.17	0	17.8	1604.267
RCP	115.195	4456	105.947	0	74.5	1413.9
RTA	205.662	13	445.454	0	4.583	1422.383
RWC	72.067	2	101.894	0.017	72.067	144.117
RWO	2.16	2486	43.881	0	0.033	1510.35
WSC	324.868	17580	335.655	0	215.267	2714.567

Appendix E

Detailed Dotted Chart Analysis

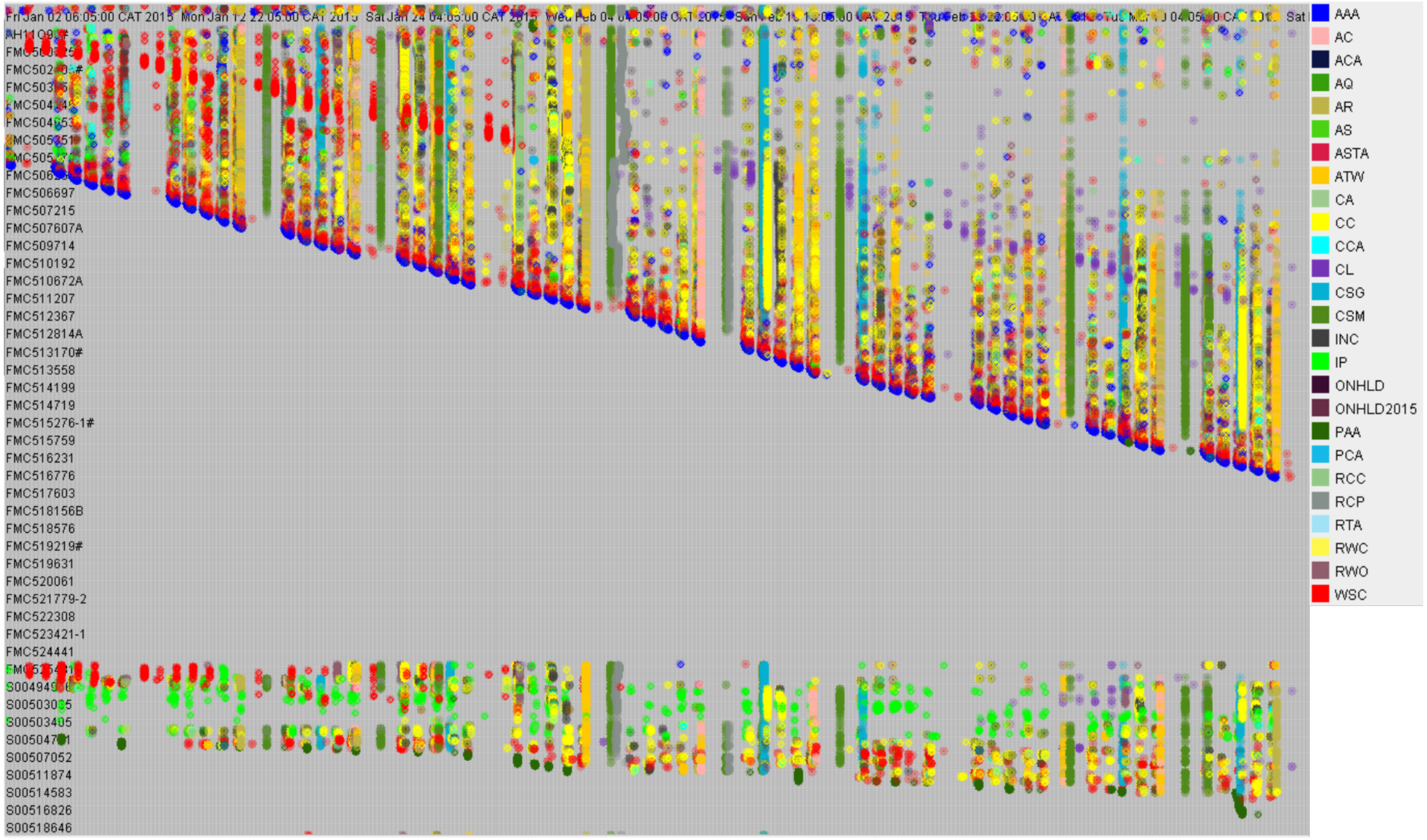


Figure E.1: Dotted chart analysis wide view, first half.

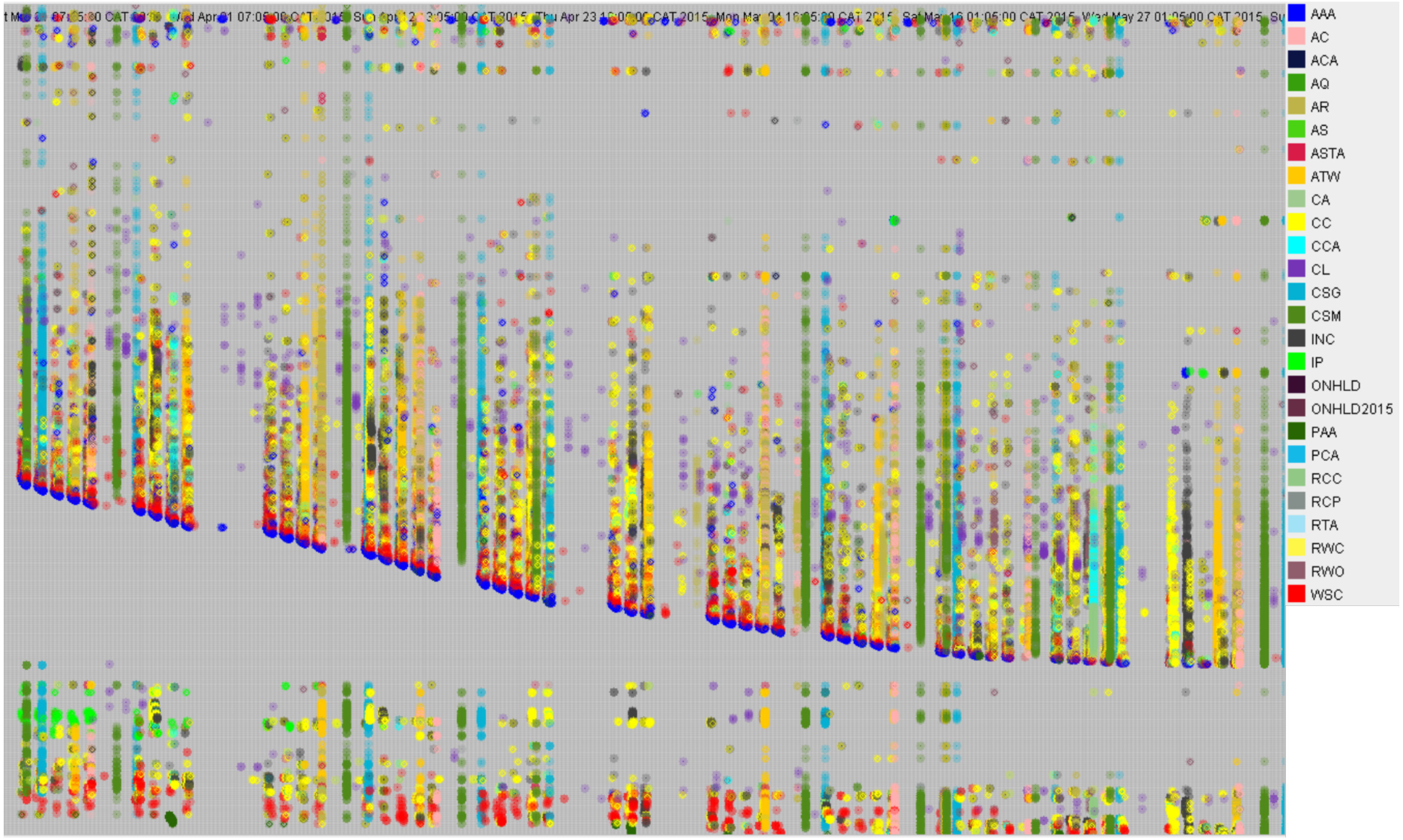


Figure E.2: Dotted chart analysis wide view, second half.

Appendix F

Conformance Checker Settings Descriptions

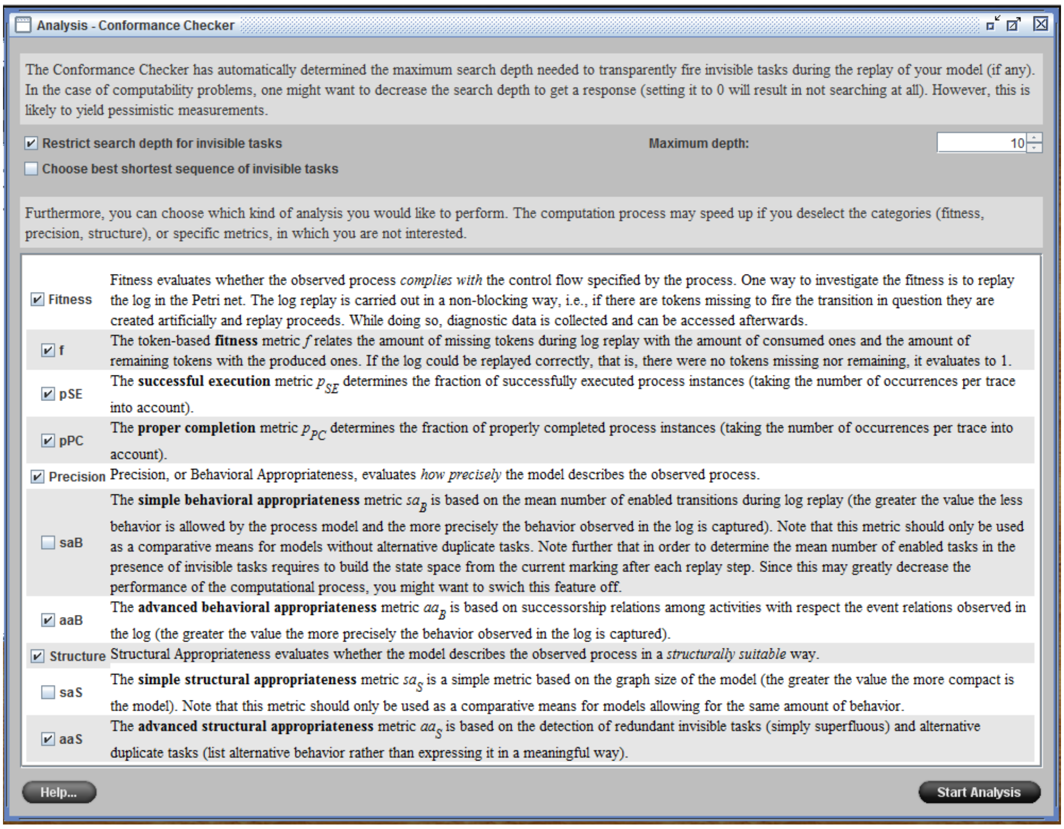


Figure F.1: Conformance checker plug-in metric descriptions.

Appendix G

Activity Impact On Process Completion Times

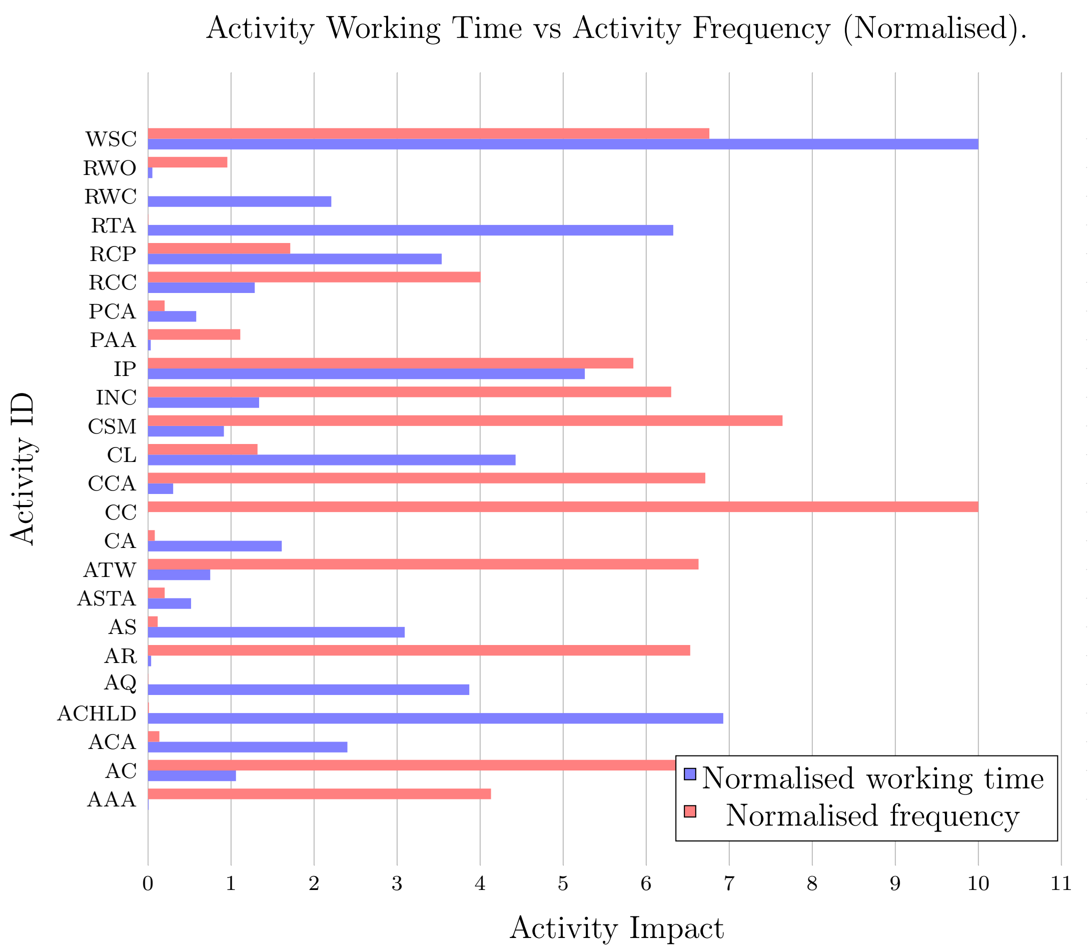


Figure G.1: Normalised working time vs the normalised frequency.