Master Thesis

Uncovering the Runtime Enterprise Architecture of a Large Distributed Organization

A Process Mining-Oriented Approach

Robert van Langerak BSc.

Master Business Informatics
Department of Information and Computing Sciences
Utrecht University

Universiteit Utrecht

Author	Robert van Langerak BSc.
	3643093
	r.p.vanlangerak@uu.nl
	Master Business Informatics
	Utrecht University
First Supervisor	dr. ir. Jan Martijn van der Werf
	J.M.E.M.vanderWerf@uu.nl
	Department of Information and Computing Sciences
	Utrecht University
Second Supervisor	dr. Sergio España Cubillo
	s.espana@uu.nl
	Department of Information and Computing Sciences
	Utrecht University
External Supervisor	Ruud Warmenhoven MSc.
	Ketenanalyse
	PostNL

I. Abstract

Many process analysis approaches exist nowadays. However, due to the complexity of todays business processes and the underlying IT it is hardly possible to visualize the process execution of a business in one overview. Often organizations consist of multiple departments that need to work together to deliver a process. Several approaches exist to analyse multiple departments in isolation, or the social network they form. Process Mining is a way to analyse process execution logs for the discovery of processes, checking the process conformance, and enhancing already existing processes. The visualizations that are part of the results of such analysis are not well designed and hard to understand. Besides, this can not be done in a dynamic real time fashion – only for post execution analysis.

As such, this thesis presents a new way of process visualization, known as the Runtime Enterprise Architecture (REA). The Runtime Enterprise Architecture is a combination of multiple existing visualization elements, presenting the results of multiple process execution analysis. The REA visualizes the flows within real life network or IT systems. It helps an operational process manager in understanding the execution of its operation; and an IT process manager understands how the systems in his process interact.

II. Acknowledgements

After almost one year later, this thesis is finally coming to an end. Many hours of hard work eventually led to this extensive document, that focusses on the application of process mining in large distributed organizations. This thesis marks the end of my academic career, which started in 2010 with my bachelor 'Informatiekunde' and ends now, in December 2016.

I would like to thank my first supervisor, Jan Martijn van der Werf, for his help and support during my graduation research. Now, numerous meetings and e-mails later, this thesis is finished. I would like to thank my second supervisor Sergio España, for all his feedback regarding, but not limited to, Business Process Management in this thesis. Additionally, I want to thank my external supervisor, Ruud Warmenhoven (PostNL) for guiding me through the complex processes of PostNL that are the basis of this study.

At last, I want to thank my mother, who passed away this year.

Robert van Langerak
Utrecht, December 2016.

III. Table of Contents

IV. List of Figures

V. List of Tables

1. Introduction

In the last decades, information systems responsible for supporting and monitoring business processes became increasingly important. These business processes became more complex over time, and nowadays are often cross-departmental or even cross-organizational. These complex processes required a more advanced way of management, modelling, and analysis. Business process management and the accompanying business process model and notation (BPMN) were introduced to answer the demand to model and analyse these processes.

In more recent times, the field of process mining emerged. Process mining focusses on the extraction of useful process information from event logs. This is done by automatically constructing process models that show the observed behaviour found in the event logs (van der Aalst et al., 2007). Process mining could be of great assistance in todays organizations, as it can help them improve their business processes. Some applications of process mining within organizations are known, but in general the technique is not widely used (van der Aalst et al., 2007, 2012). This is a loss as this technique can analyse and improve not only small but also large scale business processes. These complex large scale business processes, also known as organizational business processes, often consist of multiple departmental processes. Such a departmental process is responsible for a part of the execution of the larger organizational process. Together these processes form process architectures: the process structure of an organization.

1.1 Problem statement

Many organizations have modelled their business processes using the business process model and notation, and try to optimize their processes. However, due to the complex nature of these organizational business processes it is hard to identify possible process improvements. Therefore, these organizations should use process mining to analyse and improve these processes. The main problem in the current practice of process mining can therefore be identified as: businesses do not effectively use process mining techniques to improve their organizational process nor the departmental processes underlying this organizational process. This results in non-optimal business execution. Additionally, the available techniques are seldom used to create complete process architectures of the entire business process landscape of an organization. To summarize, the problem addressed in this study can be formulated as follows:

> Todays organizations neither do structured process analysis
> on departmental level to improve organizational processes,
> nor construct Runtime Enterprise Architectures.

In general, this problem has effect on multiple aspects of these organizations. First, they miss out on the opportunity to improve their business processes in a more advanced way using process improvements methods on multiple process levels. Second, these organizations do not use process mining to gain insight in bottlenecks and other performance issues on departmental level. At last, organizations do not visualize their process structure in a dynamic automated way using process architectures, which can be of great use internally.

1.2 Example organization

An example of an organization that does not do structured process analysis is one of the largest parcel distribution companies in the Netherlands. This company is responsible for at least 80% of all parcels that are distributed, which entails that they distribute around 700.000 parcels per day. To facilitate this process, there are 18 distribution centres throughout the Netherlands. These 18 centres are supposed to operate identically, to ensure the quality of the distribution process. This means that each centre executes the same activities, in other words, they are all a specific instance of the same distribution process. To ensure that the parcels are correctly handled, each parcel has a unique identifier, known as a barcode. This barcode is scanned multiple times in the process of distribution, often starting at acceptation and ending with the delivery. The combination of unique scans of one parcel combined are known as a scantrail. The parcels can be seen the as artefacts that travel through the distribution network. When analysing these artefacts it should be possible to identify a distribution network as a system, which can be seen as a process architecture.

The organization analyses these scantrails, for example to identify distribution problems or measure key performance indicators. This is done by analysing only a certain number of events that are present in the scantrails. The analysis therefore is not focussed on the entire process, but on a specific part of it. This is an example of a problem where process mining might come in effectively. By analysing all scantrails as a whole, errors in the departmental distribution process can be identified. This would improve the quality of the measures, and gives possibilities to think of new measures. In this case, the lack of thorough process analysis techniques causes several problems, since the organisation is:

- not able to do all the analyses necessary,
- not able to do innovative analysis that can create new insights,
- supplying incomplete information to clients,
- supplying incorrect information to clients.

New and innovative process analysis techniques can be used to improve situations as described here. Section 1.3 gives an in-depth example of the problem and a possible way to improve this.

1.3 Running Example

This running example clarifies why the problem as formulated in the problem statement indeed is a problem. In general, the lack of use of process analysis techniques results in three major problems:

1. The organization is not able to identify process problems
2. The organization is not able to visualize their process execution
3. The organization is not able to determine process exceptions

As there currently is no data-driven way to analyse and compare the departmental processes, the organization is not able to identify deficiencies or problems in these processes. However, multiple analysis are created for several divisions of the company. Each division is responsible for its own Key Performance Indicators (KPIs). KPIs are relevant measures assigned against certain objectives. They specify the information needed to measure performance (Ward & Peppard, 2002).

The KPIs are calculated based on data, which is created during the distribution process (the previously mentioned scantrail). The execution data is analysed using predefined measures, that result in a KPI. The main problem is that these KPIs all focus on a specific part of the distribution process, instead of the entire distribution process. These KPIs are used to manage people that are responsible for the distribution process, as well as to improve business processes that are part of the distribution process (see Figure 1).

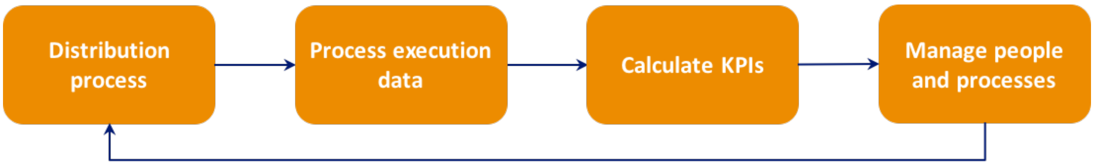


Figure 1: High level overview of Running Example

This entails that if there are errors in the distribution process, they can be seen in the distribution process data. The distribution process data is then used to calculate different KPIs. Those KPIs are used to actively manage the people and centres responsible for the departmental processes. The problem in this case is that the KPIs are insufficient in measuring everything. Additional process analysis techniques that have a different approach can improve this situation. It is important that process analysis techniques can be used to improve the analysis of the distribution processes, in order to improve the distribution process itself.

1.4 Objective, Scope, and Structure

The main objective of this research is to **identify ways businesses can use process analysis techniques to analyse and improve their business processes, and to create process architectures using business process models**. In this study the scope will be on the parcel distribution process of a large distribution company in the Netherlands. The findings will be generalized to address a wider public.

The structure of this study is as follows: in chapter 2, the research method is described. In chapters 3, 4, and 5 the related literature is analysed and reviewed, answering sub questions one, two, and three. In chapter 6, the case study company and its data is presented. Chapter 7 describes the approach that is used in the case study and relates this to the Process Mining Methodology 2. The case study itself is executed in chapter 8, presenting the analysis and results. Chapter 9 uses the results of the case study to create a new Runtime Enterprise Architecture that is able to visualize complex process executions in a well designed and clear way. The Runtime Enterprise Architecture visualization is the main deliverable of this thesis. The answer to the main research question can be found in chapter 10, conclusion. Chapter 11 presents a discussion on the results of this thesis.

2. Research Approach

To study the problem that is identified in the previous chapter, several research questions are specified in section 2.1. The research method and its implementation are discussed in section 2.2. The research design, consisting of the literature study and the case study are explained in section 2.3. Section 2.4 describes the scientific and social relevance of this study.

2.1 Research Questions

Based on the problem statement and running example as described in chapter 1, the main research of this study focusses on the use of process mining to identify static process architectures and the creation of Runtime Enterprise Architectures. These architectures show the relations between high level organizational processes and low level departmental processes, and give insight in process execution and possible improvements. Therefore, the main research question of this study can be formulated as follows:

How can artefact centric process mining be used to create
(dynamic) process architectures?

The final deliverable of this study is an approach that uses artefact centric process models to create organization specific process architectures. The following sub-questions are formulated to answer the main research question and to pave the way for such a process architecture visualization.

SQ1. What is the current state of the business process management field and how is it linked to process architectures?

There is a lot of literature available on business process management. To answer this question, the state of the art existing literature is reviewed, to create an understanding what business process management is and how it can be used. Additionally, business process modelling notations such as BPMN and Petri Nets are described.

SQ2. How can process mining be used to create process models that focus on a specific part of the organizational business process?

To answer this question, literature regarding business process mining is analysed. The analysis will be used to create a general understanding of the topic of process mining and specific implementations.

SQ3. How can business process management and process mining techniques be combined to improve business processes on organizational and departmental level?

Literature regarding process improvement and redesign is analysed to identify process optimisation methods. Together, the first three sub questions create the context of this study. The following sub questions 4, 5, and 6 focus on the execution and results of the case study, the creation of the process architectures and formulating an approach to do this.

SQ4. **What are the necessary tools, techniques, and methods to extract departmental and organizational processes from execution data?**

To answer this sub question, several methods are researched on how to extract processes based on execution data. Several possible applications of process mining will be identified here, as well as how they can be applied in practice.

SQ5. **Is it possible to identify and analyse the lifecycles of process artefacts and to combine departmental and organizational process models to construct process architectures?**

This sub question is answered by identifying the instances of the departmental processes and how they relate to each other as well as to the overall process. The final sub question uses the results of the case study to identify ways of improving the organizational process by improving the departmental processes.

SQ6. **Which steps are necessary to create a dynamic visualization of a process architecture using process models?**

This sub question identifies how a Runtime Enterprise Architecture visualization can be created using the constructed process models and static process architectures. Additionally, such a Runtime Enterprise Architecture visualization will be created. The structure of this thesis is defined based on these research questions, and is visualized in Figure 2.

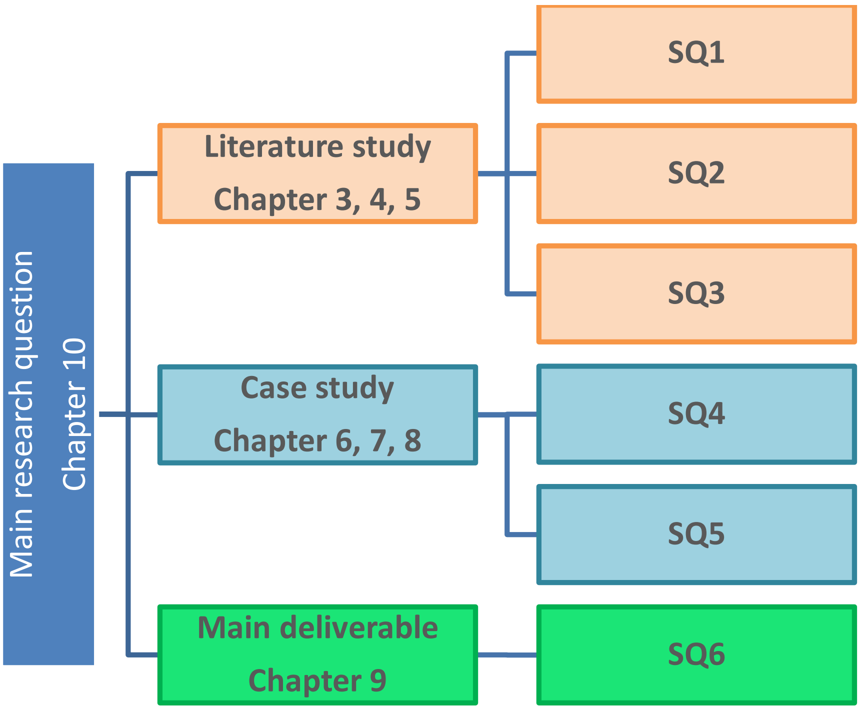


Figure 2: Visual representation of the structure of the study

2.2 Research Method

The research in this study will be performed according to the design science method, as described by Hevner, March, Park, & Ram (2004). Design science can be seen as a problem-solving method. A lot of research has been done regarding design science, for example by Hevner (2007), Peffers, Tuunanen, Rothenberger, & Chatterjee (2007), and Wieringa (2009). Wieringa (2009) identifies design science as a method to solve a set of problems, of which the top level problem is a practical problem. Additionally Wieringa (2009) states that a logical structure for solving practical problems is the regulative cycle. The regulative cycle is a general structure of a rational problem solving process, as described by van Strien (1997) and Wieringa (2006, 2009). The regulative cycle defines part of the structure of this study.

The regulative cycle contains four main phases: Real world problem investigation, Requirements analysis, Solution design, and Implementation (instantiation). Each phase of the cycle is linked to a certain set of activities. Figure 3 below is an adapted version of this regulative cycle, known as the design science cycle, which also incorporates Hevners' design science theory (van der Werf, 2015). The problem investigation can result in different underlying reasons, which lead to results in different investigation approaches. The main goal of the first phase is to describe and explain the problem. If possible, a statement can be given regarding what will happen if nothing is done with the described problem.

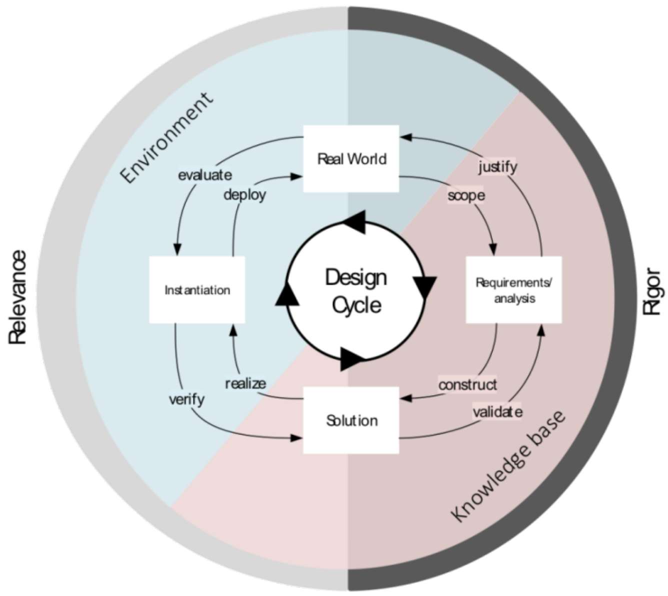


Figure 3: The design science cycle (van der Werf, 2015)

In this research, the problem investigation is a combination of a problem and solution driven approach. The research is problem-driven as there is no way to dynamically generate process architectures of complex business processes. Problem driven investigation is applied when a problem is experienced that needs to be diagnosed. It is useful to formulate and test hypotheses concerning the causes of the experienced problems, and to identify a priority in solving the problems. In addition, the investigation is solution-driven as the knowledge is present that there are new techniques that could enable more extensive analysis and visualisations. Solution driven investigation starts from possible solutions, and tries to identify problems that can be solved with a certain accompanying technology. Useful technologies can be analysed in advance, so a selection can be made based on certain functionalities.

The second phase of the design science cycle, known as requirements analysis, is concerned with the activity of designing an artefact that tries to specify how certain actions can close the gap between the experienced situation and the requested situation (from *as is* to *to be*). The main question is what is needed to solve the problem. To answer this question, a literature study and a case study are executed. Using the results from these studies, the artefact can be designed. This can for example be a model, a natural language specification, or a prototype.

In the subsequent phase solution design, the specified design is validated. This is done by investigating how to apply or adapt the solution in such a way that it can be used to solve the problem. The solution is valid if it brings the organization closer to their goals. The validation can be done by asking three questions (Wieringa, 2009):

1. Would this design, implemented in this problem context, satisfy the criteria identified in the problem investigation?
2. How would slightly different designs, implemented in this context, satisfy the criteria?
3. Would this design, implemented in slightly different contexts, also satisfy the criteria?

In the final phase, solution implementation, the proposed and validated solution is implemented. As this is the design science cycle, the process then starts again from the beginning. This enables incremental improvements. The design validation is executed using a case study. As mentioned before, the solution is valid if it brings the stakeholders who experienced the initial problem closer to their goals.

2.3 Research Design

The research design is an implementation of the research method, and uses a framework as presented by Verschuren & Doorewaard (2010). This study consists of a literature study, a case study, and a validation. The literature study presents the current state of business process management and process mining, and some additional literature. The results from this study are used in the design of the case study. The case study is focussed on the construction of process models and process architectures based on execution data. A new approach needs to be identified to correctly create process architectures using process mining. The approach is validated with process managers of the case study company. This design can be presented in the format as described by Verschuren & Doorewaard (2010). This framework presents a way to structure research consisting of four steps: theory, conceptual model to analyse, results, recommendations or improvements. When this approach is converted to the situation at hand, the framework is constructed as shown in Figure 4.

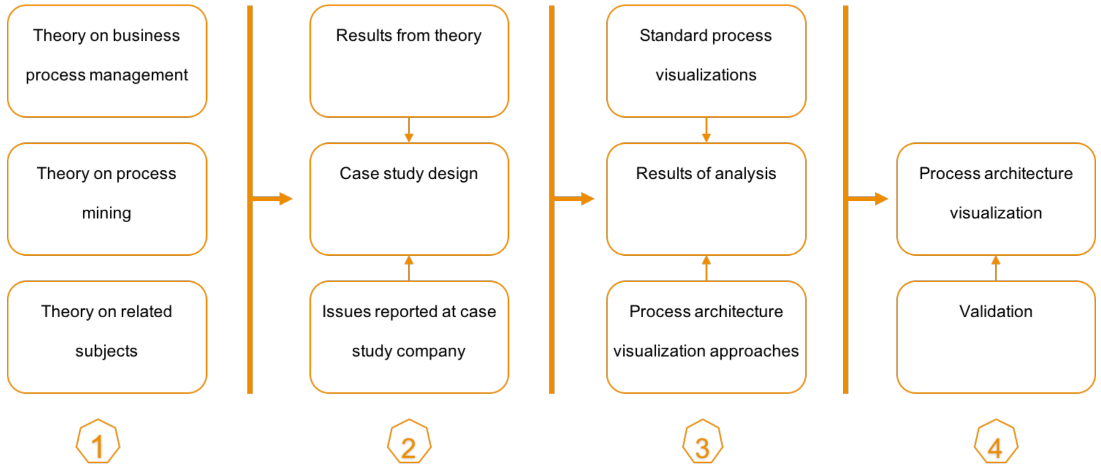


Figure 4: Research design framework

First, the literature study is executed to construct a complete overview of the theory (1). The theory is then used to design the case study, and to identify what has to be analysed (2). The third part is the review of the analysis results (3). Combining these results into a new kind of process architecture visualization and the accompanying validation is the final part of the framework, as well as the final deliverable of this thesis (4).

2.3.1 Literature study

The literature study mentioned before will review the state of the art literature for the topics business process management and process mining. This is done by starting with several well known papers in the field. Using the online computer science bibliography DBLP (Ley, Herbstritt, & Ackermann, 2015), several relevant studies were identified. In the last few years, DBLP has grown “to be the world's most comprehensive open bibliographic data service in computer science” (Ley et al., 2015). Based on this initial set of papers, additional literature was identified using the snowballing technique – which entails using the references of selected papers to find new papers (Wohlin, 2014). The total set of papers was used in the literature review, to answer sub questions one, two, and three.

2.3.2 Case study design

The case study in this research will be designed to identify process structures and process models using execution data. These models and structures are used to create insight and to construct a new kind of process architecture visualization. The organizational process (known as the distribution process) consists of 18 instances of the same departmental process. Each departmental process consists of three processes (see Figure 5): the sorting process, the shift process, and the delivery process. A more detailed description of these processes and the case study can be found in chapter 6. Each of these 18 instances can be analysed and compared, to improve the organizational process (the distribution process).

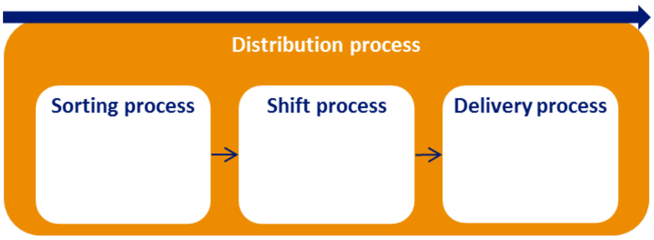


Figure 5: Case study processes

According to Yin (2009) there are three types of case studies: descriptive, exploratory, and explanatory studies. This case study is exploratory and explanatory, as it studies how process instances differ (exploratory) and tries to identify why this happens (explanatory). Besides the types, there are four possible designs for a case study. As mentioned before, multiple instances of the same departmental process are studied. This means that, according to Figure 6, the study is a single embedded case study. There are multiple units of analysis (process instances) in one context (organizational process).

The chain of evidence for this study is based on the structure of research questions of the main research. The case study relates to sub questions four, five, and six of the main research. The protocol for the case study is described in chapter 7, the case study data is extracted and analysed in chapter 8. More on the execution of this case study and the case company itself can be found in chapter 6.

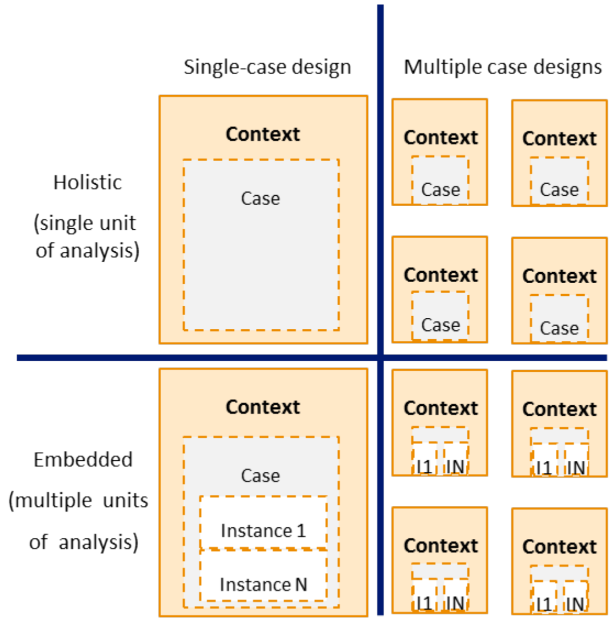


Figure 6: Case study designs (Yin, 2009)

2.3.3 Expert assessment

Using the results of the case study and the related literature, new Runtime Enterprise Architecture visualizations are created to improve the insight companies have in their processes and the relations between these processes. To validate this, the architectures are validated with several process managers of the case study organization.

2.4 Relevance

This study has scientific as well as social relevance. The scientific literature base regarding process mining is extended by this research, in several fields as business process management, process mining in general, and the application of process mining in businesses. For this study, the social relevance is mainly within the parcel handling industries. It enables companies to effectively use process mining in their day to day logistic business, and improve quality for its customers.

2.4.1 Scientific Relevance

The scientific relevance of this research focusses on how companies can apply process mining in their company. A lot of research is available on process mining, but little on the application of it in organizations. Therefore, the case study contributes to the scientific literature in general. Besides, this research adds knowledge on how to use process mining to improve processes, and to dynamically visualize business processes.

2.4.2 Social Relevance

The social relevance of this study is the practical experience gained by using process mining in the parcel industry. This added knowledge on how to apply process mining in daily operations enables organizations to improve their analysis, and make changes to processes on multiple levels. For the case company specifically, the impact of the application of process mining could improve multiple KPIs. Besides, this study can be used a preliminary guide into process mining for other organizations.

3. Business Process Management

In this section literature regarding business process management is discussed. First, business process management in general is described, together with the business process management lifecycle. In the business process management section, business process modelling notations are reviewed as well. Examples of this are the Business Process Modelling and Notation and Petri Nets. In the second part of this chapter process architectures and their use are elaborated on.

3.1 Introduction to Business Process Management

Every company has business processes to create and deliver products or offer services to customers. Van der Aalst (2011) defines a business process as "a collection of inter-related events, activities and decision points that involve a number of actors and objects, and that collectively lead to an outcome that is of value to at least one customer". The construction of these processes affect how well they fulfil the task they were created for. As companies change in this dynamic world, the business processes of the companies (should) change as well. This is not always the case, and could possibly harm the competitive position of a company (Dumas, Rosa, Mendling, & Reijers, 2013). Besides adapting processes to the changing environment, it is important to improve processes to further optimize their execution. If two companies offer the same service, one of them could outperform the other by optimizing the process execution (Dumas et al., 2013).

Business Process Management (BPM) focusses on the improvement of operational business processes, which can be done with or without the use of specific technologies. By modelling and then analysing a business process it is possible to improve a process. However, software to manage, control, support, and simulate these processes can be used to improve the execution (van der Aalst, 2011). BPM is defined as a body of methods, techniques and tools to discover, analyse, redesign, execute and monitor business processes (Fahland, de Leoni, van Dongen, & van der Aalst, 2011; Weske, 2012).

The field of BPM emerged from activities as business process redesign and re-engineering, which were common in businesses around the 1990s. Based on the process models created for redesign, Workflow Management Systems (WfMS) were created (Dumas et al., 2013). These systems distributed the work in a process model over various actors in the company. When a business process was changed, the WfMS could easily assign additional actors to new activities. This was a new way of working, as previously the business rules for executing processes were hard-coded in more complex systems (Dumas et al., 2013). Slowly, the WfMS were extended to monitor and analyse the execution of the processes they assigned actors to. The integration with other systems present in many businesses, such as Enterprise Resource Planning systems, improved. In the end this led to well integrated WfMS with sophisticated functionalities regarding monitoring and analysis of business processes. Such a system is called a Business Process Management System (BPMS).

3.1.1 Business Process Management Lifecycle

Business Process Management covers the entire lifecycle of the business processes. The lifecycle describes how business processes can be managed (see Figure 7). In the identification phase, a business problem is described and the processes relevant to the problem are identified. The result of this is an overall view of the processes in the organization. In phase two, known as process discovery, the relevant business processes are modelled. This is as-is process modelling, since the models describe the current state of the processes (Dumas et al., 2013).

In the subsequent analysis phase, the as-is process is analysed, possibly using performance measures. This results in a collection of problems and issues regarding the current business process. The identified problems and issues are prioritized, based on their impact and the effort to fix them (Dumas et al., 2013). Phase four, process redesign, focusses on the identification of process changes that would improve the issues documented in phase three. This is done by analysing multiple alternatives and comparing them based on the required performance measures. In the end, this results in a to-be process model. More on redesign can be found in section 5.2 Process improvement.

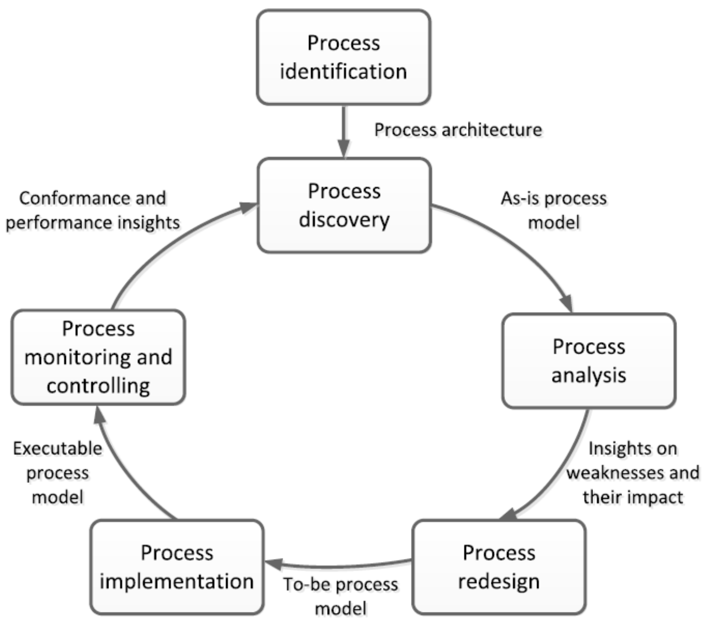


Figure 7: The BPM lifecycle (Dumas et al., 2013)

The implementation phase consists of two main activities: organizational change management and process automation. Organizational change management focusses on changing the way of working and how to involve all the actors of the changing process. It is important that the actors know what the changes are and why they are necessary to improve the process. Process automation relates to automating the suggested process changes in the IT systems that execute the process. When the improved process is in place, it should be monitored and controlled. In addition, analysis is necessary to determine if the improved version of the process is outperforming a previous version. If new issues arise, the lifecycle should start over to improve the process.

3.1.2 Business processes modelling

Business processes consist of events and activities. Activities can be seen as tasks that cost a certain amount of time. Events are things that happen, and do not consume time. An example of an event is the arrival of a container. Checking if the container contains all the ordered products is an activity, as it consumes time to do this (Dumas et al., 2013). Besides events and activities, business processes often contain decisions points. As the name suggests, at this point in the process a decision has to be made that influences the further execution of the process.

Apart from these elements, there are some additional elements that can be present in a business process. The first element is the actors, which can be humans, organizations, or software systems. They can be responsible for an entire business process, or for a certain activity. The second category of additional elements are the physical objects that can be necessary in a part of the process, like equipment, materials, or documents. The last element category contains immaterial objects, which are electronic documents or records.

Multiple notations on modelling business processes exist, for example Petri Nets, Business Process Modelling Notation (BPMN), Unified Modelling Language (UML), and Event driven Process Chains (EPCs) (van der Aalst, 2011). In business environments BPMN is widely used, as it is easy to understand by people that are not familiar with the notation. Figure 8 shows an example of a business process modelled in BPMN, containing the elements as discussed above.

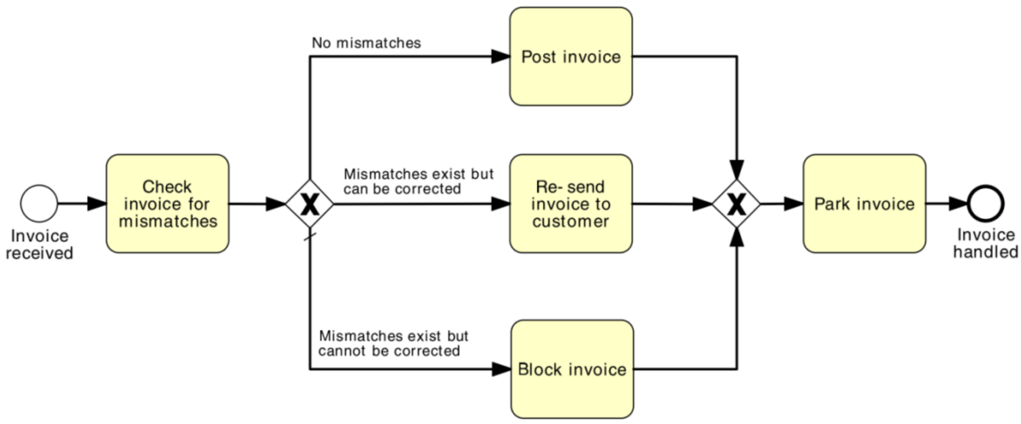


Figure 8: An example of a business process modelled in BPMN (Dumas et al., 2013)

The example shows the business process of handling an incoming invoice in a company. It starts when the invoice is received. The first activity is to check the invoice for mismatches, which results in a decision point (marked as a diamond shape). When there are no mismatches, the process continues to the activity 'post invoice'. When there are mismatches, but they can be easily corrected, the activity 'resend invoice to customer' is undertaken. If there are mismatches in the invoice that cannot be corrected, the invoice is blocked. It does not matter which of the decisions is taken, as in the end they all end with the activity 'park invoice'. Then the business process ends, as the invoice is handled.

An additional way of modelling business processes is using Petri Nets. Petri Nets are a graphical modelling tool that expresses the causal structure of processes, and consists of states and transitions (Murata, 1989). The states and transitions are connected with arrow, and every state has zero or more tokens. The directed graph of a petri net starts with at least one initial state, and based on several conditions the tokens travel over the additional states using the transitions (Girault & Valk, 2001). In graphical representation, places are drawn as circles and transitions as bars or boxes (see Figure 9).

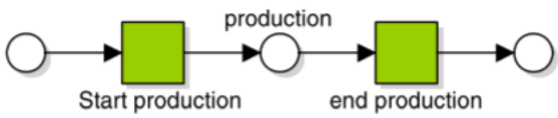


Figure 9: Example of a Petri Net (Hee, Sidorova, & van der Werf, 2013)

Transitions have a number of input and output places that are the pre- and post- conditions of the transitions. The presence of a token in a place is interpreted as holding the truth associated with the place (Murata, 1989).

3.2 Process architectures

As Indulska, Recker, Rosemann, & Green (2009) stated in their paper, business process modelling has emerged as an important instrument for the analysis and design of process-aware information systems. These models turn out to be a guidance to improve the understanding of current business operations and are often used as a foundation for improvement initiatives (Indulska et al., 2009). However, many organizations have numerous business processes that result in a large collection of process models. It is therefore important to choose a correct level of detail for process visualizations and to identify where one process stops and the next one starts. A solution for these issues is the use of process architectures, since they consist of multiple abstraction levels which represent different parts of the process.

Process architectures are part of the business process management lifecycle (as discussed in section 3.1) and are defined as "an organized overview of business processes with relations and guidelines that determine how they must be organized" (Dijkman, Vanderfeesten, & Reijers, 2011). Several studies address the use of process architectures to create a consistent and integrated collection of process models, such as (Green & Ould, 2004, 2005; Joosten, 2000; Koliadis, 2008; Scheer & Nüttgens, 2000).

An example of a process architecture is shown in Figure 10, which includes several processes and their relations. Dijkman et al. (2011) argue that there is no consensus concerning the possible relations that may exist between the business processes. They do note however that often occurring relations in literature are: 1. Decomposition, 2. Specialization, 3. Trigger, and 4. Use. The use of containers can be seen in the example architecture as well. Such a container can visualize a distinction between processes, for example between Primary processes and Support processes.

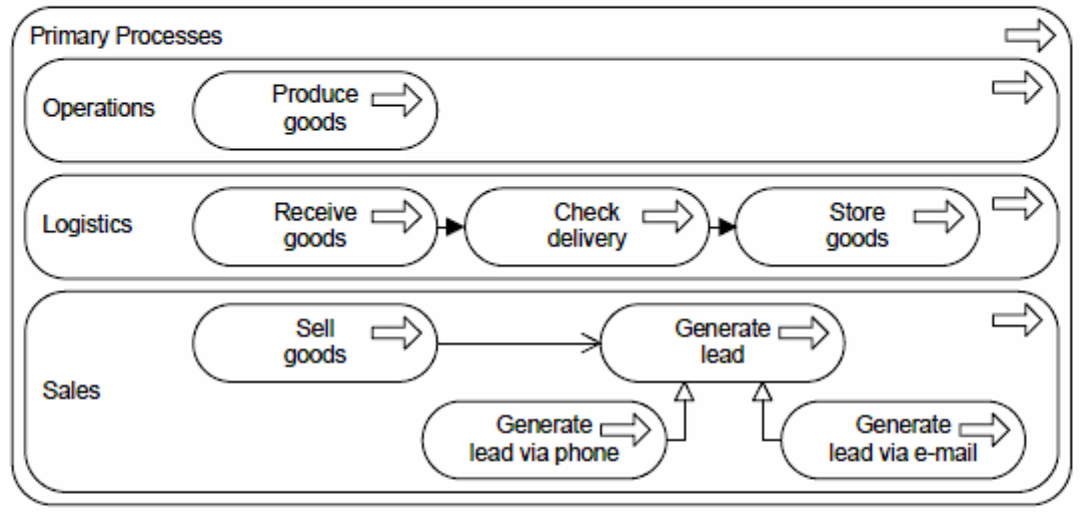


Figure 10: Example of a process architecture (Dijkman et al., 2011)

3.2.1 Approaches for designing a process architecture

Dijkman et al. (2011) executed an literature study in which they identified 48 approaches to construct a process architecture. These approaches where then grouped using the question: `On what basis are processes and their relations identified according to this approach?'. This resulted in the following five categories:

Table 1: Categories of constructing process architectures (Dijkman et al., 2011)

Approach	**Description**
goal based	using a goal structure consisting of business goals and relations, a process architecture is constructed
action based	an action structure consisting of business actions and relations is created first. A business action is a loop of activities that is executed
object based	a business object model is used as the basis for the architecture. The object model is often created in a class diagram of ER diagram.
reference model based	a reference model is adapted to create the new process architecture. Often reference models are used that do not originate from the itself, but from sources as the survey by (Fettke, Loos, & Zwicker, 2006) that analyse multiple business process reference models.
function based	this approach first identifies the business function within the organization, such as procurement. These functions are then grouped hierarchically, after which the grouping is used to create an architecture.

Dijkman et al. (2011) validated the approaches using 39 practioners who were working in the business process management area. The results of the validation are stated in the table below. In the table A stands for agree, N for neutral, and D for disagree.

Table 2: Approach evaluation results, based on (Dijkman et al., 2011)

Approach	Ease of use				Usefulness				Popularity			
	A	N	D	?	A	N	D	?	A	N	D	?
Goal based	30%	43%	22%	5%	39%	25%	28%	8%	6%	19%	33%	42%
Action based	41%	24%	32%	3%	42%	26%	21%	11%	19%	5%	35%	41%
Object based	37%	29%	29%	5%	57%	24%	19%	0%	29%	18%	18%	34%
Reference model based	67%	21%	8%	5%	62%	16%	8%	14%	56%	13%	10%	21%
Function based	45%	34%	16%	5%	50%	32%	8%	11%	38%	21%	13%	28%

It can be concluded that the reference model based approach stood out as the easiest to use (67%), most useful (62%) and most popular approach (56%). After additional validation sessions, it seems that there is a potential for an approach that enables practitioners to (graphically) represent a business process architecture along with related concepts and to apply selected design guidelines automatically.

3.2.2 Process architecture frameworks

Some frameworks exist to design process architectures. The frameworks use a specific design approach as discussed above, or combine some of the approaches. Examples of frameworks are the Architecture of Integrated Information Systems (ARIS) by Scheer & Nüttgens (2000) and Archimate by Lankhorst (2009).

The process architecture frameworks and design approaches are used to create an understanding of the business components and their interactions. Some approaches include the business processes in these models, such as the Zachman framework (Zachman, 1987), the TOGAF (Dijkman et al., 2011) , Archimate (Lankhorst, 2009) and the Dynamic Architecture (DYA) (Dijkman et al., 2011).

However, these approaches try to link business process to business goals or business functions. This enterprise architecture focus on process architectures can be of help when creating process architectures, but the creation of enterprise architectures is inherently different than the creation of process architectures. The enterprise modelling languages can be used to graphically represent business processes and their relations, but are not primarily focused on the design of a business process architecture.

3.3 Conclusion

Business process management focusses on the management and improvement of operational processes. This is done by analysing business processes that are modelled in a specific notation, for example the Business Process Modelling Notation, and suggesting improvements based on required performance measures (van der Aalst, 2004, 2011; Weske, 2012). The business process management lifecycle consists of six steps, known as identification, discovery, analysis, redesign, implementation, and monitoring and controlling.

Process architectures are part of the business process management lifecycle, and are defined as an organized overview of business processes with relations and guidelines that determine how they must be organized. Several approaches for creating process architectures exist, known as goal based, action based, object based, reference model based, and function based. In a study that evaluated these approaches using process managers, the reference model approach appeared to be the easiest to use, most useful and most popular. This approach uses a reference model to create a new process architecture.

The first sub question of this research is "*What is the current state of the business process management field and how is it linked to process architectures?*". Business process management is actively used by organizations as well as by the scientific world. Business process management uses the business process management lifecycle and the accompanying business process modelling and notation. The field is continuously extended with new ways to manage and improve business processes and helps organizations in controlling their business. Process architectures are part of the business process management lifecycle (as discussed in section 3.1) and are defined as "an organized overview of business processes with relations and guidelines that determine how they must be organized" (Dijkman, Vanderfeesten, & Reijers, 2011). Several studies address the use of process architectures to create a consistent and integrated collection of process models, such as (Green & Ould, 2004, 2005; Joosten, 2000; Koliadis, 2008; Scheer & Nüttgens, 2000).

4. Process Mining

Process mining focusses on all activities that use event data to extract process information. The goal is to discover or generate a process model by analysing recorded events in an information system, as visualized in the upper part of Figure 11 (van der Aalst, 2011). This can only be done when assuming that each event refers to a well-defined step in a process, each event refers to a case, and each event has its own unique timestamp (van Dongen, de Medeiros, Verbeek, & Weijters, 2005). In addition, events can have an actor that is responsible for executing the process steps. Using these process models, it is possible to check conformance and if necessary recommend process redesigns. It is a way to analyse the complex processes and massive amounts of data that underlie these processes (van der Aalst, 2011). For organizations it is important to use event data in a meaningful way. Meaningful depends on the purpose, but some examples are the identification of bottlenecks, providing insights, or streamline processes (van der Aalst et al., 2012).

4.1 Types of Process mining

Process mining activities can be categorized in several types, known as discovery, conformance, and enhancement (de Medeiros, Guzzo, Greco, van der Aalst, & Weijters, 2008; Song, Günther, & van der Aalst, 2009; van der Aalst, 2011). Discovery techniques take an event log, analyse it, and produce a process model without using any additional information (van der Aalst, 2011). Conformance entails the use of an existing process model, and compare this model with an event log of the same process. This technique can be used to check if the model is in line with the real process, and the other way around.

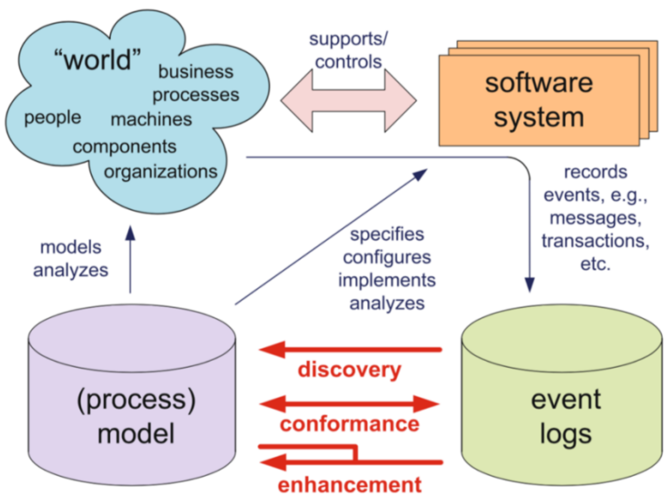


Figure 11: The three types of process mining (van der Aalst, 2011)

The last type is known as enhancement, which focusses on the extension or improvement of an existing process model by using actual process information from an event log. The first sub type of enhancement is repair, which means that the model is modified to better reflect reality (van der Aalst, 2011). The second sub type of enhancement is extension. By extending the model with new perspectives, it becomes possible to do more thorough analysis in the future.

4.2 Perspectives in Process mining

Next to the types of process mining, several perspectives are identified. These perspectives each have a different view on process mining and focus on different aspects (van Dongen et al., 2005). However, the different perspectives can overlap and are non-exhaustive. The three process mining perspectives are the process perspective, the organizational perspective, and the case perspective.

The process perspective, which is also known as the control-flow perspective, focusses on the ordering of activities (van der Aalst, 2011; van Dongen et al., 2005). When process mining with this perspective in mind, the main goal is to identify all possible paths. These paths are preferably shown in a notation, for example BPMN, UML Activity Diagrams, or Event-driven Process Chain (van der Aalst, 2011). The goal of the organizational perspective is to identify organizational resources and their relations. Using this perspective, social networks can be created and the organizational structure can be revealed (van Dongen et al., 2005). The perspective shows who is responsible for each part of the process. This can be important when bottlenecks are identified: if a process always hangs at a specific department, additional resources might be needed to improve the throughput time. At last, the case perspective is applied. Cases can be characterized by their flow through the process, but also by the values of corresponding data elements in the event log (van der Aalst, 2011; van der Aalst et al., 2007; van Dongen et al., 2005).

Van der Aalst (2011) identifies an additional (fourth) perspective that is not yet mentioned in literature by (van Dongen et al., 2005): the time perspective. This perspective focuses on all aspects of the process regarding timing and frequency of events. The availability of timestamps is a necessary data element for this perspective, as using timestamps makes it possible to identify time related bottlenecks or the remaining process execution time.

4.3 Example of process mining

To further clarify process mining an example of an event log is showed in Figure 12. This table shows a log containing 19 events, 5 activities, and 6 actors. Often, event logs contain more data elements than shown here (van Dongen et al., 2005). The example focusses on the discovery of a process model using an event log.

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	Register request	Pete	50	...
	35654424	31-12-2010:10.06	Examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	Check ticket	Mike	100	...
	35654426	06-01-2011:11.18	Decide	Sara	200	...
	35654427	07-01-2011:14.24	Reject request	Pete	200	...
2	35654483	30-12-2010:11.32	Register request	Mike	50	...
	35654485	30-12-2010:12.12	Check ticket	Mike	100	...
	35654487	30-12-2010:14.16	Examine casually	Pete	400	...
	35654488	05-01-2011:11.22	Decide	Sara	200	...
	35654489	08-01-2011:12.05	Pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	Register request	Pete	50	...
	35654522	30-12-2010:15.06	Examine casually	Mike	400	...
	35654524	30-12-2010:16.34	Check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	Decide	Sara	200	...
	35654526	06-01-2011:12.18	Reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	Examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	Check ticket	Pete	100	...
	35654531	09-01-2011:09.55	Decide	Sara	200	...
	35654533	15-01-2011:10.45	Pay compensation	Ellen	200	...

Figure 12: Example of an event log (van der Aalst, 2011)

Depending on the applied process mining technique, the type of process mining, and the perspective, only a selection of the information is used (van der Aalst, 2011). As mentioned in the introduction of this chapter, the minimal requirements for process mining are that any event can be related to both an activity (e.g. a process step) and a case, and that the events within a case are ordered. If each case consist of a sequence of activities it is referred to as a trace (van der Aalst, 2011). Examples of traces are shown in Figure 13, where a = register request, b = examine thoroughly, c = examine casually, d = check ticket, e = decide, f = reinitiate request, g = pay compensation, and h = reject request.

Case id	Trace
1	$\langle a, b, d, e, h\rangle$
2	$\langle a, d, c, e, g\rangle$
3	$\langle a, c, d, e, f, b, d, e, g\rangle$
4	$\langle a, d, b, e, h\rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h\rangle$
6	$\langle a, c, d, e, g\rangle$
...	...

Figure 13: Compact representation of an event log (van der Aalst, 2011)

Using process mining techniques it is possible to discover a business process based on the information in Figure 13. All the traces that are present in this table are also present in the process model (in Figure 14). When a trace fits the model, *it conforms* to the model (van der Aalst, 2011). This can be checked by replaying the traces, in other words, checking if the sequence of events that is present in the case are possible in the model.

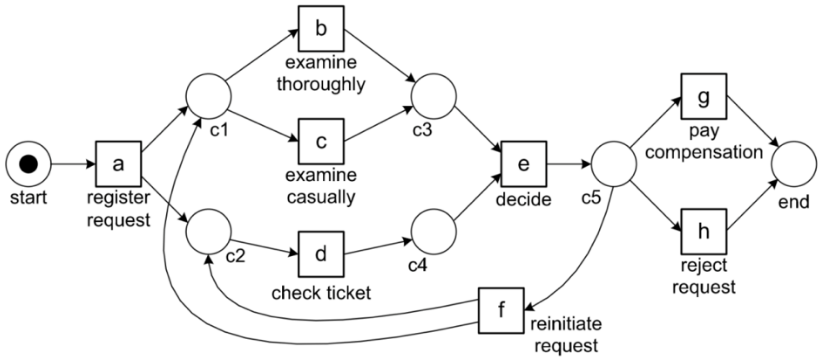


Figure 14: Process model based on the traces in Figure 13.

The discovered process model also allows for traces that are not present in Figure 13. This is the case because the process mining technique tries to generalize the behaviour that can be found in a log, which means that the process model allows for others combinations of activities than current present in the log. When doing this, a model can be overfitting, thus being too specific and not allowing for accidental behaviour. On the other hand, a model can be underfitting, which means that the model is too general and allows for behaviour that is not related to the behaviour observed. One should note that it is hard to balance this (van der Aalst, 2011).

Besides fitting, conformance consists of the following aspects: simplicity, precision, and generalization (Adriansyah, Munoz-Gama, Carmona, Van Dongen, & van der Aalst, 2013). In literature, the focus is mainly on fitting, as described above. Precision is about penalizing a process model for allowing behaviour that is not likely to occur given the observed behaviour in the event log. When a process allows for behaviour that is not related to the behaviour that is seen in the event log, the model is too general. A model that is too general is hard to work with and scores low on precision. Therefore Adriansyah et al. (2013) proposed an approach to pre-align the log with the model, making it possible to measure precision more accurately and ensuring that the model does not allow for too much unrelated behaviour.

4.4 Trace clustering

Traditional process mining techniques as described in section 4.1 are not able to deal well with unstructured processes or processes that can have numerous paths (Bose & Van Der Aalst, 2009). The process models that are mined using these techniques are often referred to as spaghetti-like models, as they are hard to comprehend.

A known approach to avoid this is to cluster certain traces in such a way that each of the clusters corresponds to a coherent set of traces. This improves the mined process model as well as the understandability of the model. Bose & Van Der Aalst (2009) propose an approach that is able to construct process models based on clustered traces. In most of the cases, there are sets of traces that represent expected or normal process behaviour and sets of traces that represent other paths. Figure 15 shows multiple process models based on the same event log. The spaghetti model contains all traces, the other models all represent one set of traces (a cluster).

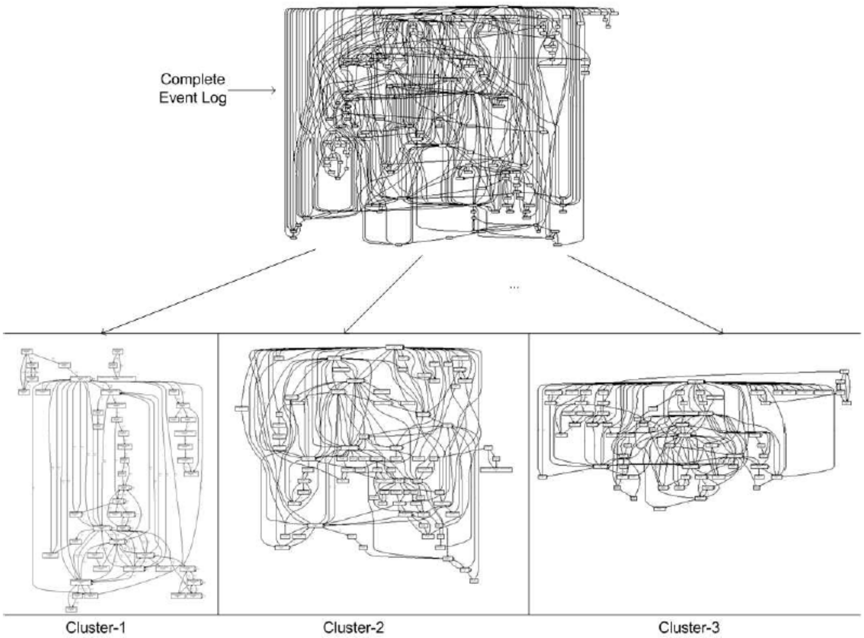


Figure 15: Cluster based process models (Bose & van der Aalst, 2010)

When executed correctly, Bose & Van Der Aalst (2009) state that "process mining based on clusters of good traces should generate models that have a high degree of fitness and a low degree of structural complexity". This means that the constructed process models are not overfitting (they are not too general) and not underfitting (they are not too specific)(van der Aalst, 2011).

Several clustering techniques exist, and Bose & Van Der Aalst (2009) used a technique called agglomerative clustering (also known as hierarchical clustering). This technique places each event trace into a different cluster, and then iteratively combines clusters that are closest to each other. Various criteria exist to define which clusters are closest, Bose & Van Der Aalst (2009) use the minimum variance criteria that tries to minimize the variance within a cluster. This leads to compact clusters, which eases the discovery of different process models. Bose & Van Der Aalst (2009) state that good clusters are clusters that consist of process models which show a high degree of fitness and are less complex than the original models. In other words; the constructed process models should be less spaghetti like. This is the goal of clustering: the ability to generate several simpler models each explaining a coherent group of process instances (Bose & van der Aalst, 2010).

4.5 Artefact centric mining

An additional way of process mining is the artefact centric approach, which focusses on data objects and the manipulations of those data objects in a process (Nooijen, van Dongen, & Fahland, 2013). These data objects are the key entities driving a company's operations (Cohn & Hull, 2009; Popova, Fahland, & Dumas, 2013). As they drive the operations, they define the lifecycle of the overall business process of an organization. The relations between artefacts are defined by the interactions between the artefacts during their lifecycles. An artefact-centric process model describes a process as multiple collaborative artefacts, each with its own life cycle and interactions.

As Canbaz, van der Aalst, Fahland, & Weijters (2011) described in their research "Discovering artifact-centric processes: mining assistance handling process of a first aid company using an artifact-centric approach", classical process mining techniques can not be used to discover what is happening in a process with multiple objects that have complex relationships. Artifact centric mining is used to discover a process by using the artifacts that are present in the process and is therefore often used to create better process models for real life or physical processes (Fahland, De Leoni, Van Dongen, & Van Der Aalst, 2011). Figure 16 shows how traditional process mining is related to artefact centric mining. It visualizes how case instances in process mining are identified as artefact instances in artefact centric mining (Lu, 2013). This is important, as the lifecycle of the artefact instances and their interactions are the basis for an artefact centric process model.

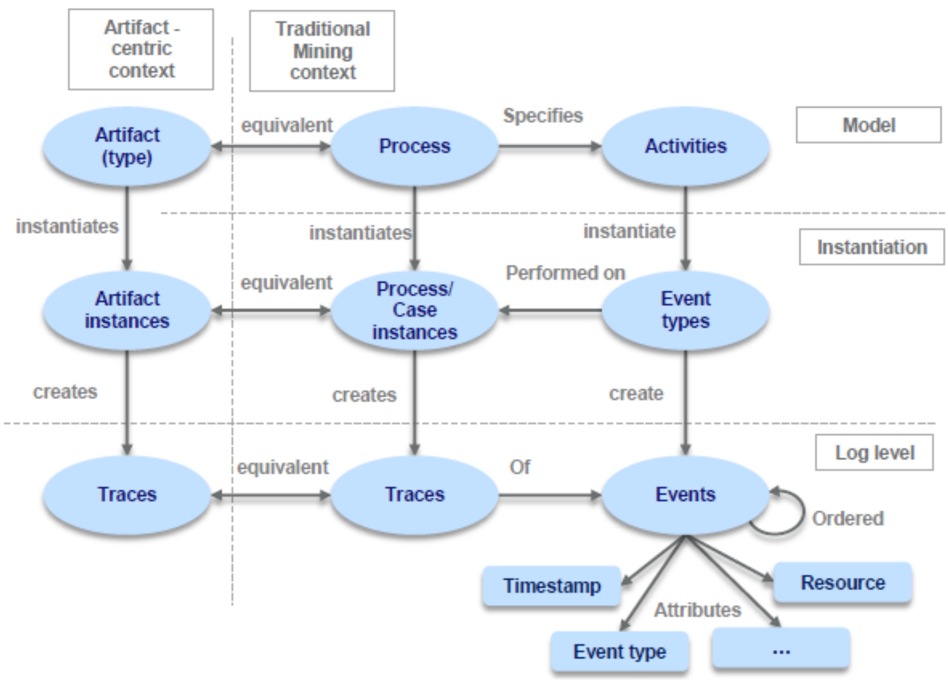


Figure 16: Traditional process mining versus artefact-centric context (Lu, 2013)

Nooijen et al. (2013) present a technique for discovering artefact centric processes from structured data sources. An example of such a structured data source is a relational database. To correctly mine these processes, it is necessary to automatically identify the case identifiers in the raw log of the data source, as they represent an artefact. For each artefact an event log is than extracted from the data source, which can be used to discover artefact life cycle models (Nooijen et al., 2013). The technical details regarding algorithms and artefact schema extraction are outside the scope of this thesis, the interested reader is referred to Nooijen et al. (2013).

Popova et al. (2013) extend this approach with the Guard-Stage-Milestone (GSM) meta-model. GSM is a notation that can visualize the behaviour that is present in an artefact event log. The complete approach can be seen in Figure 17. The approach starts with a raw artefact log. Such a log consists of all events with a timestamp, and additional data attributes. To determine the instances of artefacts, an identifier is chosen to structure the log. Using these identifiers it is possible to decompose the raw log in artefact centric logs that consist of one artefact and its trace. These artefact centric logs can be used to discover the artefact centric lifecycles, visualized as Petri Nets using standard process mining methods. The last step of the Guard-Stage-Milestone approach is to convert the petri net model into the GSM notation.

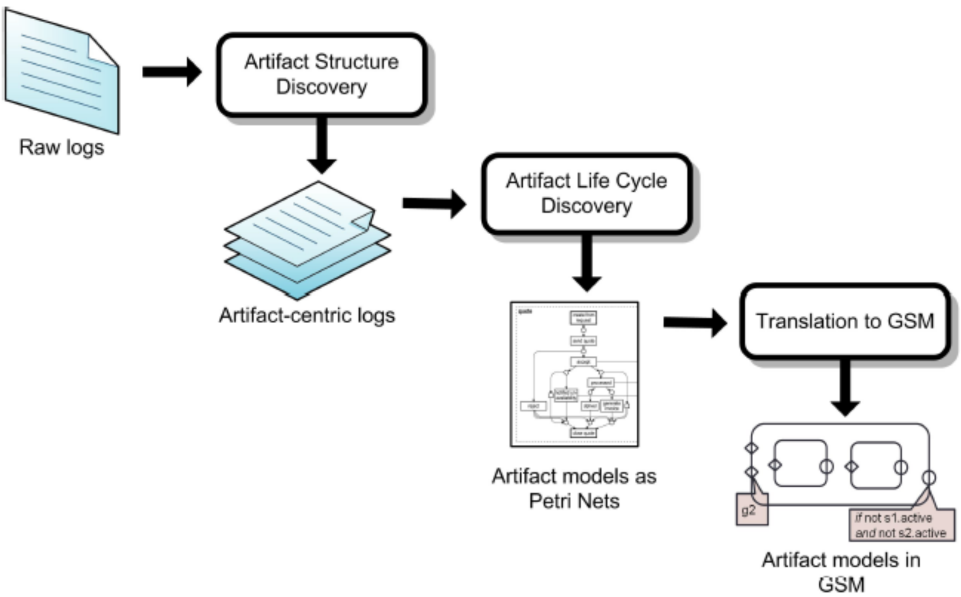


Figure 17: The artefact lifecycle discovery process (Popova et al., 2013)

4.6 Configurable process models

A configurable process model is a combination of a set of process variants, based on a collection of event logs (Buijs, van Dongen, & van der Aalst, 2013). Such a model is able to show commonalities and differences among the different variants. A user of a configurable process model has the possibility to configure the model by making configuration choices: this is known as an process instantiation (Schunselaar, Verbeek, van der Aalst, & Reijers, 2012).

There is need for these models, because different units within an organization may need to execute similar business processes. These organizations can use configurable process models to share development efforts, analyse differences, and learn best practices across organizations (Buijs et al., 2013; Schunselaar et al., 2012).

Buijs et al. (2013) proposed four approaches to create configurable process models from collections of event logs: (1) merge individually discovered process models, (2) merge similar discovered process models, (3) discover a single process models then discover configurations, (4) discover multiple process model configurations at the same time. From these approaches, the fourth approach where both the control flow and the configuration options are discovered together seems to be the most flexible (Buijs et al., 2013).

4.7 Process mining methodologies

The execution of process mining techniques at companies is not yet very common. To assist companies in doing so, three process mining methodologies exist (van Eck, Lu, Leemans, & van der Aalst, 2015). The first one is known as the Process Diagnostics Method (PDM), and focusses mainly on providing a broad overview of a process (Bozkaya, Gabriels, & Werf, 2009). The L* life cycle model is somewhat more in depth, and considers aspects as process improvement and operational support (van der Aalst, 2011).

However, both methodologies are not designed for iterative process analysis, which is often necessary in process analysis situations (van Eck et al., 2015). The third methodology is known as PM^2: a process mining project methodology, and is more extensive than the previously mentioned methodologies. It is constructed to analyse structured as well as unstructured processes using several process mining techniques and applications. The idea behind the methodology is to guide companies in performing process mining projects.

Figure 18 shows the PM^2 methodology and its stages. The methodology starts with Planning, a stage which is used to set up the project and to determine what to research. The goal for starting a process mining project is performance improvement or compliance checking. The second stage, Extraction, focusses on the extraction of event logs and (optionally) process models. The event logs needed are identified based on the research questions formulated in the planning stage. The data processing stage is necessary to optimize the extracted data for analysis. Four types of activities are present in this stage to improve the data quality: creating views (focus on a specific part of the data), aggregating events (reduce complexity), enriching logs (computing additional event data), and filtering logs (remove unnecessary attributes).

The fourth stage is known as mining and analysis, where process mining techniques are used to answer the research questions formulated in stage one. Using these techniques it is possible to gain insight in process performance. The activities in this stage consist of the process mining types mentioned before in section 4.1: process discovery, conformance checking, and enhancement. In addition, process analytics is possible, which consists of data mining and visual analytics (de Leoni, van der Aalst, & Dees, 2014).

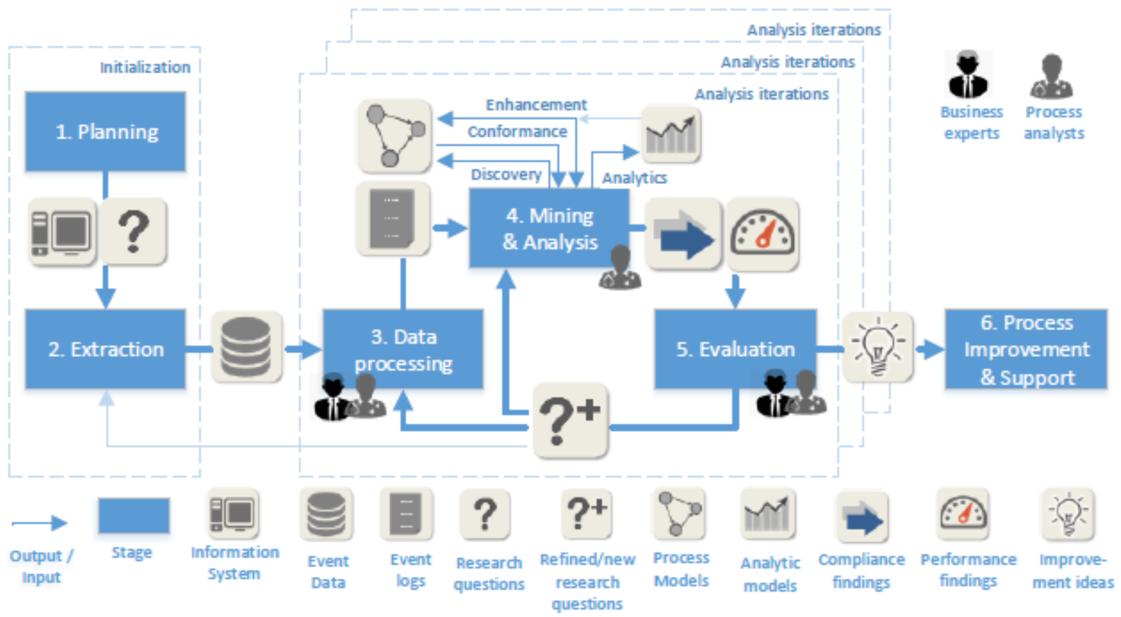


Figure 18: The PM² process mining methodology (van Eck et al., 2015)

The goal of the evaluation stage is to use the results of the analysis stage to think of ways to improve the process. This can be done by distinguishing interesting or unusual results, understanding the process model that is discovered by the software, and verifying and validating the findings. Discussing the results of the analysis should create new insights in the process (van Eck et al., 2015). These new insights are used in the process improvement & support stage, where the actual process modifications are determined (as described in section 5.2.1 Process improvement). However, the implementation of process changes often is not part of the process analysis project, but of subsequent projects.

4.8 Conclusion

To answer to the second sub question of this research, "*How can process mining be used to create process models that focus on a specific part of the organizational business process?*", it can be concluded that Business process mining is a technique that can be used to generate process models based on event logs. An event log is a list of events, recorded by an information system. Each event should refer to a specific case to enable process mining.

Multiple types of process mining exist, for various purposes. Discovery purely focusses on the discovery of a process model; conformance checks a discovered process model with an existing process model; and enhancement focusses on the improvement and extension of an existing process model using an actual discovered process model. There are several perspectives on process mining which are defined as the organizational perspective, the process perspective, and the case perspective. Each perspective has its own view on process mining and focusses on different aspects. An additional way of process mining is the artefact-centric approach, which focusses entirely on data objects and the manipulations of the data objects. An artefact-centric process model describes a process as multiple collaborative artefacts, each with its own life cycle and interactions.

Using process mining, configurable process models that include a set of process variants can be constructed. Such a model is able to show commonalities and differences among the different variants, and can be used to execute similar business process instances in an organization. This is related to trace clustering, which can be used if mined process models are "spaghetti-like": the models have numerous paths and are hard to comprehend. Using trace clustering it is possible to identify clusters in these spaghetti like models, and these clusters can be identified as process variants.

The PM^2: a process mining project methodology can be used to analysis structured as well as unstructured processes using several process mining techniques and applications. The idea behind the methodology is to guide companies in performing process mining projects.

5. Positioning and related literature

This chapter positions the described literature regarding business process management and process mining in their context. Many related concepts, methods, and techniques exist that might play a role further on in this study. Besides the positioning, related fields as general process improvement, functional architectures, and social network analysis are discussed here as well.

5.1 Positioning

In this section the disciplines of business process management and process mining are positioned within relevant fields. In addition, related literature and concepts are described. In 2012, the process mining manifesto was released by van der Aalst et al. (2012). The manifesto is created by the IEEE Task force on Process mining, and it aims to promote the topic of process mining. It contains a set of guiding principles regarding process mining and it lists important challenges. The manifesto serves as a guide for scientists, consultants, software developers, and end-users (van der Aalst et al., 2012). According to the manifesto, process mining is a young research discipline, where additional research and education is required. The idea of process mining is situated between the fields of data mining and process modelling & analysis, which are closely related to Business Intelligence (BI). This can be seen in Figure 19.

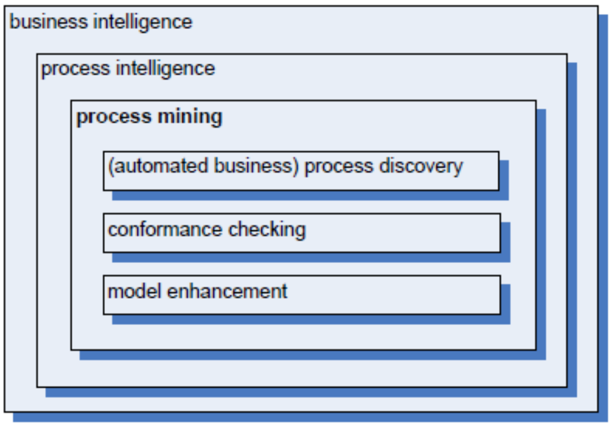


Figure 19: Positioning of Process mining (van der Aalst et al., 2012)

BI consists of multiple concepts, such as Business Activity Monitoring, Complex Event Processing, Corporate Performance Management, Continuous Process Improvement, Business Process Improvement, Total Quality Management, Six Sigma, and Lean. Process mining is a technique that can support any of these concepts as it is able to give new insights in an organizations processes. Of course, this also relates to the fields of Business Process Management and process redesign. In this thesis the focus is on the application of process mining using process mining methodologies. Process mining is also related to other fields outside the BI scope. An example of such an application is described in section 5.3. Section 5.2 gives a general overview on process improvement.

5.2 Process improvement

As mentioned in the Business Process Management section (3.1), process redesign is an important aspect of process management and focusses on the adaptation of an existing process. Multiple methods and approaches for redesign exist, which can differ on level of abstraction: methods, techniques, and tools (Reijers, Dumas, van der Aalst, & ter Hofstede, 2005). Methods can be defined as a collection of problem solving methods, consisting of a set of principles (Reijers et al., 2005). These methods completely cover a redesign project from the identification and diagnosis until the implementation. Methods generally can be used in two ways: to adapt an existing process or to start over (the clean sheet approach). More specific process improvement methods using process mining are already described in section 4.7 Process mining methodologies. In this section general process improvement techniques and methodologies are discussed. These techniques are a set of procedures that are aimed to achieve a task and are often incorporated in a method. An example of such a technique is diagramming, a technique that can be used for process diagnosis. A technique is often supported by a tool: a software program that is able to assist with a certain technique. Many tools exist that can evaluate process models – less are available to assist the redesign process.

5.2.1 Process improvement techniques

Reijers et al. (2005) describes multiple techniques to redesign processes. It is recommended that the process is already modelled before these techniques can be applied. This makes it easier to apply the methods and create insight in the processes. The approaches that can be used for process redesign are listed below, with a short explanation:

- Task Elimination
    - The idea of this approach is to eliminate tasks in the process that do not add value. This can increase the speed of processing, also known as throughput.
- Task Composition
    - Composition focusses on merging small tasks into one larger task, or decomposing large tasks into smaller ones. This can have a positive effect on the quality of the executed tasks.
- Task Automation
    - The purpose of automation is to improve execution times with less costs, as no human resources are required to execute the task.
- Resequencing
    - Resequencing can be used to move tasks in a process to another location in the process, thus improving the process flow.
- Parallelism
    - This practice can be used to see if some tasks could be executed in parallel, decreasing the process time.
- Numerical involvement
    - Numerical Involvement focusses on the minimization of the number of units involved in a business process. These units can be persons, departments, companies, etcetera. Reducing this number could lead to fewer coordination issues and less shared responsibilities.

- Integration
    - This approach considers the integration of the business process with a business process of a client or a supplier. In the end this approach could lead to synergies, where the result of the integrated business process is larger than the sum of the parts (the individual business processes).
- Outsourcing
    - Outsourcing is a well known practice to reduce costs of a certain business process. A business process can be outsourced in whole or in part, but the overall quality of the process execution may decrease.
- Interfacing
    - Interfacing can be used when it seems useful to use a standardized interface with clients and partners. An interface can reduce the probability of mistakes and unclear communication. Interfacing can be seen as an implementation of the integration approach mentioned above.
- Case types
    - The case types practice can be used to distinguish new business processes. This can be necessary when the current superflow (the umbrella business process) is not capable of correctly handling a specific process execution.

This shows that many approaches exist to modify an existing business process. These approaches can be applied independently or combined, with their purpose being to reduce throughput time and/or the costs of the process (Reijers et al., 2005). However, it should be noted that nowadays electronic processes are often complex and have multiple dependencies, so redesigning these processes is a time consuming activity.

5.2.2 Process improvement methodologies

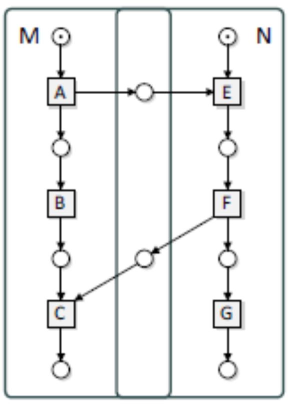
Besides these techniques, several process improvement methodologies exist which do not specifically focus on process mining. A well known method is Six Sigma, a method based on certain techniques that tries to optimize business processes (Pyzdek, 2003). The sigma in the name refers to the Greek letter that is used by statisticians to measure the variability in business processes. The idea behind six sigma is that the variation that occurs in a business process lies within six standard deviations (sigma's) of the normal process execution. The methodology was originally created by Motorola to improve their process execution and consists of quality management methods, including statistical methods. These methods in combination with experts (several expert levels exist, referred to as Black belts, Green belts, etc.) that can apply them creates a special infrastructure within the organization focussed on process improvement (Pyzdek, 2003). Each six sigma project has several pre-defined steps that are executed to further improve a process. Often, these projects have related financial targets to reduce costs/increase revenue.

Another well known improvement method is Lean Manufacturing (Naylor, Naim, & Berry, 1999). The focus of lean operations is to create maximum value for customers with the least amount of wasted resources. Reducing the amount of wasted resources lowers the operational costs, which in the end improves the revenue. The lean method consists of several techniques to eliminate unnecessary activities and to create a production flow (Krafcik, 1988). These techniques are applied when one of the following three problems is identified: Muri, Muda, and Mura. Muri focusses on overload – moments where the demand is higher than machine or men can work. Muda is keeping in mind that every activity that is executed in the process should add value for the customer. When this is not the case, the activity is identified as Muda. At last, Mura tries to identify inconsistencies in the process. An example of an inconsistency is product demand and the accompanying process effects throughout the year.

5.3 Functional architectures from events logs

Recent research of van der Werf & Kaats (2015) focusses on the use of process discovery techniques to construct functional architectures. This is partly based on work of van der Werf & Verbeek (2015), in which execution data was used to map events to features. By doing this it is possible to construct a software architecture. Van der Werf & Kaats (2015) tried to use process mining techniques to discover the functional architecture from an event log. They determined if it was possible to derive which features of a system interact, based on an event log. In addition, research was done to discover if process discovery techniques can be used to identify the order of execution of system features. Using an event trace, it is possible to create a petri net diagram of the possible interaction. The communication between the modules is a-synchronous, so messages are sent between the features to complete their functionality. It is important to consider the order of communication when creating the petri net.

Using an event log of system execution data, it is possible to create a Petri net for the modules M and N. The petri net shows the interaction between the two software modules, and which features are responsible for the interaction. Figure 20 shows the event log on the right, and the corresponding petri net on the left.



Case	Trace
1	A, E, F, G, B, C
2	A, E, F, B, G, C
3	A, E, B, F, G, C
4	A, E, B, F, C, G
5	A, E, F, B, C, G
6	A, B, E, F, G, C
7	A, B, E, F, C, G

Figure 20: Petri net of modules M and N, based on event log (van der Werf & Kaats, 2015)

The functional architecture model (FAM) of Brinkkemper & Pachidi (2010) is used to model the overview of a system. This functional architecture represents at a high level the major functions of a software product, and the interactions between these functions. Using the petri net and the event log it is possible to create such a functional architecture, as can be seen in Figure 21.

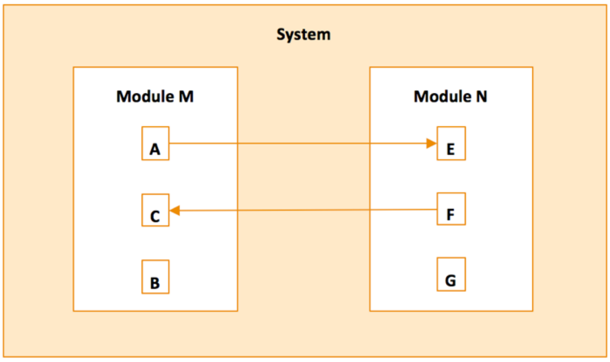


Figure 21: Functional Architecture Model based on an event log

The approach presented by van der Werf & Kaats (2015) provides useful insights for software architects and software architecture (re)construction. Furthermore, there are possibilities to adapt this approach for the discovery of business processes based on event logs. The Business Process Modelling and Notation can then be used to model the interaction between these processes. The use of process mining to create functional architectures is related to the previously mentioned artefact-centric approach. Both approaches focus on the interactions between within execution data. Additional research seems necessary to further study this relation.

5.4. Social network analysis

Process analysis and process diagrams are related to social network analysis and social network diagrams, since both give insight in relations and interactions that exist between nodes. Social network analysis (SNA) can give new, more socially focussed insights in process structures and relations. According to (Serrat, 2009) social networks are nodes of individuals, groups, organizations, and related systems with some kind of relationship. These relationships can for example be of social contact, financial exchange, shared ideas, or any other possible relation. The focus of social network analysis is on the structure of these relationships and assumes that these relationships are important. It tries to map the actors in a social network and indicates their relations (Serrat, 2009). Otte & Rousseau (2002) describe social network analysis as a strategy for investigating social structures. They state that traditional individualistic social theory analysis does not take into account the behaviour of multiple actors, thus ignoring the social context of an actor. Social network analysis does focus on these inter-actor relations by focussing on the relational data.

A social network consists of nodes (N) and a set of relationships or links (L). Since social network analysis often focusses on relations between humans, the nodes are seen as people. However, it is possible to create social networks where the nodes are different from humans. An example is a social network showing the interactions of several organizations that are spread throughout a country.

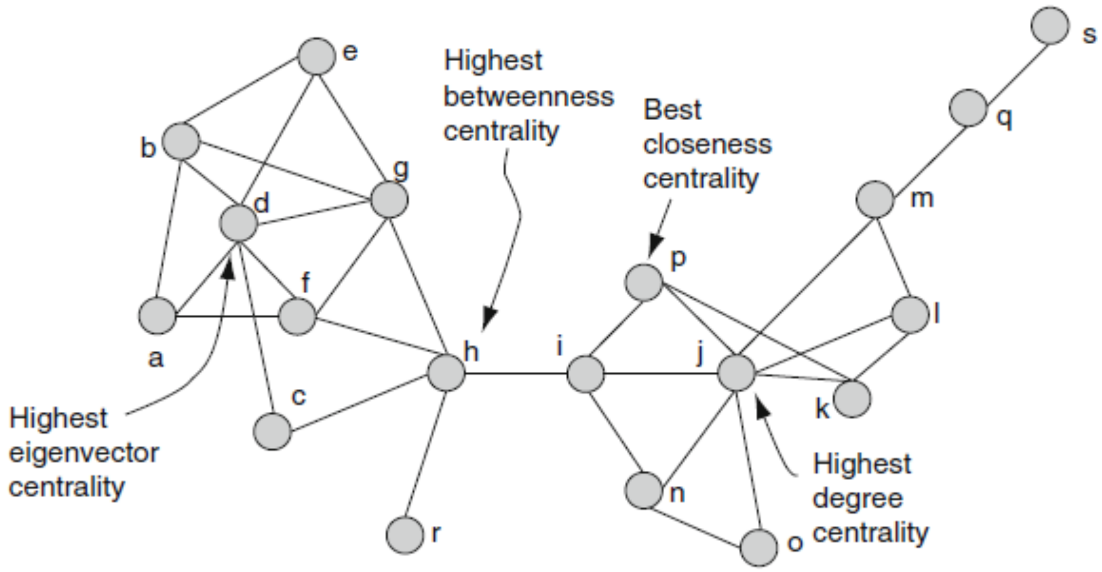


Figure 22: Diverse centrality measures (Manuel, 2013)

To describe the structure of a social network, some centrality measures are available (see Figure 22). These measures are known as degree centrality, closeness centrality, and betweenness centrality. Degree centrality focuses of the number of ties a node has compared to other nodes. The higher the number, the more central the node is in het network. Closeness centrality is measured by the distance from a certain node to all other nodes. The lower the total distance is, the more central the node is. The betweenness centrality measure is defined as the number of paths that pass through a node. The higher the number of paths that go through a certain node, the more central it is. Less used is the eigenvector centrality, which calculates the influence of a node on the network. Higher scores mean a node can influence (touch) many other important nodes.

Besides visualizing social relations, social networks can also be used to visualize any other kind of network structure. An example of this is the visualization of supply chain networks. These networks are often thought of as some kind of flow from input to output. However, an organization can have multiple suppliers and buyers. This is hard to visualize using just a simple supply chain flow – such a visualization fails to capture the complexity of the network (Kim, Choi, Yan, & Dooley, 2011). Many relations can be identified using social network analysis, such as relations between multiple suppliers and buyers. SNA therefore can be used for identifying and studying patterns within an organizations network and enables organizations to use these patterns to create competitive advantages (Kim et al., 2011).

5.5 Conclusion

Process mining is a relatively new research discipline with many developments in the last few years. It can be situated between data mining and process mining, and is seen as a Business Intelligence (BI) technology. BI encompasses other concepts such as Business Activity Monitoring, Complex Event Processing, Corporate Performance Management, Continuous Process Improvement, Business Process Improvement, Total Quality Management, Six Sigma, and Lean.

A related concept of process mining is one of the latest developments where event logs are used to create functional architecture models. Besides software architectures, this development can also be applied in physical processes. This enables the user to visualize functional architectures of real life processes. The same yields for social network analysis: although it is often used to visualize networks of social actors in a certain context, it is also possible to show the relations and interactions of different kinds of nodes in different kinds of networks. Combining these techniques should create a new kind of interactive visualization that is able to present process information in innovative ways.

To answer sub question 3; "*How can business process management and process mining techniques be combined to improve business processes on organizational and departmental level?*", it can be stated that it is necessary to confirm the existing business process models using process mining. By combining the analytical elements and the business process management principles it is possible to understand process hierarchies in complex processes. This is necessary to create a process architecture visualization that gives innovative insights in the process execution, which can be used for improvements.

6. Case organization and data

In this section the case organization is described. The organization studied in this research is PostNL. PostNL is the largest mail and parcel distributor in the Netherlands. In 2014, PostNL addressed 2,705 million mail items and 142 million parcels (see Figure 23). The total revenue was 4,251 million euros, of which the parcel distribution generated 854 million (PostNL, 2014). PostNL is also active in Germany, Italy, the United Kingdom, and since 2015 in Belgium. In each country they operate, PostNL wants to provide a broad range of solutions for their customers that are active in e-commerce. Customers are provided with flexible delivery options for parcels, for example 'before 10:00' and 'before noon'. The aim for the coming years is to profitably grow the parcel business, and sustaining the mail business. In previous years, it can be seen that the parcel business has grown significantly (see Figure 23). Besides growing the parcel business PostNL aims to offer excellent service and customer focus.

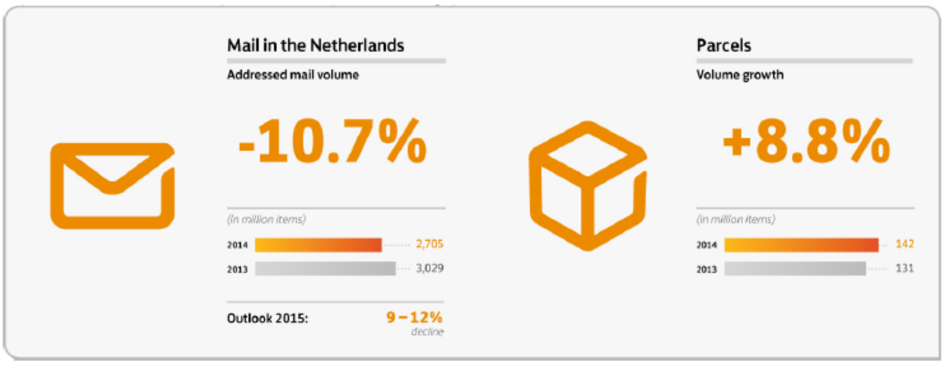


Figure 23: Growth rates of Mail and Parcel divisions of PostNL

6.1 Processes

As mentioned earlier, PostNL has 18 distribution centres throughout the Netherlands. Each of these centres contains a sorter, which is a large conveyor belt with multiple barcode scanners. On the long side of each sorter, there is the possibility to load parcels on small conveyors. These conveyors lead up to the sorter. Using the conveyors and the barcode scanners, the parcels are sorted.

In the distribution centre two main processes are executed: an evening sorting and a morning sorting. In the evening, the centre sorts all parcels that are received during the day in shifts. Each shift contains multiple postal codes and thus covers a specific area. In most of the cases, the evening sorting results in 8 shifts. This depends however on the size of the area the sorting centre has to cover. The larger the area, more parcels need to be distributed, so more shifts might be necessary for some centres. In the morning, each shift is sorted for multiple delivery mans. When the parcels are sorted over the delivery mans, they will try to deliver the parcels to the clients. This third process is the responsibility of the delivery men, not of the distribution centre.

6.1.1 Cross docks

Three of the 18 distribution centres are so called cross docks. They do not only distribute parcels to clients, but are hubs in PostNLs network between distribution centres. During the evening sorting process in these cross docks, parcels are not only sorted in shifts but also to other regions in the Netherlands. The three cross dock centres (CD) form the centre of PostNLs distribution network. Each of these centres is responsible for around one third of the Netherlands. Figure 24 shows a network graph of sorting centres throughout the Netherlands.

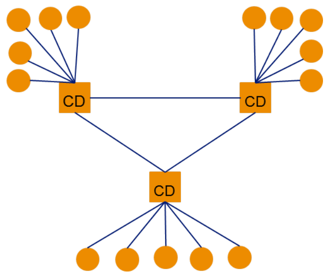


Figure 24: Network graph of sorting centres

6.1.2 Process instances

As mentioned earlier, the 18 centres each are responsible for their own distribution process. This means that there are 18 instances of the same process, that each generate unique data. This makes it possible to research, analyse, and compare all these instances. During these distribution processes, the parcels are scanned multiple times. When the parcels are scanned and what such a scan means is explained in the following section. The scan data is used in the case study to compare the distribution centres and to create organizational and departmental process models.

6.2 Data generation

Each parcel gets assigned different values through multiple scans and each value has its own meaning and possible consequences for the further distribution of this parcel. In this section a description of possible values that can be assigned during the process is given. Additionally, it is shown how the process log of one day of one distribution centre looks like. Figure 25 below shows the different processes as part of the overall organizational distribution process. Each process creates its own scans.

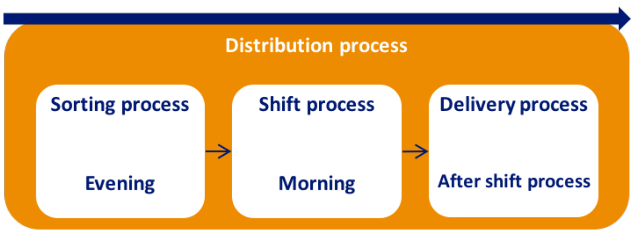


Figure 25: Overview of processes of one sorting centre

6.2.1 Data generation during evening sorting process

During the day a distribution centre receives multiple containers with parcels. The number of parcels depends on the location of the centre – a rough estimate is between 30,000 and 40,000 parcels per day. From 18:00 o'clock onwards, all the parcels go on the conveyer. The parcels receive there first scan, which assigns the parcels a B1 value. B1 stands for 'accepted in sorting centre'. The parcels are now divided in the shifts, covering the areas where they need to be delivered. When parcel 1 needs to be delivered in area A, the conveyer shoves the parcel into the sub-sorter for area A. This is done with all the 40,000 parcels that are on the sorter in the evening. In the end all the parcels are divided over the correct areas.

6.2.2 Data generation during morning shift process

In the morning, the shifts are further divided. Starting with shift 1 and ending with 8. First, all parcels for shift one are placed on the conveyer. All the parcels are scanned again, and given the value J1. When a parcel is at the right conveyer, the sorter shoves it down. The delivery man for that particular area can then take the parcel and place it in his van. When all the parcels for shift 1 are sorted, all delivery men are notified. Then they can go deliver their parcels for their specific region. In the centre, they will now start with sorting shift 2. This will be repeated until all the shifts are sorted.

6.2.3 Data generation during delivery process

Each delivery man has a small computer as a hand scanner. With this terminal he can scan a parcel when delivering the parcel. The delivery man can choose from multiple options when he has scanned the parcel. If someone is home, the delivery man for example selects "delivered". The hand scanner then assigns the I1 value to the scan. If the delivery man delivers the parcels with the neighbours, the I10 value is assigned to the scan. If the delivery man delivers the parcel at the nearest retailer, for example a supermarket or post office, scan J11 is assigned. When the delivery man is going to try to deliver the parcel tomorrow, he assigns scan J8. When a delivery man returns to the sorting centre, all the parcels that could not be delivered are scanned. These parcels will be added to the evening sorting process, such that they can be delivered the next day.

6.3 Scantrails

As stated before, numerous values can be assigned to a parcel when it is scanned. All these values combined for one parcel is known as a scantrail. A value always consists of a character and a number. These scantrails know several happy flows: flows where nothing went wrong and the parcel was delivered. Often such a flow consists of the scans as shown in Table 3.

Table 3: Scan values

Value	Definition	Location
A1	Pre-notification of shipment	System
B1	Shipment is accepted and in sorting process	Distribution centre
A4	Billing notification	System
J1	Shipment is sorted	Distribution centre
J40	Shipment is loaded onto truck	Distribution centre
J5	Shipment is out for delivery	Distribution centre
I1	Shipment is delivered	Delivery men

There are numerous combinations of values in a scantrail, possibly with duplicates or wrong scans. Table 4 below shows an selection of these values and their description. The entire list of possible scans consists of more then 150 scans.

Table 4: Most occurring scan values

Value	Definition
A1	Shipment is reported. Systems are notified of the arrival of an incoming parcel from a retailer.
A16	Client matched as receiver of parcel. Systems of PostNL linked a parcel to a known client.
A4	Pre-notification generated by PostNL. Check that the parcel still is in the process.
B1	Shipment is accepted and in sorting process.
I1	Parcel is delivered to the client
I10	Parcel is delivered at neighbours of client
J1	Parcel is sorted
J10	Parcel received at sorting centre
J11	Parcel is delivered to retailer
J30	Parcel collected by 'planbalie'
J4/J40	Parcel is loaded onto truck
J5	Shipment is out for delivery. Parcel is on route to the client
J8	Second delivery. Parcel could not be delivered, will be tried again tomorrow

6.3.1 Example export of process log

An export of a process log is shown in Table 5. This log is exported from Track and Trace, a software system that contains most of the data regarding the parcels. This export only shows one trace, for the parcel with identifier 3SBOL1671337. In addition, only some elements are shown in the log: the date, time, the value (consisting of a character and a number), and the description of the value. As can be seen in this log, the scantrails follows the happy flow as described above.

Table 5: Process log

Barcode	Date	Time	Value	Reason	Description
3SBOL1671337	9-12-2015	18:08:07	A	01	Zending is voorgemeld
3SBOL1671337	10-12-2015	0:49:25	B	01	Geaccepteerd
3SBOL1671337	10-12-2015	8:00:00	A	04	VERZENDTARIEF
3SBOL1671337	10-12-2015	9:34:25	J	01	Zending gesorteerd
3SBOL1671337	10-12-2015	9:34:26	J	40	Voorgemeld en gesorteerd op rit
3SBOL1671337	10-12-2015	10:06:03	J	05	Voorgemeld en out for delivery
3SBOL1671337	10-12-2015	12:28:45	I	01	Afgeleverd expl. vastgelegd

In the case study large amounts of these scantrails are analysed. The case study approach is described in the following chapter.

6.4 Conclusion

PostNL is the largest parcel distributor in the Netherlands. During the distribution of parcels, an event log is created, which enables PostNL to trace all their unique parcels throughout the process. Each parcel has a unique identifier, known as a barcode. Each parcels is scanned multiple times and this creates the event log. Useful attributes in the event logs are Barcode, date, time, value, and location code. Some scans that are present in the event log do not influence the process: these scans need to be removed before the analysis. Chapter 7 will elaborate on the methodology for the execution of the case study.

7. Process mining methodology and proposed approach

The case study is executed according to the PM² process mining methodology as described in literature section 3.3.2. The methodology is visualized in Figure 26. Each of the six steps is implemented in this study, if necessary, additional steps are defined since this study tries to extend current process mining possibilities. In the following subsections the execution of each step of the case study is described.

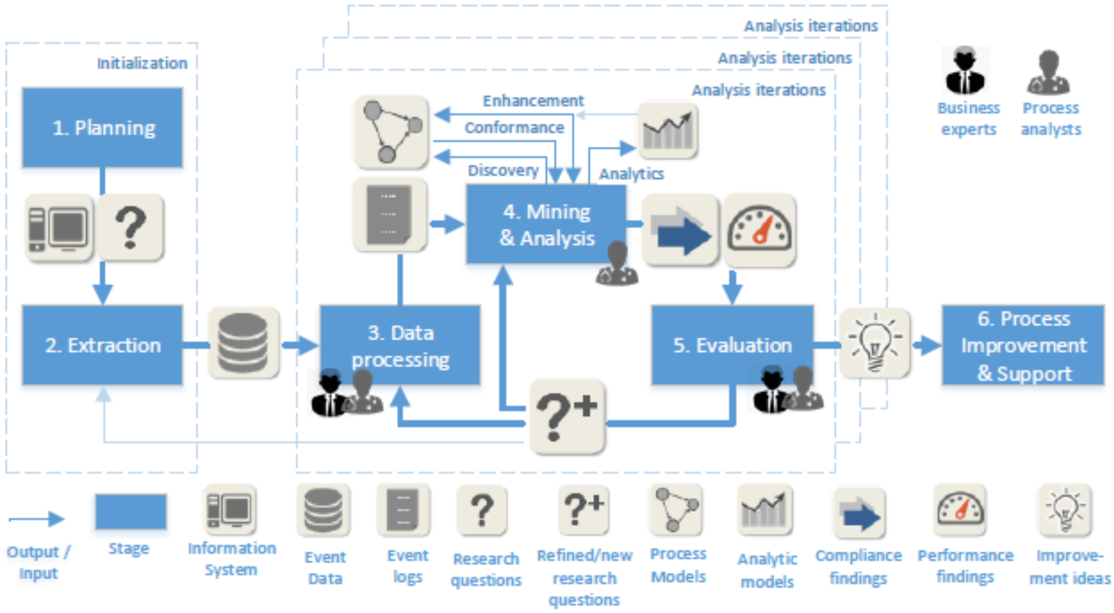


Figure 26: The PM² process mining methodology

7.1 Approach

7.1.1 Planning

The data required for this research consists of 19 datasets containing process information. For each sorting centre, one dataset is required. Additionally, one dataset is needed that covers the Netherlands as a whole. It is important that the selected datasets are covering the same dates, avoiding that certain datasets cover certain exceptions that changed the process. Examples of exceptions are failures in the systems, bad weather conditions, or strikes. The case study is structured and described in the following chapters:

- Preparatory (Chapter 7)
    - Data extraction
    - Data processing
- Execution (Chapter 8)
    - Mining
    - Evaluation
- Improvement (Chapter 9)
    - Process improvement using process architectures

7.1.2 Data extraction

The necessary data can be gathered using the parcel tracking software Track & Trace. This software contains all data regarding parcels in PostNLs distribution network. The data is originating from software systems in all the distribution centres and is combined by an external party. The data in Track & Trace contains many attributes, but only a few are necessary for this study. To avoid unnecessary data in the datasets the software enables the user to create a certain *view* so only the required data attributes are shown. The data attributes that are used in this study are depicted in Table 6. This is a selection of six attributes from an available total of 61. Using these attributes it is possible to extract complete event logs for process mining.

Table 6: Data attributes

Attribute	Description
Barcode	Unique ID
Waarnemingdatum	Date
Tijd	Time
Soort waarneming	Value category
Reden waarneming	Value identifier
Lokcode	Location identifier

The data that is selected covers the whole month February in 2016. For each distribution centre, 29 daily datasets were extracted. Such a daily dataset was constructed by retrieving the parcels that have a B1 value (proof of acceptance scan) on a specific day. When combined with a certain date as limitation, it is ensured that a scantrail started on this date. However, due to the large numbers of parcels in the distribution process it was not possible to extract the entire scantrail of all these parcels on a given day per distribution centre. This would lead to a dataset of approximately 30.000 parcels x average of 10 rows per parcel x 29 days = 8.700.000 rows per distribution centre. Combined into one dataset that covers the Netherlands this would lead to 156.600.000 rows. Therefore a restriction of 500 unique parcels is set to the number of parcels extracted per day. The 29 separate datasets were combined into one dataset for a distribution centre. This process is repeated 18 times. The total number of rows per distribution centre is indicated in Table 7. At last, the 18 distribution centres are combined into a 19th overall dataset. Since Microsoft Excel is not able to handle more than 1.000.000 rows, the datasets were managed using a database construction. This also simplifies the extraction of necessary subsets if necessary.

Table 7: Details of dataset

Distribution centre	Total number of rows in dataset
Amersfoort	111.001
Born	84.309
Breda	85.935
Den Bosch	103.887
Den Hoorn	72.345

Elst	90.116
Goes	81.362
Halfweg	124.722
Hengelo	66.348
Kolham	79.234
Leeuwarden	74.715
Opmeer	83.476
Ridderkerk	85.049
Sassenheim	87.188
Son	80.159
Utrecht	77.148
Waddinxveen	88.714
Zwolle	79.784
Combined dataset	1.555.492

7.1.3 Data processing and preparation of analysis tools

Each dataset has to be prepared before it can be analysed. As mentioned, an event log extracted from Track & Trace consists of six attributes. For correct analysis however it is necessary that the Date and Time attributes are combined to form a *timestamp* since this is a prerequisite for process mining (see section 4.1 Types of Process mining). The two attributes that both contain a part of the value, 'soort waarneming' and 'reden waarneming', are combined as well. By combining these values it is possible to analyse the process on different levels of detail. Analysing only the 'soort waarneming' is more high level than analysing the 'combinedscan', since the 'reden waarneming' divides the 'soort waarneming' in multiple parts. Therefore, the original attributes are not deleted, since it might be necessary to analyse only one part of the scan. The final structure of the dataset is shown in Table 8.

Table 8: Final structure of dataset

Attribute	Description (example)
Barcode	Unique ID (3SCOOL1050383)
Timestamp	Date and time (dd/mm/yyyy hh:mm:ss)
Soort waarneming	Value category (B)
Reden waarneming	Value identifier (01)
Combinedscan	Category and identifier (B01)
Lokcode	Location identifier (AMF)

Additionally, there are certain values that can occur in a scantrail that have nothing to do with the actual physical process. An example of this is the value A4, which is a billing notification. This is a value assigned by a system to the barcode of a parcel, such that the sender of the parcel is billed. The value is given to all parcels at 8 o'clock in the morning, meaning that its 'place' in the scantrail can differ from parcel to parcel (as different parcels can be in different steps of the process at this time). As this will eventually create multiple (unnecessary) process variants, it is better to omit the value as a whole. Besides the A4 value, there are multiple A values that are omitted as well, as they do not influence the physical process.

For the analysis of the datasets, the open source software ProM 6 (process mining framework) is used. ProM is often used and supported by research and has a diverse range of functionalities, due to the availability of many plugins. The software is developed by the University of Eindhoven. It allows for the import from and the export to a wide variety of formats and systems and provides advanced visualization and verification capabilities (van der Aalst, 2016). Besides ProM, another tool originated within the University of Eindhoven, known as Disco. Disco is created by Fluxicon, a company of two PHD students of professor W. van der Aalst (Rozinat & Gunther, 2016). W. van der Aalst is known for his large contribution to the field of process mining, as represented by but not limited to (Buijs et al., 2013; van der Aalst et al., 2007, 2012; van der Werf, Verbeek, & van der Aalst, 2012). The main difference between the tools is on usability (Disco) and on functionality (ProM). As ProM offers more functionality due to the plug in framework, it is used for this study. However, if necessary, Disco can be used as well.

When importing an event log in ProM, it has to be converted to an XES file. XES is an XML-based format, and its name is an acronym for eXtensible Event Stream. The conversion can be done easily using an available plugin (*Convert CSV to XES*) as shown in Figure 27.

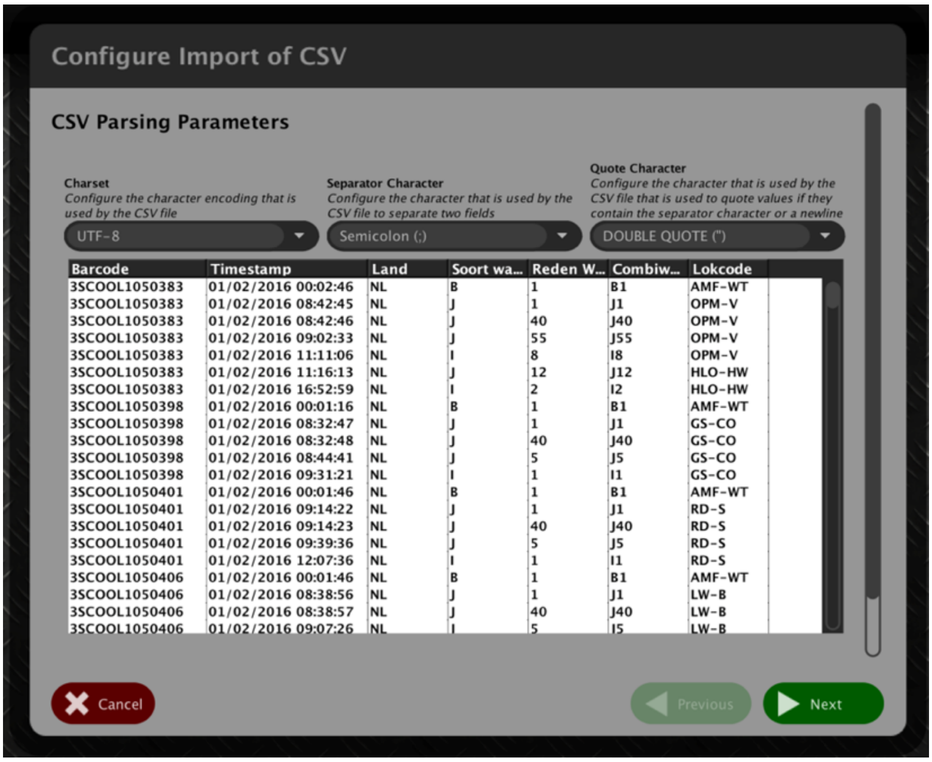


Figure 27: Prom plugin Convert CSV to XES

The XES file can be filtered using certain parameters, for example only taking into account traces that start or end with a specific activity. The filtering is done using the *Filter using simple Heuristics* plugin. The datasets for this study are filtered so that they all end with a I-value, which means that the parcels in the dataset are delivered. This is done to avoid scantrails of parcels that are still in the delivery process, as this might give unwanted and incomplete results.

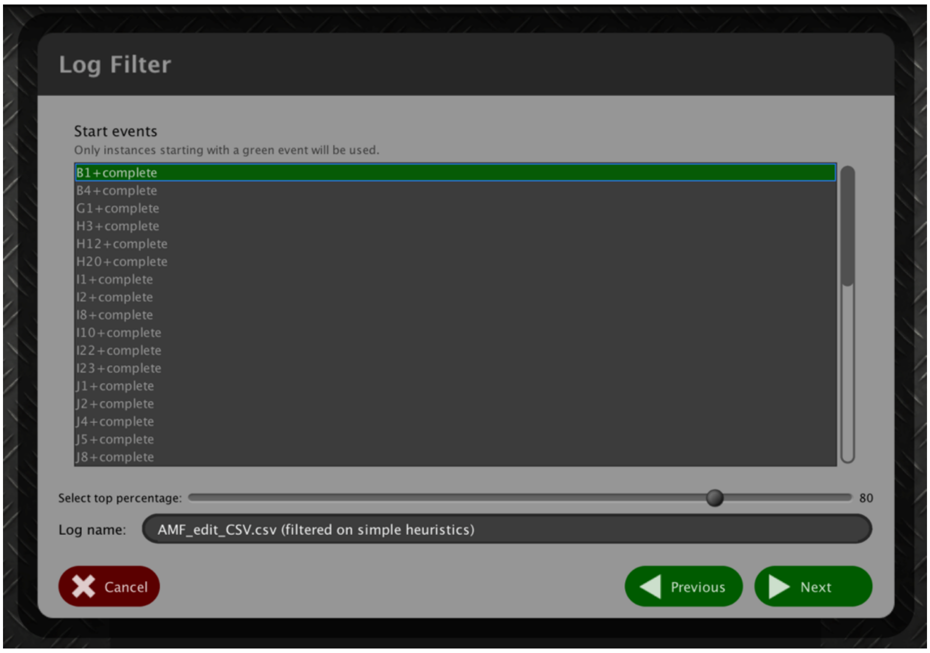


Figure 28: Filtering start events in ProM

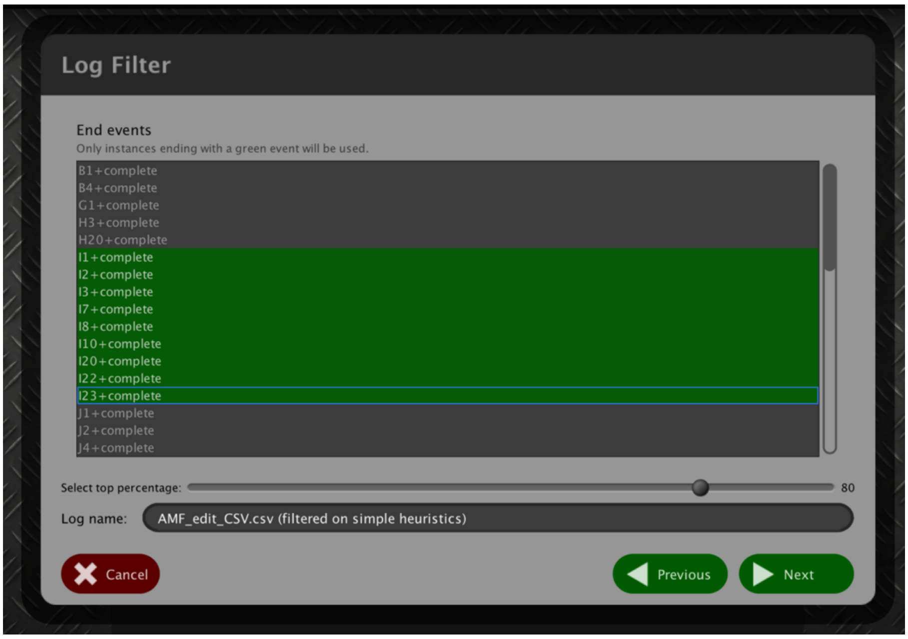


Figure 29: Filtering end events in ProM

The filtered files are used for the analysis. The subsequent steps mining & analysis and evaluation are described and executed in chapter 8: Analysis & results.

7.2 Proposed approach to identify process architectures

As the goal of this study is to create an approach that can be used to identify process architectures using process mining, an initial approach is proposed here. The results of the case study (chapter 8) will be used in chapter 9 to improve the proposed approach and to show a Runtime Enterprise Architecture example. The analysis focusses on identifying process structures on departmental and organizational level, as well as their interaction. The interaction is important to create fully functional process architectures. As described in the related literature, a process architecture is "an organized overview of business processes with relations and guidelines that determine how they must be organized" (Dijkman et al., 2011). The approach can be seen as an extended implementation of step of the PM^2 process mining methodology: Mining and analysis.

(Dijkman et al., 2011) stated that the reference model approach for creating a process architecture was the most easy to use, most popular, and most useful (see literature section 3.2.1 Approaches for designing a process architecture). Therefore, the reference model approach is chosen as starting point for the creation of process architectures. This is done by combining knowledge extracted from literature (specifically, the parts on business process management (chapter 3), process mining in general and artefact centric mining (chapter 4), process improvement (5.2), social network analysis (5.4), and creating functional architectures based on an event log (5.3)), and reference models created using process mining techniques and social network analysis. The initial approach consists of the steps as listed below this paragraph. In chapter 8 steps one to three will be executed. Using these results, in chapter 9 a static process architecture is constructed. Based on this static architecture, a Runtime Enterprise Architecture visualization is proposed.

1. construct an organizational model using process mining
2. construct a departmental model using process mining
3. construct process models using social network analysis
4. combine the models to manually create a static concept process architecture
5. optimize the model to create a final Runtime Enterprise Architecture

7.3 Conclusion

To answer sub question 4, "What are the necessary tools, techniques, and methods to extract departmental and organizational processes from execution data", it can be concluded that execution data often needs data preparation first, due to business rules or other restrictions. The prepared data can be analysed using Process Mining software. Using this software, it is possible to apply additional filters to the data, and to analyse the data. Using the correct plugins, it should be possible to extract departmental and organizational processes from execution data. The PM^2 process mining methodology describes the approach that is executed in the case study. This is done in chapter 8: analysis & results. The analysis that are executed in chapter 8 are used to create several process models. These process models and their interactions are used to create a process architecture in chapter 9.

8. Analysis & Results

In this chapter the results of the case study, that is based on the Process Mining Methodology 2 (described extensively in chapter 7), are presented. Besides executing the case study using the proposed approach, it is important to identify necessary additional steps that are possibly missing. First, basic process mining models are constructed using ProM. These basic process mining models are divided in organizational models and departmental models. Then, using a social network analysis tool a social network is created on organizational as well as on departmental perspective. All these models will be combined into a Archimate model, that shows the enterprise architecture. This model can be seen as a preliminary version of a static process architecture. This static version finally is adapted to a dynamic version. To recap, the main analysis questions are stated below:

1. Basic process mining models
    a. Is it possible to create organizational models based on the gathered execution data?
    b. Is it possible to create departmental models based on the gathered execution data?
2. Social network model
    a. Is it possible to create an artefact centric social network model?
3. Is it possible to create a static process architecture using Archimate?
4. Is it possible to create process architectures based on the data and additional tools?
5. Is it possible to create Runtime Enterprise Architectures? (chapter 9)

8.1 Basic process mining models

In this section the executed analysis are described and the results are shown for analysis questions one and two. Several models are created, and if necessary multiple techniques or tools are used to improve these models. In the ideal situation a hierarchical model is eventually created, which makes it possible to zoom in on the process from high level overview to low level detail. This can be used as starting point for the process architecture.

8.1.1 Creating an organizational process model

The organizational model is created by analysing a dataset containing the location data of the process covering the Netherlands. The 18 distribution centres and additionally some retail locations are present in this dataset. The dataset is filtered to show only the interactions between the distribution centres. This model (Figure 30) is created using the Inductive Visual Miner in ProM, which can be seen as standard process mining analysis. It is clear that this plugin is not made to visualize these kinds of process interactions on organizational level. The interactions are not shown correctly and as much as the model is unreadable it is unusable.

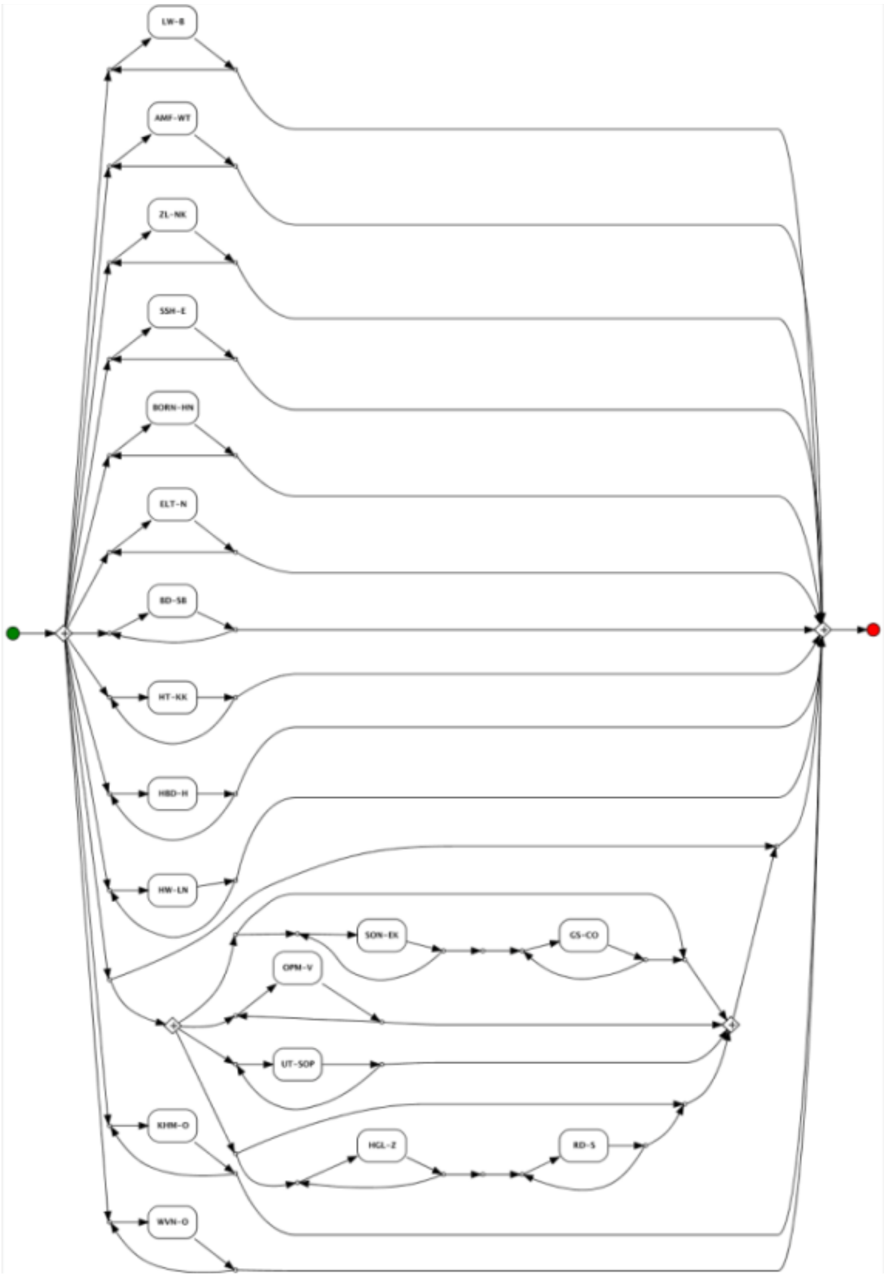


Figure 30: Organizational model created using the inductive miner in ProM

As this model clearly is not the best way to represent the organizational structure, another plugin of ProM was used to create an organizational model. The plugin is known as the social network miner, and is described in the paper *Discovering social networks from event logs* (Van Der Aalst, Reijers, & Song, 2005). The model created using this miner still is not very clear, as it shows all the routes between all sorting centres (see Figure 31). In the ideal situation, the thickest routes are the routes with the highest number of packages, and the size of the circle should indicate the number of incoming and outgoing routes. This kind of analysis can not be done with ProM, since the focus is on the process execution instead of creating some kind of a logistics network. Besides, the original format of the data can not be used to do such an analysis. More on this and on the conversion in of the data is described in section 8.2.

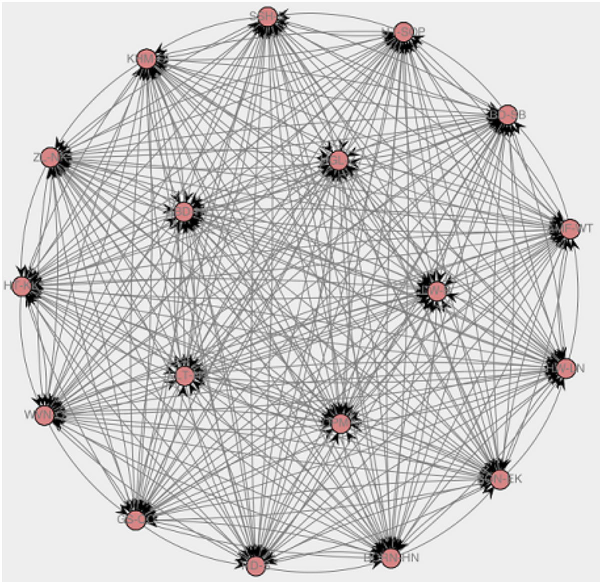


Figure 31: Location/organizational model based on social network miner

It can be noted that it is possible to create organizational models using the execution data gathered. However, the models are not clear and not useful. These models do not create new insights into the process, and they do not enable process managers to identify process problems or bottlenecks. The reason for this is partly based on the structure of the data, and partly on the limitations by the used plugins. Other ways to create an organizational model need to be identified, which is done in section 8.2.

8.1.2. Creating a departmental process model

The same process mining techniques are now applied to create a departmental model. The dataset that is used is constructed based on the most occurring traces of the 18 distribution centres. This is done by analysing the organisational dataset and then selecting the most occurring traces. ProM shows the frequencies of occurred traces, which makes it easy to select the right traces. An example of the frequencies visualization by ProM is shown in Figure 32 below. The figure shows the first five most frequent traces.

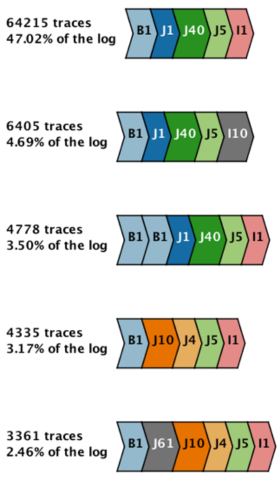


Figure 32: Most occurring traces in organisational dataset

Table 9 shows the 20 most occurring traces, that together cover 72,43% of all the scantrails. The log in total contains 136575 traces.

Table 9: Scantrails of 20 most occurring traces

Number	% of the log	Cumulative %	Trace (scantrail)
1	47.02%	47,02%	B1 > J1 > J40 > J5 > I1
2	4.69%	51,71%	B1 > J1 > J40 > J5 > I10
3	3.50%	55,21%	B1 > B1 > J1 > J40 > J5 > I1
4	3.17%	58,38%	B1 > J10 > J4 > J5 > I1
5	2.46%	60,84%	B1 > J61 > J10 > J4 > J5 > I1
6	1.70%	62,54%	B1 > J1 > J40 > J55 > I8 > J12 > I2
7	1.58%	64,12%	B1 > S2 > V90 > B1 > B1 > J40 > J5 > I1
8	1.18%	65,30%	B1 > J1 > J40 > J5 > J8 > J30 > J1 > J1 > J40 > J5 > I1
9	1.15%	66,45%	B1 > J1 > J1 > J40 > J5 > I1
10	1.00%	67,45%	B1 > J1 > J40 > J5 > J8 > J30 > J1 > J1 > J40 > J5 > K70 > J30 > B1 > J1 > J40 > J55 > J21 > J2 > I2
11	0.82%	68,27%	B1 > K50 > J1 > J40 > J5 > I1
12	0.77%	69,04%	B1 > J1 > J5 > I1
13	0.69%	69,73%	B1 > J1 > J40 > J5 > J32 > J44 > I1
14	0.44%	70,17%	B1 > J5 > J1 > J40 > I1
15	0.41%	70,58%	B1 > J1 > J40 > J5 > J8 > J30 > K33 > B1 > J1 > J1 > J40 > J55 > J21 > J2 > I2
16	0.40%	70,98%	B1 > V90 > S2 > B1 > B1 > J40 > J5 > I1
17	0.39%	71,37%	B1 > J1 > J40 > J4 > J5 > I1
18	0.38%	71,75%	B1 > J17 > J1 > J1 > J40 > J5 > I1
19	0.35%	72,10%	B1 > I8 > J12 > J1 > J40 > J55 > I2
20	0.33%	72,43%	B1 > B1 > J1 > J10 > J19 > J20 > J20 > J5 > I1

For each of the 20 traces, a barcode is selected representing the trace. These 20 barcodes and their events are used to create a new dataset that only represents these 20 paths. Using ProM, a petri net is created based on those most occurring traces. This is done by visualizing the paths using the Inductive Visual Miner, and then saving the visualisation as a petri net. The same is done using only the first 10 traces, since these traces together already cover 67,45% of all scantrails. The constructed petri net representing the 20 traces is shown in Figure 34, the petri net using the 10 most occurring traces is shown in Figure 33.

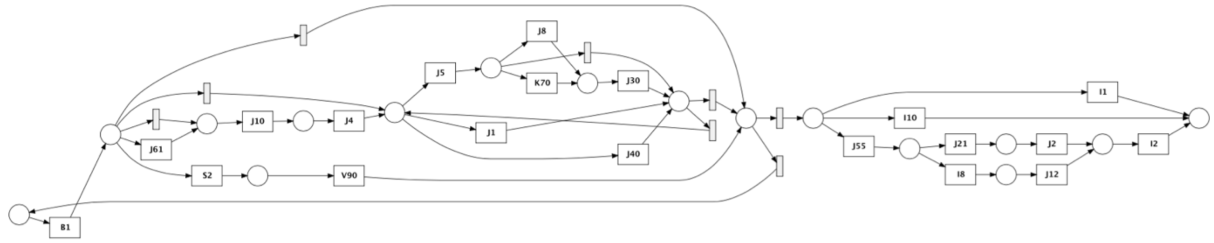


Figure 33: Departmental model created using 10 most occurring traces

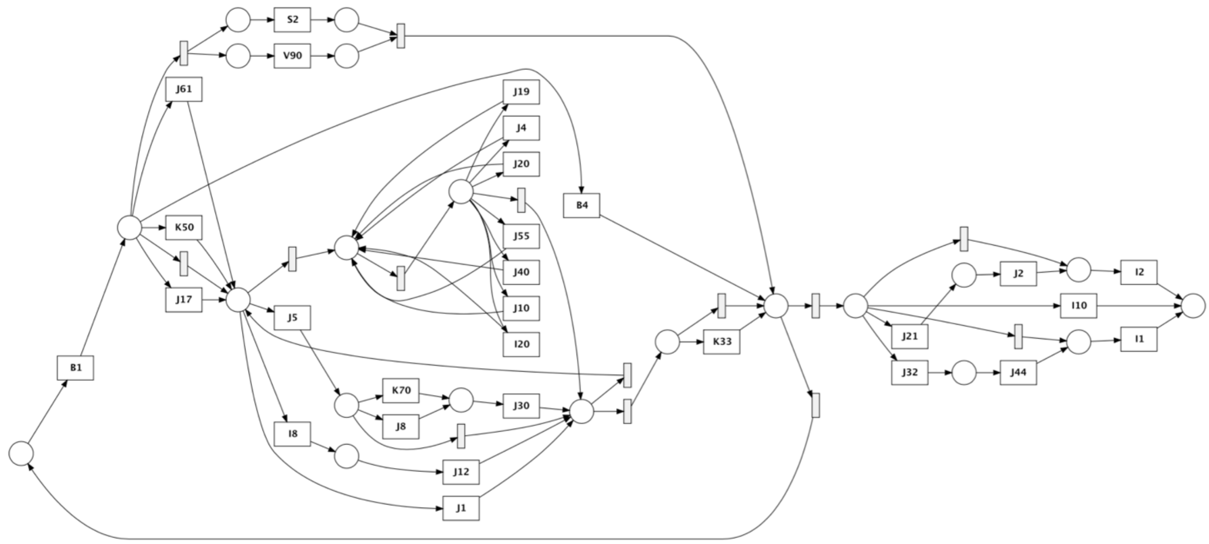


Figure 34: Departmental model created using 20 most occurring traces

As seen, this is model is not very clear. The number of deviations is high, and this petri net still does not represent all the possible paths. To check how well the processes of the distribution centres conform to the created petri net, the logs are *replayed* through the petri net. This is done using the *Replay for conformance checking* plugin in ProM. This results in fitness values: as mentioned in the literature, the fitness value indicates how well a model represents a log (see section 4.4). A fitness value of around 0.8/0.9 indicates a good model. The fitness values of these petri nets are shown in Table 10.

Table 10: Fitness values per distribution centre

Distribution centre	Fitness value (20 most occurring traces)	Fitness value (10 most occurring traces)
Amersfoort	0.990	0.911
Born	0.986	0.956
Breda	0.983	0.960
Den Bosch	0.975	0.916
Den Hoorn	0.975	0.964
Elst	0.900	0.870
Goes	0.991	0.957

Halfweg	0.968	0.943
Hengelo	0.978	0.954
Kolham	0.970	0.941
Leeuwarden	0.984	0.954
Opmeer	0.979	0.955
Ridderkerk	0.983	0.968
Sassenheim	0.973	0.954
Son	0.980	0.952
Utrecht	0.986	0.949
Waddinxveen	0.980	0.948
Zwolle	0.971	0.949

To simplify the model, the main categories of the values are combined. Thus, no distinction is made between for example a J4 value and a J5 value: these are both seen as a J. This version of the departmental model is shown below. It can be seen that it consists of 3 main activities: the distribution centre (evening), the distribution centre (morning), and the delivery part. All the 'B' values are seen as distribution (evening), all the J values as distribution (morning), and all the 'I' values as delivery. This is visualized in Figure 35.

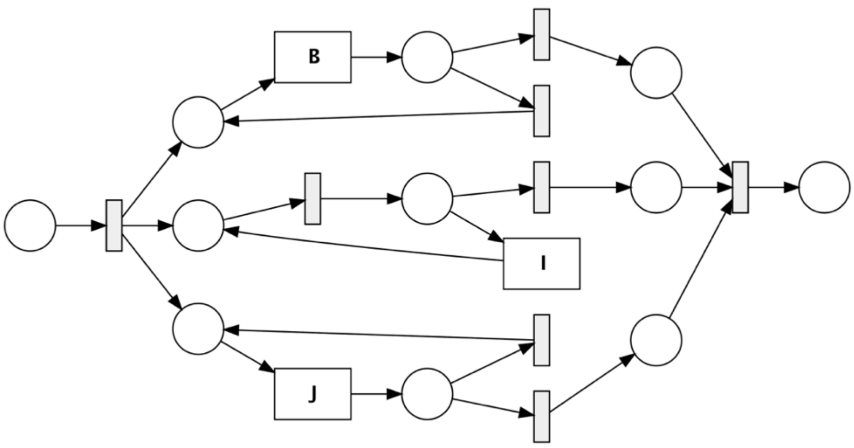


Figure 35: Departmental model using only main categories

The process of one distribution centre contains more activities than shown in the figure above. However, these "missing" activities are part of one of the three main categories mentioned. Some parcels for example are returned by clients. These parcels receive some additional values, but are still part of the main distribution centre process.

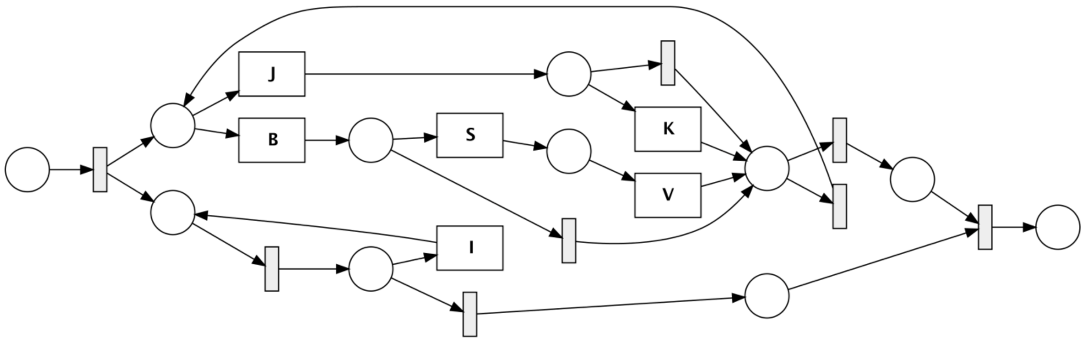


Figure 36: Extended departmental model using only main categories

Figure 36 shows these additional flows between the sub processes. The additional distribution scans are mainly concerning errors in the process that need to be corrected (V, S). Like the other models, this model does not show the flow of the parcels – it only shows possible combinations of scans in a scantrail.

Figure 37, Figure 38 and Figure 39 represent the petri nets of the sub processes of respectively B, J, and I.

Table 11 shows the accompanying fitness values of these processes.

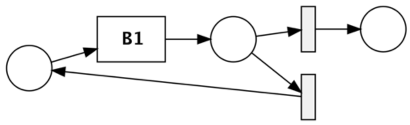


Figure 37: Petri net of sub process B

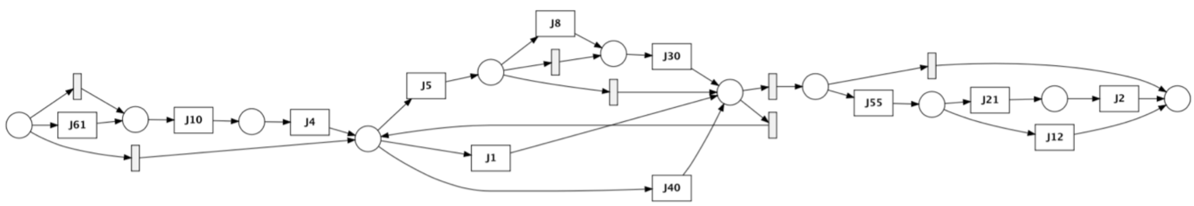


Figure 38: Petri net of sub process J

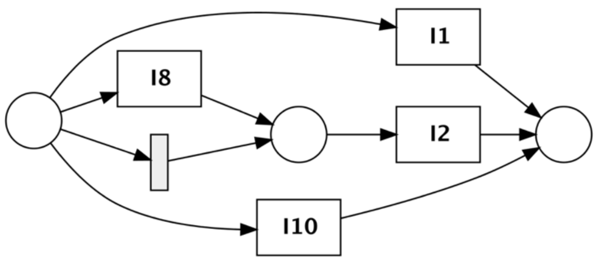


Figure 39: Petri net of sub process I

Table 11: Fitness values of sub processes

Number	Distribution centre	Fitness of B	Fitness of J	Fitness of I
1	Amersfoort	1.0	0.912	0.936
2	Born	1.0	0.963	0.993
3	Breda	1.0	0.976	0.986
4	Den Bosch	1.0	0.909	0.937
5	Den Hoorn	1.0	0.983	0.990
6	Elst	1.0	0.944	0.978
7	Goes	1.0	0.969	0.997
8	Halfweg	1.0	0.942	0.990
9	Hengelo	1.0	1.0	0.985
10	Kolham	1.0	0.935	0.970
11	Leeuwarden	1.0	0.962	0.984
12	Opmeer	1.0	0.975	0.973
13	Ridderkerk	1.0	0.993	0.950
14	Sassenheim	1.0	0.962	0.956
15	Son	1.0	0.956	0.956
16	Utrecht	1.0	0.959	0.969
17	Waddinxveen	1.0	0.945	0.996
18	Zwolle	1.0	0.954	0.988

It can be noted that all these departmental models are not very clear. As the process that is analysed is inherently a process guided by packages, the artefact centric approach seems more appropriate - since artefact centric process mining focusses on the lifecycle of physical artefacts. Besides, the use of petri nets is not the best way to visualize the process flow. In the following section a combination of artefact centric mining and social network analysis is made, with the expectation that this leads to improved models.

8.2 Combining artefact centric mining and social network analysis

To improve the models a combination of artefact centric mining and social network analysis is made. Using artefact centric mining it is possible to identify the transitions between distribution centres. A transition is a parcel that travels from A to B. These transitions can be converted to weights which are necessary to construct social network diagram. To correctly do this, all the transitions between distribution centres have to be analysed.

Since the data is not structured like this, some changes have to be made. Using a SQL database, the organizational dataset containing all the data of the Netherlands is adapted. First, all unique barcodes are grouped in such a way that for every barcode there is only one row with multiple columns. These columns all represent a location. Each row now looks like this: barcode; location 1; location 2; location N. This is restructured to unique rows with only 2 locations per row, which leads to: row 1: barcode; location 1; location 2; row 2: barcode; location 2; location N. This data is now used in Excel to create a PivotTable. The PivotTable shows the relative values between all sorting centres, as shown in Table 12.

Table 12: Artefact centric mined location transition weights

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0,000	0,598	0,536	0,554	0,017	0,361	0,044	1,529	0,031	0,392	0,018	0,023	0,013	0,010	0,034	0,041	1,080	0,559
2	0,745	0,000	0,559	0,453	0,065	0,222	0,058	0,759	0,121	0,306	0,047	0,126	0,067	0,068	0,132	0,131	0,631	0,492
3	0,804	0,519	0,000	0,513	0,053	0,222	0,067	0,762	0,131	0,267	0,026	0,350	0,121	0,078	0,320	0,102	0,682	0,476
4	1,082	0,607	0,745	0,000	0,122	0,364	0,163	0,964	0,250	0,370	0,050	0,627	0,171	0,131	0,486	0,237	0,834	0,678
5	0,422	0,316	0,271	0,296	0,000	0,097	0,000	0,344	0,003	0,188	0,018	0,001	0,001	0,000	0,003	0,000	0,337	0,246
6	1,080	0,687	0,783	0,617	0,210	0,000	0,271	0,924	0,625	0,360	0,087	1,485	0,515	0,163	1,011	0,402	1,222	0,594
7	0,785	0,388	0,381	0,389	0,003	0,196	0,000	0,506	0,006	0,298	0,006	0,003	0,000	0,000	0,004	0,006	0,416	0,510
8	1,107	0,718	0,624	0,442	0,011	0,304	0,009	0,000	0,016	0,415	0,030	0,001	0,014	0,026	0,009	0,043	0,748	0,470
9	0,995	0,478	0,422	0,307	0,000	0,270	0,007	0,728	0,000	0,249	0,018	0,010	0,013	0,001	0,018	0,010	0,593	0,362
10	0,853	0,523	0,701	0,517	0,098	0,296	0,162	0,603	0,257	0,000	0,084	0,849	0,259	0,166	0,607	0,252	0,769	0,687
11	0,620	0,304	0,394	0,372	0,044	0,166	0,065	0,418	0,135	0,271	0,000	0,601	0,186	0,064	0,343	0,132	0,488	0,476
12	0,729	0,519	0,438	0,451	0,001	0,190	0,003	0,551	0,010	0,294	0,007	0,000	0,028	0,000	0,007	0,000	0,497	0,324
13	0,610	0,495	0,483	0,360	0,003	0,213	0,000	0,540	0,009	0,237	0,001	0,001	0,000	0,000	0,003	0,020	0,613	0,540
14	0,561	0,409	0,469	0,355	0,000	0,200	0,003	0,556	0,001	0,252	0,013	0,003	0,000	0,000	0,001	0,003	0,465	0,468
15	0,853	0,495	0,398	0,352	0,003	0,213	0,004	0,620	0,010	0,257	0,017	0,010	0,003	0,001	0,000	0,006	0,571	0,411
16	1,011	0,681	0,590	0,583	0,006	0,290	0,007	0,787	0,011	0,515	0,018	0,014	0,018	0,001	0,006	0,000	0,648	0,618
17	1,482	0,611	0,658	0,557	0,088	0,446	0,094	1,148	0,213	0,350	0,041	0,209	0,152	0,088	0,244	0,266	0,000	0,529
18	1,018	0,594	0,696	0,573	0,090	0,297	0,271	0,750	0,233	0,388	0,087	0,483	0,136	0,128	0,283	0,252	0,675	0,000
Tot	14,76	8,94	9,15	7,69	0,81	4,35	1,23	12,49	2,06	5,41	0,57	4,80	1,70	0,93	3,51	1,90	11,27	8,44

8.2.1 Creating a social network based organizational model

The social network diagram shows the most important distribution centres as well as the relations between distribution centres. The model can be seen as a new variant of standard organizational models, as it represents the organizational structure including the artefact flow that is present in the organization. In the model below, created using the social network analysis software Gephi, the structure is depicted using the data that is depicted in Table 12 (M. Bastian, S. Heymann, 2009).

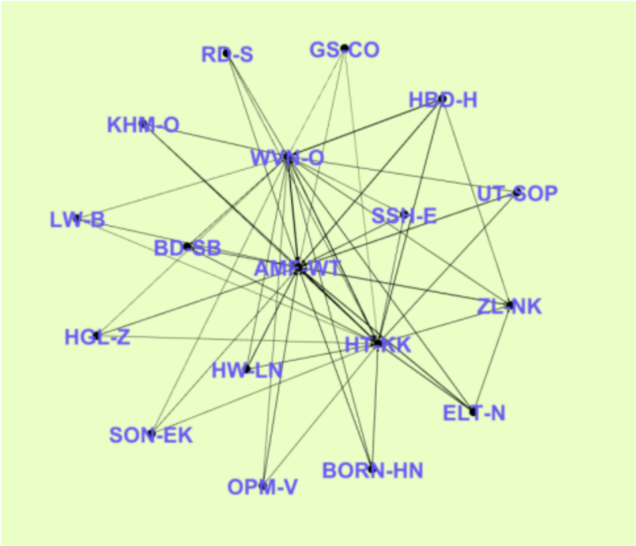


Figure 40: Organizational model using Social Network Analysis

In this figure (Figure 40) the circles indicate the centrality of the sorting centre, and the thickness of the relations indicates the number of packages that flow between these centres. This model shows the relations between the distribution centres and the flow of the parcels. However, it can not be used to improve the execution of the process or get innovative new insights in the process. The important parts of this visualization that can be used for further research are the usage of relation thickness to indicate the strength of the relation, and the size of the distribution centre to indicate its importance in the network.

8.2.2 Creating a retail based location model

The final location model is created based on just one distribution model and its local retailers. Retailers are shops were customers can bring and receive packages. These retailers play an important role in the overall network. However, due to the high number of retailers it is impossible to show them in the location network with all the other sorting centres and retailers of the entire Netherlands. Therefore, a selection is made to create an idea of this process flow. AMF is the distribution centre, ALR locations are retailers.

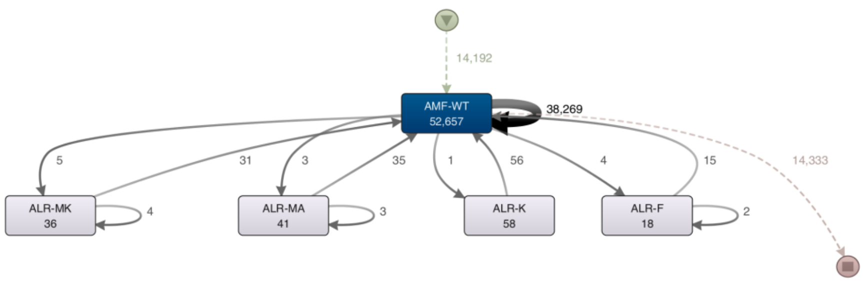


Figure 41: Location model with some retailers

All models and used techniques of sections 8.1 and 8.2 are now combined to create a general process architecture in section 8.3 – using the modelling language Archimate.

8.3 Static process architecture

A static process architecture can be created using Archimate. As noted in section 3.2.1 Approaches for designing a process architecture, process architecture frameworks and design approaches such as the Archimate framework are used to create an understanding of the business components and their interactions. These approaches try to link business process to business goals or business functions. This enterprise architecture focus on process architectures can be of help when creating process architectures, but the creation of enterprise architectures is inherently different than the creation of process architectures. This is due to the fact that enterprise architectures focus on a static representation of the organizations structure, while process architectures aim to represent some kind of process flow. The enterprise modelling languages can be used to graphically represent business processes and their relations, but are not primarily focused on the design of a business process architecture.

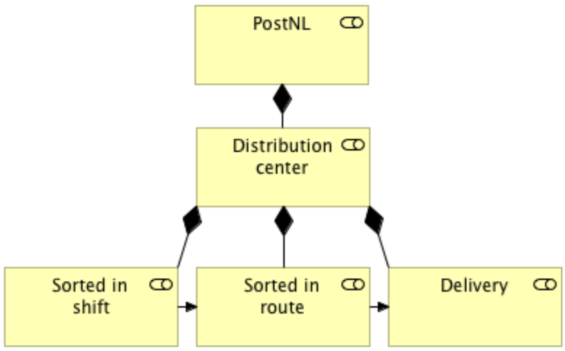


Figure 42: Static process architecture created using Archimate

Figure 42 below shows a basic version of PostNL architecture represented using Archimate. It visualizes the architecture as it is, but there is no possibility to correctly and dynamically show the interactions and process information in this model. Many details are missing, such as the numbers of parcels travelling from one distribution centre to another. Additionally, there is no possibility to dynamically change this model due to specific circumstances.

8.4 Conclusion

The models presented in this chapter are the result of an extensive case study. Process execution logs were analysed using the software tool ProM. Multiple models were created, on departmental and organizational level as well as social network variants. Each model gives a certain insight in the process, but the consistency between the models is missing. The Archimate models combines most of the information presented in the other models – but it simplifies the complex process since Archimate it not made to visualize process architectures.

Some of the models show how the parcels travel throughout the distribution network. These models are artefact centric: they are constructed by analysing how the parcels move instead of creating a process models that shows all the possible process routes based on the event log. The answer to sub question 5, "Is it possible to identify and analyse the lifecycles of the artefacts that are the subject of a business process?" is yes. In this case it is useful to use the artefact process information to create process models and eventually a process architecture – since the entire process focusses on these artefacts (parcels). The models presented are a step forward in visualizing the entire process architecture as one model, which is the final step of this study. The Runtime Enterprise Architecture should be able to present process information like median duration times, fitness, and process flows. Mock ups of such a visualization are created and presented in chapter 9.

9. Creating a Runtime Enterprise Architecture

The easiest way to create a process architecture is using a reference model (Dijkman et al., 2011). When there is no reference model present, other approaches should be used. These are however more time consuming and harder to understand (Dijkman et al., 2011). To simplify the approach, the focus of the proposed method (chapter 7) is to construct reference models using process mining and social network analysis software. These models can then be combined to create a process architecture.

The initial approach as proposed in chapter 7 consisted of the following steps:

1. construct an organizational model using process mining
2. construct a departmental model using process mining
3. construct a social network model that shows multiple process levels
4. combine the models to create a concept process architecture
5. optimize the model to create a Runtime Enterprise Architecture

The steps one, two, and three of the proposed method were executed in section 8.1 and 8.2. Several organizational and departmental models were constructed, and additionally a retail model is created. At last, in chapter 8 a simplified process architecture is created using Archimate. In this chapter, the previously created models will be used to design a concept process architecture (step four). Finally, an interactive mock up of the final process architecture is made using the concept process architecture (step 5).

9.1 Creating a conceptual process architecture

In the following conceptual architecture different models are combined to show process combinations on different levels. This model is a simple representation of the architecture based on the event logs. The most important aspect that is missing here is information regarding the process execution. For example, the numbers of parcels that is distributed from location one to location two, or the average waiting time of this event log at a certain point.

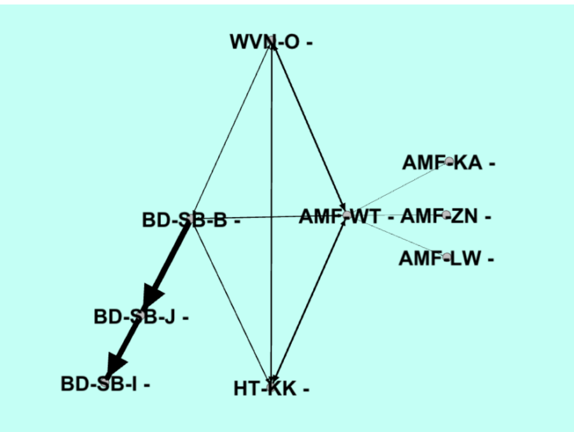


Figure 43: Conceptual Process architecture foundation

In Figure 43 on the left there is the distribution centre Breda, which is divided in three main activities: B (night sorting), J (morning sorting), and I (delivery). This creates insight in the process on low level. The distribution centres Waddinxveen (WVN-O) and Den Bosch (HT-KK) are represented without specific detail. Distribution centre Amersfoort is shown with some additional detail, showing three retailers. The only level that is missing in this architecture is the specific detail within a main activity in a distribution centre (such as J1, J4, J5 instead of J). Concluding, there is a lot to improve on the current design.

9.1.1 Finalizing the process architecture

The conceptual process architecture foundation shows the process structure of the distribution process of PostNL. However, the model is not clear and too ambiguous to understand. Nonetheless it is used as foundation of the creation of the final process architecture. It is important to identify all the important aspects of the process architecture, in order to create a complete and interactive version. Figure 44 shows the most basic variant of the process architecture: distribution centres are represented by the orange circles, their relations by the orange arrows. The architecture shows that in essence every distribution centre can be connected to every other distribution centre. This has to be taken into account, since visualizing 18 distribution centres that all can be connected to each other is a challenge.

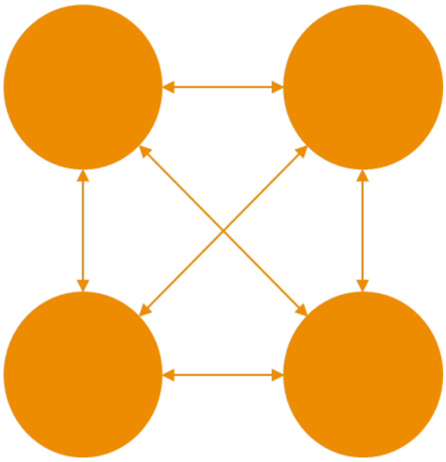


Figure 44: High level process architecture

Figure 45 shows more detail in the process architecture. One of the distribution centre illustrates the three main activities executed in every distribution centre: Evening sorting (B), Morning sorting (J), and Distribution (I). It is important to note here that all 18 distribution centres consist of these sub processes. This is the second challenge in the design of the final process architecture: it should visualize the 18 distribution centres uniquely, as well as the performance of the sub processes in those distribution centres.

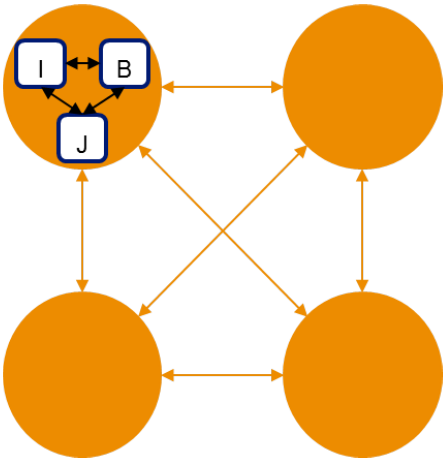


Figure 45: Process architecture with some detail

When zooming in further on the activities of one distribution centre, more internal sub-activities can be identified. This is shown in Figure 46. As described in chapter 4, the morning sorting for example consists of multiple activities. A parcel first receives a J1 scan, then a J4, and eventually a J5. Thereafter, a parcel receives at least one of the following I scans, indicating that the parcel is delivered. In the ideal situation, the final process architecture can visualize the 18 distribution centres and their "higher" level sub process performance. This could be extended by also being able to show the performance of the sub processes (for example the J1, J4, and J5 scans in sub process 'J').

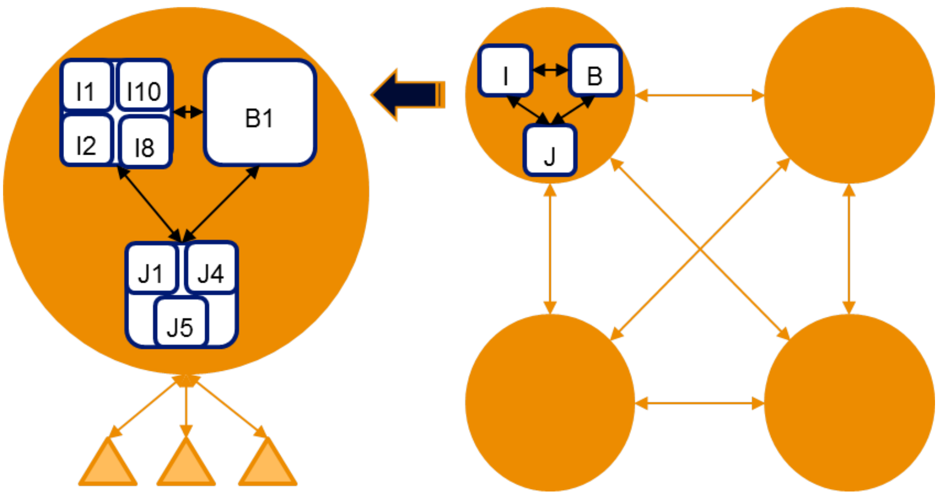


Figure 46: Detailed process architecture

Additionally, retail locations are depicted in this architecture. Retail locations are shops were customers can retrieve or deliver their parcels. All the delivered parcels are retrieved by the delivery man at the end of the day. These retail locations are not the main focus of the final process architecture – since the most important part of the process execution is the main process that consist of the interactions within and between distribution centres. Figure 46 shows all the possible process level in one architecture, but the structure is not clear. In section 9.1.2 a more hierarchical version of the process architecture is presented. Additional improvements that can be made for the final architecture is changing the size of distribution centres as well as the size of the connection arrows to indicate the volumes flowing through the network.

9.1.2. Extended process architecture (top-down)

An example extended process architecture is shown below. It connects the models or adaptions of the models shown in the previous section. This figure can also be seen a 'process system', which can be described as a system of processes consisting of processes. In such a process system a hierarchy can be identified, such as visualized below. Studying systems in a hierarchical way makes it easier to understand the complexity of the structure.

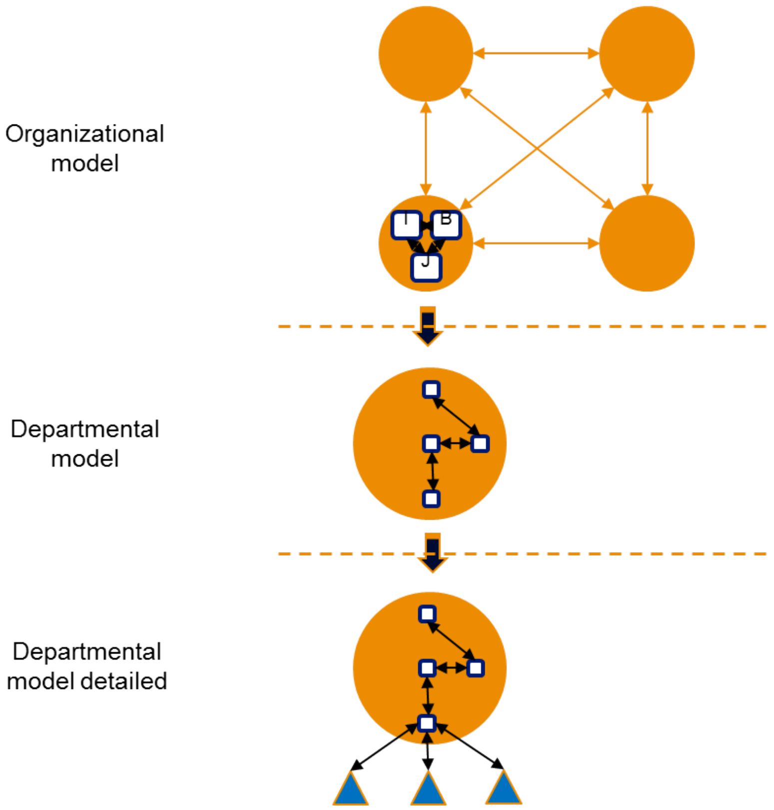


Figure 47: Extended process architecture

9.1.3 Design principles for the final process architecture visualization

In section 9.2 the Runtime Enterprise Architecture (REA) will be designed. Therefore, it is important to summarize the design principles that are identified throughout the first section of this chapter, or elsewhere in this thesis. These principles are listed below:

- The REA should visualize 18 distribution centres
- The REA should visualize the flows between the 18 distribution centres
- The size of the flows in the REA should represent the volumes travelling through the process
- The REA should visualize sub process performance of these 18 distribution centres
- The REA design should be consistent and clear

9.2 Creating a Runtime Enterprise Architecture

The Runtime Enterprise Architecture is an interactive representation of the conceptual process architecture that is discussed in section 9.1. This interactive representation has to be easy to adapt, so it can be reused with new data. The design that is shown in this section is based on real life event log data. Several elements of these event logs are used in the visualization, in different ways. By using the data in several ways it is possible to give a high level as well as a detailed insight in the process in one visualization.

9.2.3 Basis of the visualization

Many ideas for the Runtime Enterprise Architecture are already coined throughout this thesis. They are summarized in section 9.1.3. Nowadays many visualization possibilities exist, using specific software packages or visualization libraries. An innovative way to do this is using D3.JS, which is a JavaScript library for "producing dynamic, interactive data visualizations" (Bostock, 2016). Many examples of interactive visualizations can be found on their website, and the Runtime Enterprise Architecture is a combination of ideas and these examples.

The basis of the visualization is known as the 'Chord diagram'. The chord diagram is a visualization that is able to represent a network in such a way that the flows within the network are shown without creating a chaos. This is necessary in the case of process execution visualizations since there are often many flows to visualize. An example of such a Chord diagram visualization is shown below.

Figure 48: Example of a Chord diagram (Bostock, 2016)

This visualization covers several design principles that were mentioned earlier. This design is able to visualize the 18 distribution centres, and the flows between these distribution centres. Additionally, the size of the flows represents the volumes travelling through the process. The size of the nodes in this design represent how 'important' such a node is in the network: in social network analysis terminology known as the 'centrality'.

Since the idea is to show high level process information (such as shown in the chord diagram) and low level detailed information in the same Runtime Enterprise Architecture visualization, the chord diagram had to be extended. This is done by using the idea of the 'sunburst' visualization, as shown in Figure 49.

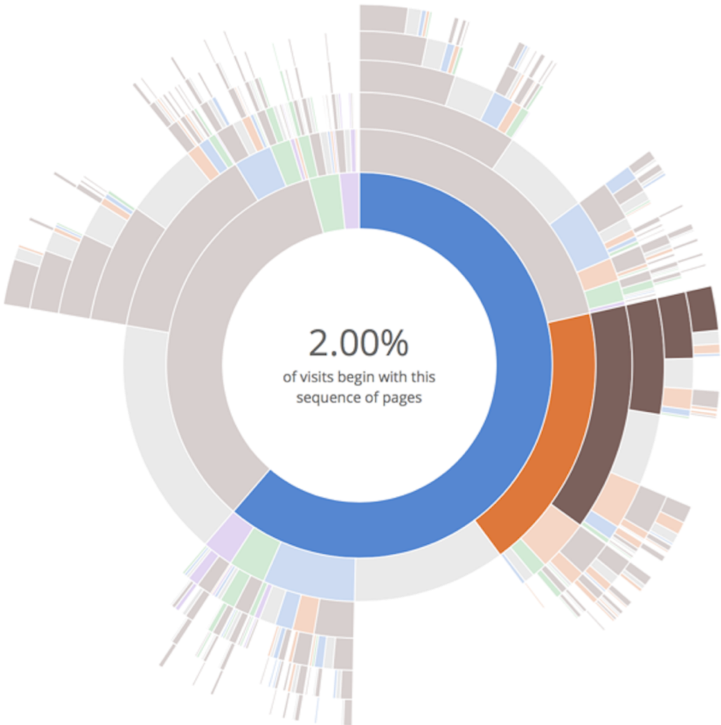


Figure 49: Example of Sunburst visualization (Bostock, 2016)

This sunburst visualization example can be used to zoom in on a certain process execution. In this case, it shows the sequences of webpages of a website that the visitors of the website navigate through. The further from the centre, the more detailed the information becomes. This idea, where 'sunbursts' can be used at the outer ring of the Chord diagram, create the possibility to visualize high level information (the chord part) as well as detailed information (the sunburst part). Since this thesis focusses on parcels and the distribution data of these parcels, data preparation is necessary to create this Runtime Enterprise Architecture visualization.

9.2.2 Data Preparation

To correctly visualize the event data, some calculations have been executed using ProM and if necessary with spreadsheet software such as Excel. The information that is necessary is:

- Determining the flows between distribution centres (the chord part)
- Determining the flows within distribution centres (the sunburst part)
- Determining the median duration time of a distribution centre (the sunburst part)

The basis for this information is the same as for the analysis that have been executed in chapter 8. Using the data of one month covering all the 18 distribution centres, the flows can be identified by focussing on the location changes that are present in the data. Then all these location changes (for example, 100x from Amersfoort to Utrecht and 200x from Kolham to Leeuwarden) are normalised using the total number of location changes. All location changes are now relative to each other, and when summed they account for 100% of all changes.

Table 13: Normalised values for distribution flows

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0,000	0,598	0,536	0,554	0,017	0,361	0,044	1,529	0,031	0,392	0,018	0,023	0,013	0,010	0,034	0,041	1,080	0,559
2	0,745	0,000	0,559	0,453	0,065	0,222	0,058	0,759	0,121	0,306	0,047	0,126	0,067	0,068	0,132	0,131	0,631	0,492
3	0,804	0,519	0,000	0,513	0,053	0,222	0,067	0,762	0,131	0,267	0,026	0,350	0,121	0,078	0,320	0,102	0,682	0,476
4	1,082	0,607	0,745	0,000	0,122	0,364	0,163	0,964	0,250	0,370	0,050	0,627	0,171	0,131	0,486	0,237	0,834	0,678
5	0,422	0,316	0,271	0,296	0,000	0,097	0,000	0,344	0,003	0,188	0,018	0,001	0,001	0,000	0,003	0,000	0,337	0,246
6	1,080	0,687	0,783	0,617	0,210	0,000	0,271	0,924	0,625	0,360	0,087	1,485	0,515	0,163	1,011	0,402	1,222	0,594
7	0,785	0,388	0,381	0,389	0,003	0,196	0,000	0,506	0,006	0,298	0,006	0,003	0,000	0,000	0,004	0,006	0,416	0,510
8	1,107	0,718	0,624	0,442	0,011	0,304	0,009	0,000	0,016	0,415	0,030	0,001	0,014	0,026	0,009	0,043	0,748	0,470
9	0,995	0,478	0,422	0,307	0,000	0,270	0,007	0,728	0,000	0,249	0,018	0,010	0,013	0,001	0,018	0,010	0,593	0,362
10	0,853	0,523	0,701	0,517	0,098	0,296	0,162	0,603	0,257	0,000	0,084	0,849	0,259	0,166	0,607	0,252	0,769	0,687
11	0,620	0,304	0,394	0,372	0,044	0,166	0,065	0,418	0,135	0,271	0,000	0,601	0,186	0,064	0,343	0,132	0,488	0,476
12	0,729	0,519	0,438	0,451	0,001	0,190	0,003	0,551	0,010	0,294	0,007	0,000	0,028	0,000	0,007	0,000	0,497	0,324
13	0,610	0,495	0,483	0,360	0,003	0,213	0,000	0,540	0,009	0,237	0,001	0,001	0,000	0,000	0,003	0,020	0,613	0,540
14	0,561	0,409	0,469	0,355	0,000	0,200	0,003	0,556	0,001	0,252	0,013	0,003	0,000	0,000	0,001	0,003	0,465	0,468
15	0,853	0,495	0,398	0,352	0,003	0,213	0,004	0,620	0,010	0,257	0,017	0,010	0,003	0,001	0,000	0,006	0,571	0,411
16	1,011	0,681	0,590	0,583	0,006	0,290	0,007	0,787	0,011	0,515	0,018	0,014	0,018	0,001	0,006	0,000	0,648	0,618
17	1,482	0,611	0,658	0,557	0,088	0,446	0,094	1,148	0,213	0,350	0,041	0,209	0,152	0,088	0,244	0,266	0,000	0,529
18	1,018	0,594	0,696	0,573	0,090	0,297	0,271	0,750	0,233	0,388	0,087	0,483	0,136	0,128	0,283	0,252	0,675	0,000
Tot	14,76	8,94	9,15	7,69	0,81	4,35	1,23	12,49	2,06	5,41	0,57	4,80	1,70	0,93	3,51	1,90	11,27	8,44

This matrix contains all distribution flows of all 18 distribution centres. Of course, from distribution centre 1 to distribution centre 1, the flow is indicated as 0,000 since this is not possible. The bottom row values sum up to one hundred percent.

To determine the flows within a distribution centre, it is necessary to look at the ratio between the number of parcels with a B scan (accepted in distribution centre) and parcels with a I scan (delivered). If the ratio is high, this means that most of the parcels that arrive at this distribution centre (coming from retail companies for example) are distributed in this region. An example of such a distribution centre is number 9. An example of a distribution centre where the ratio is rather low is distribution centre 10. This distribution centre is located near the border of the Netherlands, and many parcels enter the network here but are not distributed in the same area.

Table 14: Ratio between B/J/I and total sunburst

	# of B	# of J	# of I	Total	Share B	Share J	Share I	Share of total
1	22852	17340	16882	57074	40%	30%	30%	8%
2	16102	13774	11880	41756	39%	33%	28%	6%
3	16007	13057	12332	41396	39%	32%	30%	6%
4	15065	11861	11107	38033	40%	31%	29%	5%
5	21260	13589	10202	45051	47%	30%	23%	6%
6	17703	13841	11209	42753	41%	32%	26%	6%
7	15958	12370	12124	40452	39%	31%	30%	5%
8	20621	17786	11059	49466	42%	36%	22%	7%
9	11984	11983	10892	34859	34%	34%	31%	5%
10	18937	10103	9654	38694	49%	26%	25%	5%
11	14372	10512	10022	34906	41%	30%	29%	5%
12	16793	10593	10686	38072	44%	28%	28%	5%
13	18327	12835	9707	40869	45%	31%	24%	6%
14	17083	12382	10234	39699	43%	31%	26%	5%
15	15414	12029	10159	37602	41%	32%	27%	5%
16	13636	13523	9840	36999	37%	37%	27%	5%
17	16011	14825	11192	42028	38%	35%	27%	6%
18	16452	11927	10353	38732	42%	31%	27%	5%
			Total	738441			Total	100%

To use this information in the visualization, each sunburst consists of one "block" of 100%, divided in a B, a J, and an I part. Each block however should also be relative to the blocks in other sunburst parts. The 'sunburst size ratio' calculates the total number of B, J and I scans per distribution centre, and shows in a percentage the share of each distribution centre (Table 14). Distribution centre #1 accounts for the highest combined number of B, J and I scans, thus this 'block' in the sunburst is relatively seen the largest (8%). Of this sunburst, the B part has a size of 40%, and the J and I parts of 30%.

The median duration time is the time that a parcel is within the distribution centre. This data is extracted from the event logs using the process mining software Disco and is shown in the Table below. The table is sorted from shortest median duration (14,3 hours) to longest median duration (23,3 hours). Based on this data, all distribution centres are colour labelled. The shortest median duration has been given the colour green in HEX format, the longest duration has been given the colour red in HEX format. All colours between these two received a colour somewhere in between according to their ranking, using a gradient going from green to red.

Table 15: Median duration time per distribution centre

Distribution centre ID	Median duration of total process in hours	HEX colour
3	14,3	00FF00 (green)
9	15,1	22FF00
16	15,1	11FF00
1	15,8	66FF00
7	15,8	44FF00
6	15,9	88FF00
2	16	CCFF00
17	16	AAFF00
8	16,8	EEFF00
15	17,6	FFEE00
18	17,6	FFFF00
5	17,9	FFCC00
4	18,6	FFAA00
14	18,7	FF8800
10	19,1	FF6600
11	19,6	FF4400
13	19,7	FF2200
12	23,3	FF0000 (red)

9.3 Runtime Enterprise Architecture visualization

Combining the chord diagram with the sunburst parts and additional information presents a complete Runtime Enterprise Architecture. This REA visualizes the process, the flows within the process, it compares the performance of multiple nodes on median duration, and it compares sub processes on size. The mock up below is created with real data, as described in 9.2.2. The REA visualizes all the essential process information in one overview.

The internal chord diagram shows the flows of parcels between distribution centres. The size of the flows represents the relative volume of parcels that is part of that flow. All the flows together account for 100% of the parcels that is distributed. The outer ring of sunbursts shows a variety of process information. Each sunburst has three main elements and an additional ring. The three elements represent the three sub process B, J, and I. The colour of these elements represent the median duration time of each sub process within each distribution centre. All these values are relative to each other (as represented in Table 15). The outer ring of the sunburst represents the median duration time of the entire distribution centre. Distribution centre number three has the lowest median duration time, of approximately 14,3 hours. Therefore, its colour is rather fluorescent green. Distribution centre 12 has a median duration time of 23,3 hours, which is the highest number of all distribution centres. Therefore, it is coloured red.

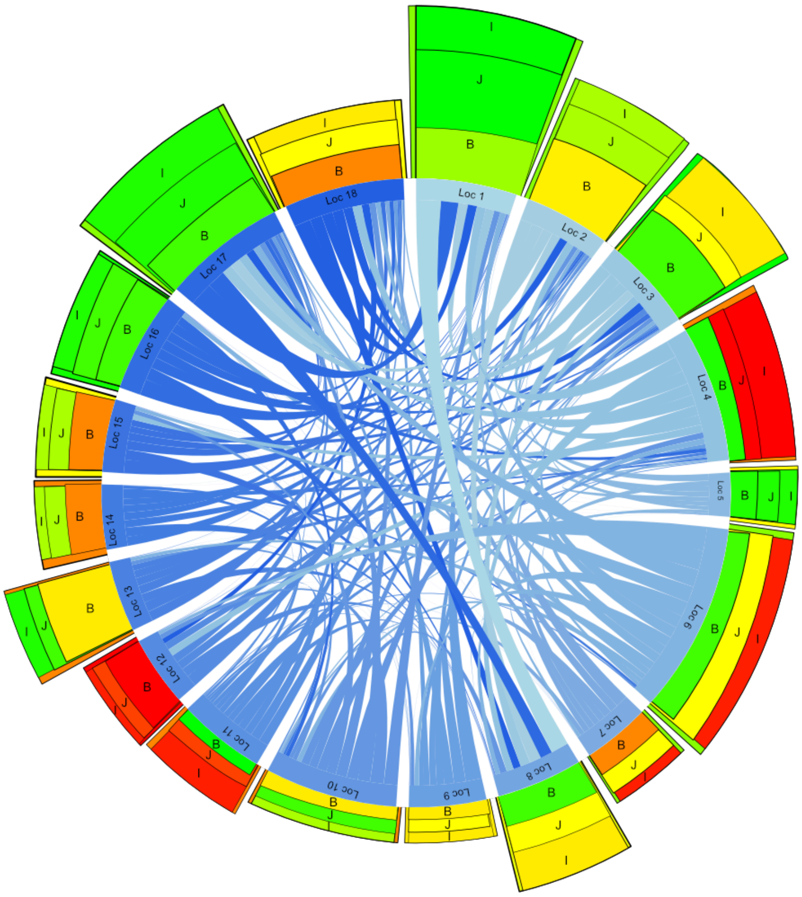


Figure 50: Runtime Enterprise Architecture visualization

When analysing further, it can be observed that location 6 sends out many more parcels than that it handles, as its size is large, while its length is relatively low. The location of this distribution centre is such that many parcels arrive here from outside the Netherlands, and enter the network here. Location 1 handles most parcels, as it has the largest length. Another observation in Figure 50 is that at 5 centres (i.e., 28%) the delivery process have a longer than average median duration (red), and that at 3 centres (i.e., 17%) the process takes shorter than average. Only at one location, the sorting to shifts process (B) takes much longer than average.

It is also possible to create this REA with different data (Figure 51). Instead of using the median duration times to colour the visualization, the fitness values as presented in chapter 8 can be used. Analysing this version of the REA it can be observed that all centres conform the B process, whereas the other two processes have more deviations. Only location 1 and 4 show outlying fitness values for the I and J processes.

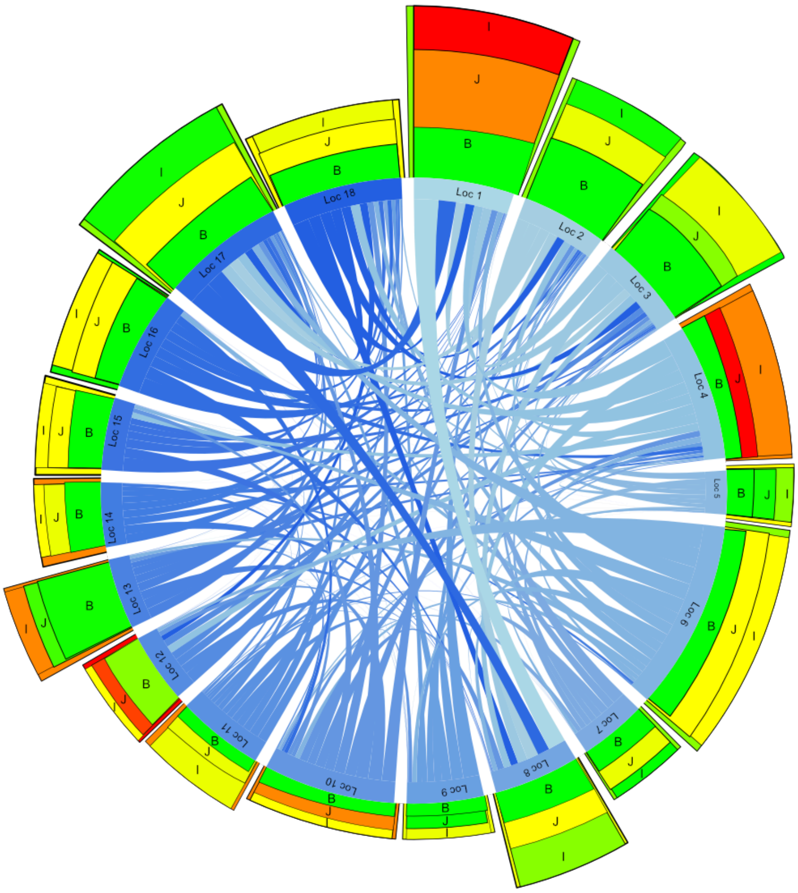


Figure 51: Runtime Enterprise Architecture visualization with fitness data

The majority of work on this Runtime Enterprise Architecture visualization above is still done by hand. Further work that is possible to do is to automate the conversion of the data into a format that fits the backed of the visualization. Additionally, some kind of interactive version should be created. This should make it possible to change for example the date that the visualization represents. By doing this, it is possible to compare process execution in an easy and understandable way.

9.4 Validation

As stated in chapter 2. Research Approach, the created design is validated. This is done by investigating how to apply or adapt the solution in such a way that it can be used to solve the problem as defined in the problem statement. The solution is valid if it brings the organization closer to their goals. The validation can be done by asking the three validation questions as mentioned by (Wieringa, 2009):

1. Would this design, implemented in this problem context, satisfy the criteria identified in the problem investigation?
2. How would slightly different designs, implemented in this context, satisfy the criteria?
3. Would this design, implemented in slightly different contexts, also satisfy the criteria?

These question were used to guide the validation session which was held with PostNL. The goal of the validation was to identify the perceived usefulness of the design, possible improvements, and additional feedback. The goal of the validation was made clear to the two stakeholders, being one process manager and one process data analyst. The technical details of the design were discussed, as well as the possible business applications of the solution. Additionally, the feasibility within PostNL was part of the discussion.

The validation started with discussion the standard process mining models, which are also presented in chapter 8. Analysis & Results. The stakeholders mentioned that these models were *"interesting to see, but are not really useful in day to day practice"*. There is no real chance for *active management* using only the petri nets. However, they could be useful to get some *"low level process insight"*.

The Runtime Enterprise Architecture visualization was perceived very useful. The idea and the ability to present real time process data to the "Operations Management Division" of the organization was met with an enthusiastic response. This would lead to more proactive management of the distribution process throughout each day. Currently, most of the improvements actions that are undertaken are based on negative outliers of the past week. Using the REA on a real time basis would enable Operations management to proactively correct problems that occur in a distribution centre.

Additionally, the REA could be used to bridge the gap between Operations and the organizations management. The management is not aware and not interested in low level process information. Their main objective is to fulfil the KPI's that were set for a certain period. Using the REA, it is possible to convert low level process data to high level process information. This information subsequently can be used to actively used when discussion the current execution and performance of the entire organization.

Concluding, the REA was received positively. Some remarks were made how the REA could be adapted to better fit the organization, but this is a matter of implementation that varies per organization. The stakeholders mentioned that "in todays world the trick is to create something that can covert the abundance of data into information, so it can be used by someone who is not familiar with the data. This is what the Runtime Enterprise Architecture does.".

9.5 Conclusion

The Runtime Enterprise Architecture visualization presents a simple, well designed visualization of a complex structure of processes and their performance. Instead of using multiple process models to illustrate how the process works, all these models are combined into one representation. Besides illustrating how these processes work, the Runtime Enterprise Architecture also shows the performance of these processes by using colours. The combination of the Chord diagram in the centre and the Sunburst parts on the outer ring is unique and an innovative way of showing process executions.

Sub question 6: "*Which steps are necessary to create a dynamic visualization of a process architecture using process models?*", can now be answered. At first, an idea is needed how to visualize the process architecture in an attractive way. This can depend on the data, however, often process execution logs of real life processes covering multiple instances consist of a number of nodes and their interactions. The created visualization consisting of a chord diagram and the additional sunbursts represents these nodes and interactions. Additionally, it shows performance indicators using several colours, ranging from median duration time to fitness. It is necessary to convert the process execution logs that are subject of the visualization to a format that can be read by the backed of the visualization. Future work is needed to automate this process, making it easier to create visualizations on a daily or even real time basis.

10. Conclusion

Based on the problem statement that was posed in chapter 1 and the main research question that was formulated based on this problem statement, the main research of this study focussed on the use of process mining to identify static process architectures. These architectures show the relations between high level organizational processes and low level departmental processes, and give insight in process execution and possible improvements. Using these static process architectures, it was possible to create a Runtime Enterprise Architecture visualization, by coming the static architectures and innovative visualizations. The main research question was formulated as follows:

How can artefact centric process mining be used to create (dynamic) process architectures?

To answer this main research question, sub-questions were formulated that individually address parts of the main research question. Together these sub questions paved the way for the answer on the main research question and the accompanying process architecture visualization. Each sub question is answered in the chapter were it was addressed, but the answers are summarized here.

10.1 Answering the sub questions

10.1.1. What is the current state of the business process management field and how is it linked to process architectures?

Business process management focusses on the management and improvement of operational processes. This is done by analysing business processes that are modelled in a specific notation, for example the Business Process Modelling Notation, and suggesting improvements based on required performance measures (van der Aalst, 2004, 2011; Weske, 2012). The business process management lifecycle consists of six steps, known as identification, discovery, analysis, redesign, implementation, and monitoring and controlling. Process architectures are part of the business process management lifecycle, and are defined as an organized overview of business processes with relations and guidelines that determine how they must be organized. Several approaches for creating process architectures exist, known as goal based, action based, object based, reference model based, and function based. In a study that evaluated these approaches using process managers, the reference model approach appeared to be the easiest to use, most useful and most popular. This approach uses a reference model to create a new process architecture.

The first sub question of this research is "*What is the current state of the business process management field and how is it linked to process architectures?*". Business process management is actively used by organizations as well as by the scientific world. Business process management uses the business process management lifecycle and the accompanying business process modelling and notation. The field is continuously extended with new ways to manage and improve business processes and helps organizations in controlling their business. Process architectures are part of the business process management lifecycle and are defined as "an organized overview of business processes with relations and guidelines that determine how they must be organized" (Dijkman, Vanderfeesten, & Reijers, 2011). Several studies address the use of process architectures to create a consistent and integrated collection of process models, such as (Green & Ould, 2004, 2005; Joosten, 2000; Koliadis, 2008; Scheer & Nüttgens, 2000).

10.2 How can process mining be used to create process models that focus on a specific part of the organizational business process?

Multiple types of process mining exist, for various purposes. Discovery purely focusses on the discovery of a process model; conformance checks a discovered process model with an existing process model; and enhancement focusses on the improvement and extension of an existing process model using an actual discovered process model. There are several perspectives on process mining which are defined as the organizational perspective, the process perspective, and the case perspective. Each perspective has its own view on process mining and focusses on different aspects. An additional way of process mining is the artefact-centric approach, which focusses entirely on data objects and the manipulations of the data objects. An artefact-centric process model describes a process as multiple collaborative artefacts, each with its own life cycle and interactions. Using process mining, configurable process models that include a set of process variants can be constructed. Such a model is able to show commonalities and differences among the different variants, and can be used to execute similar business process instances in an organization. This is related to trace clustering, which can be used if mined process models are "spaghetti-like": the models have numerous paths and are hard to comprehend. Using trace clustering it is possible to identify clusters in these spaghetti like models, and these clusters can be identified as process variants.

To answer to the second sub question of this research, "How can process mining be used to create process models that focus on a specific part of the organizational business process?", it can be concluded that Business process mining is a technique that can be used to generate process models based on event logs. An event log is a list of events, recorded by an information system. Each event should refer to a specific case to enable process mining.

10.3 How can business process management and process mining techniques be combined to improve business processes on organizational and departmental level?

Process mining is a relatively new research discipline with many developments in the last few years. It can be situated between data mining and process mining, and is seen as a Business Intelligence (BI) technology. BI encompasses other concepts such as Business Activity Monitoring, Complex Event Processing, Corporate Performance Management, Continuous Process Improvement, Business Process Improvement, Total Quality Management, Six Sigma, and Lean. A related concept of process mining is the use of event logs for the creation of functional architecture models. Besides software architectures, this development can also be applied in physical processes, which makes it possible to visualize functional architectures of real life processes. The same yields for social network analysis: although it is often used to visualize networks of social actors in a certain context, it is also possible to show the relations and interactions of different kinds of nodes in different kinds of networks. Combining these techniques should create a new kind of interactive visualization that is able to present process information in innovative ways. To answer sub question 3; "How can business process management and process mining techniques be combined to improve business processes on organizational and departmental level?", it can be stated that it is necessary to confirm the existing business process models using process mining. By combining the analytical elements and the business process management principles it is possible to understand process hierarchies in complex processes. This is necessary to create a process architecture visualization that gives innovative insights in the process execution, which can be used for improvements.

10.4 What are the necessary tools, techniques, and methods to extract departmental and organizational processes from execution data?

To answer sub question 4, "what are the necessary tools, techniques, and methods to extract departmental and organizational processes from execution data", it can be concluded that execution data often needs data preparation first, due to business rules or other restrictions. The prepared data can be analysed using Process Mining software. Using this software, it is possible to apply additional filters to the data, and to analyse the data. Using the correct plugins, it should be possible to extract departmental and organizational processes from execution data. The PM^2 process mining methodology describes the approach that is executed in the case study, and consist of the following steps: planning, extraction, data processing, mining & analysis, evaluation, process improvement & support.

10.5 Is it possible to identify and analyse the lifecycles of process artefacts and to combine departmental and organizational process models to construct process architectures?

This sub question is answered by identifying the instances of departmental processes and how they relate to each other as well as to the overall process. Process execution logs were analysed using the software tool ProM. Multiple models were created, on departmental and organizational level as well as social network variants. Some of the models show how the parcels travel throughout the distribution network. These models are artefact centric: they are constructed by analysing how the parcels move instead of creating a process models that show all the possible process routes based on the event log. Each model gives a certain view on the process, but the consistency and connection between the models is missing. The Archimate models presented later on combine most of the information presented in the created models – Archimate however simplified the complex process since the modelling language is not made to visualize Runtime Enterprise Architectures. The answer to sub question 5, "Is it possible to identify and analyse the lifecycles of the artefacts that are the subject of a business process?" is yes. In this case it is useful to use the artefact process information to create process models and eventually a process architecture – since the entire process focusses on these artefacts (parcels). The models presented are a step forward in visualizing the entire process architecture as one model, which is the final step of this study.

10.6 Which steps are necessary to create a dynamic visualization of a process architecture?

The Runtime Enterprise Architecture visualization presents a simple, well designed visualization of a complex structure of processes and their performance. Instead of using multiple process models to illustrate how the process works, all these models are combined into one representation. Besides illustrating how these processes work, the Runtime Enterprise Architecture also shows the performance of these processes by using colours. The combination of the Chord diagram in the centre and the Sunburst parts on the outer ring is unique and an innovative way of showing process executions. Sub question 6: "Which steps are necessary to create a dynamic visualization of a process architecture using process models?", can now be answered. At first, an idea is needed how to visualize the process architecture in an attractive way. This can depend on the data, however, often process execution logs of real life processes covering multiple instances consist of a number of nodes and their interactions. The created visualization consisting of a chord diagram and the additional sunbursts represents these nodes and interactions. Additionally, it shows performance indicators using several colours, ranging from median duration time to fitness. It is necessary to convert the process execution logs that are subject of the visualization to a format that can be read by the backed of the visualization. Future work is needed to automate this process, making it easier to create visualizations on a daily or even real time basis.

10.2 Answering the main research question

The main research question of this thesis was "How can artefact centric process mining be used to create process architectures?". Artefact centric process mining is able to identify process models of complex physical processes. These physical processes are part of certain hierarchy and consist of multiple nodes. To identify the hierarchy of the process structure it is necessary to create multiple models on different levels of abstraction. Eventually, several process models can be combined on several levels creating a static process architecture. Creating process architectures using reference models is a proven method, known for its reliability and ease of use. However, static process architectures are not able to cover the complexity of todays processes and are only useful for general process insight.

The static process architectures that are created using the mining and analysis activities can be extended to a new, dynamic way of process visualization. This Runtime Enterprise Architecture visualization combines multiple innovative graphical elements to represent process information as volumes, flows, median duration times, and fitness in one overview. The Runtime Enterprise Architecture visualization consists of a Chord diagram that visualizes the nodes, flows, and volumes, as well as an outer ring of sunbursts that represents unique information coming from all the original nodes. To do this, it is necessary to convert the process execution data into several formats that can be used in the visualization. Further work is needed to automate this data conversion which makes it possible to automate the process of Runtime Enterprise Architecture Visualization.

11. Discussion

Every research is subject to certain limitations in its execution. In this chapter some of these limitations are discussed using the constructs of internal validity, external validity, and reliability.

To ensure the internal validity of this research, some steps are taken to limit the chance of errors in the data. All datasets cover the exact same month and are selected using the same selection criteria. However, it is always possible that there some inconsistencies or imperfections in the extracted dataset. The data is reviewed, but it is not possible to ensure the quality of over 1,5 million entries. Besides, each dataset consists of 29 smaller datasets, that each cover one day. These daily sets are limited to 500 parcels per day, since the software of the case study organization limited us in extraction larger amounts of data.

External validity is related to generalizing the outcomes of this research to hold in similar studies. Since this entire study is built around the data of one company, there are some limitations to the generalization of the results. Several assumptions are made with the operational execution of the case study company in mind. However, the main deliverable of this study is the Runtime Enterprise Architecture. Since this REA focusses on the visualization of complex processes consisting of multiple instances and multiple (data) flows, the underlying background of the data plays no role. This means that it does not matter if the data originates from real life processes or from a system, it can be presented in the REA. Therefore, the REA can be generalized to a wider public.

At last, reliability refers to the repeatability of the study. In other words, researchers must be able to perform exactly the same experiment, under the same conditions and generate the same results. This should be the case, since the entire study is based on real data that is analysed using open source process mining software. All the data preparation steps are mentioned in this report, as well as which plugins are used to analyse this data.

References

Adriansyah, A., Munoz-Gama, J., Carmona, J., Van Dongen, B. F., & van der Aalst, W. M. P. (2013). Alignment based precision checking. *Lecture Notes in Business Information Processing*, *132 LNBIP*(1), 137–149. http://doi.org/10.1007/978-3-642-36285-9_15

Bose, R. P. J. C., & Van Der Aalst, W. M. P. (2009). Context aware trace clustering: Towards improving process mining results. In *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics* (Vol. 1, pp. 397–408).

Bose, R. P. J. C., & van der Aalst, W. M. P. (2010). Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models, 170–181. http://doi.org/10.1007/978-3-642-12186-9_16

Bostock, M. (2016). D3 Data Driven Documents. Retrieved from https://d3js.org

Bozkaya, M., Gabriels, J., & Werf, J. M. Van Der. (2009). Process Diagnostics: A Method Based on Process Mining. *2009 International Conference on Information, Process, and Knowledge Management*, 22–27. http://doi.org/10.1109/eKNOW.2009.29

Brinkkemper, S., & Pachidi, S. (2010). Functional Architecture Modeling for the Software Product Industry. *Software Architecture SE - 16*, *6285*, 198–213. http://doi.org/10.1007/978-3-642-15114-9_16

Buijs, J. C. A. M., van Dongen, B. F., & van der Aalst, W. M. P. (2013). Mining configurable process models from collections of event logs. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8094 LNCS*, 33–48. http://doi.org/10.1007/978-3-642-40176-3_5

Canbaz, S., van der Aalst, W. M. P., Fahland, D., & Weijters, A. J. M. M. (2011). *Discovering Artifact - Centric Processes 07*. Technische Universiteit Eindhoven.

Cohn, D., & Hull, R. (2009). Business artifacts: A data-centric approach to modeling business operations and processes. *IEEE Data Eng. Bull*, *32*(3), 3–9. http://doi.org/10.1007/978-3-540-88873-4_17

de Leoni, M., van der Aalst, W. M. P., & Dees, M. (2014). A General Framework for Correlating Business Process Characteristics. In *Business Process Management* (pp. 250–266). http://doi.org/10.1007/978-3-319-10172-9_16

de Medeiros, A. K. A., Guzzo, A., Greco, G., van der Aalst, W. M. P., & Weijters, A. J. M. M. (2008). Process Mining Based on Clustering : A Quest for Precision. *Business Process Management Workshops*, *4928*, 17–29. http://doi.org/10.1007/978-3-540-78238-4_4

Dijkman, R., Vanderfeesten, I., & Reijers, H. a. (2011). The Road to a Business Process Architecture: An Overview of Approaches and their Use. *Architecture*, *350*(July), 1–17. Retrieved from http://cms.ieis.tue.nl/Beta/Files/WorkingPapers/wp_350.pdf

Dumas, M., Rosa, M. La, Mendling, J., & Reijers, H. A. (2013). *Fundamentals of Business Process Management*. Retrieved from http://fundamentals-of-bpm.org/wp-content/uploads/2013/03/errata_Dumas.pdf

Fahland, D., de Leoni, M., van Dongen, B. F., & van der Aalst, W. M. P. (2011). Behavioral Conformance of Artifact-Centric Process Models. *Lecture Notes in Business Information Processing*, *87 LNBIP*, 37–49. http://doi.org/10.1007/978-3-642-21863-7_4

Fahland, D., De Leoni, M., Van Dongen, B. F., & Van Der Aalst, W. M. P. (2011). Conformance checking of interacting processes with overlapping instances. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6896 LNCS*, 345–361. http://doi.org/10.1007/978-3-642-23059-2_26

Fettke, P., Loos, P., & Zwicker, J. (2006). Business Process Reference Models: Survey and Classification. *Business Process Management Workshops : BPM 2005 International Workshops, BPI, BPD, ENEI, BPRM, WSCOBPM, BPS, Nancy, France, September 5, 2005. Revised Selected Papers*, *3812*, 469–483. http://doi.org/10.1007/11678564_44

Girault, C., & Valk, R. (2001). Petri Nets for Systems Engineering, 607. Retrieved from http://www.amazon.com/Petri-Systems-Engineering-Claude-Girault/dp/3540412174\nhttp://lossofgenerality.devzing.com/files/petribook.pdf

Green, S., & Ould, M. (2004). The Primacy of Process Architecture. *CAiSE Workshops (2)*, 154–159.

Green, S., & Ould, M. (2005). A framework for classifying and evaluating process architecture methods. *Software Process Improvement and Practice*, *10*(4), 415–425. http://doi.org/10.1002/spip.244

Hee, K. M. Van, Sidorova, N., & van der Werf, J. M. E. M. (2013). LNCS 7480 - Business Process Modeling Using Petri Nets, (257129), 116–161.

Hevner, A. R. (2007). A Three Cycle View of Design Science Research A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2), 87–92.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. http://doi.org/10.2307/25148625

Indulska, M., Recker, J., Rosemann, M., & Green, P. (2009). Business Process Modeling: Current Issues and

Future Challenges. *Advanced Information Systems Engineering. 21st International Conference, CAiSE 2009, Amsterdam, The Netherlands, June 8-12, 2009. Proceedings.*, *5565*(June), 501–514. http://doi.org/10.1007/978-3-642-02144-2

Joosten, S. (2000). Why modellers wreck workflow innovations. *Business Process Management*, 289–300. Retrieved from http://link.springer.com/chapter/10.1007/3-540-45594-9_18

Kim, Y., Choi, T. Y., Yan, T., & Dooley, K. (2011). Structural investigation of supply networks: A social network analysis approach. *Journal of Operations Management*, *29*(3), 194–211. http://doi.org/10.1016/j.jom.2010.11.001

Koliadis, G. (2008). Towards an Enterprise Business Process Architecture standard Towards an Enterprise Business Process Architecture standard, 239–246.

Krafcik, J. F. (1988). Triumph of the lean production system. *Sloan Management Review*, *30*(1), 41. http://doi.org/10.1108/01443570911005992

Lankhorst, M. (2009). *Enterprise Architecture at Work: Modelling, Communication and Analysis (The Enterprise Engineering Series)*. BOOK, Springer. Retrieved from citeulike-article-id:9666817

Ley, M., Herbstritt, M., & Ackermann, M. (2015). dblp: computer science bibliography. Retrieved January 26, 2016, from http://dblp.uni-trier.de/

Lu, X. (2013). *Artifact-Centric Log Extraction and Process Discovery.*

M. Bastian, S. Heymann, M. J. et al. (2009). Gephi: an open source software for exploring and manipulating networks. *Proceedings of International AAAI Conference on Web and Social Media*, 361–362.

Manuel, A. (2013). Social Network Analysis.

Murata, T. (1989). Petri Nets: Properties, Analysis and Applications. *Proceedings of the IEEE*, *77*(4), 541–580. http://doi.org/10.1109/5.24143

Naylor, J. Ben, Naim, M., & Berry, D. (1999). Leagility: integrating the lean and agile manufacturing in the total supply chain. *International Journal of Production Economics*, *62*, 107–118.

Nooijen, E. H. J., van Dongen, B. F., & Fahland, D. (2013). Automatic Discovery of Data-Centric and Artifact-Centric Processes, 316–327.

Otte, E., & Rousseau, R. (2002). Social Network Analysis: A Powerful Strategy, Also for the Information Sciences. *Journal of Information Science*, *28*(6), 441–453. http://doi.org/10.1177/016555150202800601

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems (JMIS)*, *24*(3), 45–77.

Popova, V., Fahland, D., & Dumas, M. (2013). Artifact Lifecycle Discovery. *arXiv Preprint arXiv:1303.2554*, 1–27. Retrieved from http://arxiv.org/abs/1303.2554

PostNL. (2014). Jaarverslag 2014.

Pyzdek, T. (2003). *The Six Sigma handbook*. *Search*. http://doi.org/10.1036/0071415963

Reijers, H. A., Dumas, M., van der Aalst, W. M. P., & ter Hofstede, A. (2005). Process-Aware Information Systems. *Wiley Online Library*, 409. http://doi.org/10.1002/0471741442.ch10

Rozinat, A., & Gunther, C. (2016). Process Mining and Automated Process Discovery Software for Professionals - Fluxicon Disco. Retrieved February 10, 2016, from http://fluxicon.com/disco/

Scheer, A.-W., & Nüttgens, M. (2000). ARIS Architecture and Reference Models for Business Process Management. *Business Process Management: Models, Techniques, and Empirical Studies*, *1806*, 376–389. http://doi.org/10.1007/3-540-45594-9_24

Schunselaar, D. M. M., Verbeek, E., van der Aalst, W. M. P., & Reijers, H. A. (2012). Creating Sound and Reversible Configurable Process Models Using CoSeNets. *Business Information Systems SE - 3*, *117*, 24–35. http://doi.org/10.1007/978-3-642-30359-3_3

Serrat, O. (2009). Solutions, (February).

Song, M., Günther, C. W., & van der Aalst, W. M. P. (2009). Trace clustering in process mining. *Lecture Notes in Business Information Processing*, *17 LNBIP*, 109–120. http://doi.org/10.1007/978-3-642-00328-8_11

van der Aalst, W. M. P. (2004). Business process management demystified: A tutorial on models, systems and standards for workflow management. *Lectures on Concurrency and Petri Nets*, *3098*(3098), 1–65. http://doi.org/10.1007/b98282

van der Aalst, W. M. P. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. *Media* (Vol. 136). http://doi.org/10.1007/978-3-642-19345-3

van der Aalst, W. M. P. (2016). ProM: Process Mining. Retrieved February 10, 2016, from http://www.processmining.org/

van der Aalst, W. M. P., Andriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., & Buijs, J. C. A. M. (2012). Process Mining Manifesto. *BPM 2011 Workshops Proceedings*, 169–194.

Van Der Aalst, W. M. P., Reijers, H. A., & Song, M. (2005). Discovering social networks from event logs. *Computer Supported Cooperative Work*, *14*(6), 549–593. http://doi.org/10.1007/s10606-005-9005-9

van der Aalst, W. M. P., Reijers, H. A., Weijters, A. J. M. M., van Dongen, B. F., Alves de Medeiros, A. K., Song, M., & Verbeek, H. M. W. (2007). Business process mining: An industrial application. *Information Systems*, *32*(5), 713–732. http://doi.org/10.1016/j.is.2006.05.003

van der Werf, J. M. E. M. (2015). Design science research. Presentation. Retrieved from https://sites.google.com/site/advresearchmethods/schedule

van der Werf, J. M. E. M., & Kaats, E. (2015). Discovery of Functional Architectures From Event Logs. *International Workshop on Petri Nets and Software Engineering (PNSE'15)*, 227–243.

van der Werf, J. M. E. M., & Verbeek, H. M. W. (2015). Online Compliance Monitoring of Service Landscapes. *Business Process Management Workshops*, *202*, 89–95. http://doi.org/10.1007/978-3-319-15895-2_8

van der Werf, J. M. E. M., Verbeek, H. M. W., & van der Aalst, W. M. P. (2012). Context-aware compliance checking. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7481 LNCS*, 98–113. http://doi.org/10.1007/978-3-642-32885-5_7

van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., & Weijters, A. J. M. M. (2005). The ProM Framework: A New Era in Process Mining Tool Support. *In Applications and Theory of Petri Nets 2005*, (i), 444–454. http://doi.org/10.1007/11494744_25

van Eck, M. L., Lu, X., Leemans, S. J. J., & van der Aalst, W. M. P. (2015). PM2 : A Process Mining Project Methodology, 297–313. http://doi.org/10.1007/978-3-319-19069-3_19

van Strien, P. J. (1997). Towards a Methodology of Psychological Practice: The Regulative Cycle. *Theory & Psychology*, *7*(5), 683–700. http://doi.org/10.1177/0959354397075006

Verschuren, P., & Doorewaard, H. (2010). *Designing a Research Project: Project Design*. *Designing a Research Project*. http://doi.org/10.1007/s13398-014-0173-7.2

Ward, J., & Peppard, J. (2002). *Strategic planning for information systems*. *Wiley series in information systems*. http://doi.org/10.1016/0024-6301(90)90122-K

Weske, M. (2012). *Business Process Management: Concepts, Languages, Architectures*. *Business Process Management* (Vol.

54). http://doi.org/10.1007/978-3-540-73522-9

Wieringa, R. J. (2006). Requirements Engineering : Frameworks for Understanding. *Development*.

Wieringa, R. J. (2009). Design Science as Nested Problem Solving.

Wohlin, C. (2014). Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In *18th International Conference on Evaluation and Assessment in Software Engineering (EASE 2014)* (pp. 1–10). http://doi.org/10.1145/2601248.2601268

Yin, R. K. (2009). *Case Study Research: Design and Methods. Essential guide to qualitative methods in organizational research* (Vol. 5). http://doi.org/10.1097/FCH.0b013e31822dda9e

Zachman, J. A. (1987). A framework for information systems architecture. *IBM Systems Journal*. http://doi.org/10.1147/sj.263.0276

Uncovering the Runtime Enterprise Architecture of a Large Distributed Organisation
A Process Mining-Oriented Approach

Robert van Langerak, Jan Martijn E.M. van der Werf, and Sjaak Brinkkemper

Department of Information and Computing Science
Utrecht University
{r.p.vanlangerak, j.m.e.m.vanderwerf, s.brinkkemper}@uu.nl

Abstract. Process mining mainly focuses on analyzing a single process that runs through an organization. However, often organisations consist of multiple departments that need to work together to deliver a process. ArchiMate introduced the Business Process Cooperation Viewpoint for this. However, such models tend to focus on modeling design time, and not the runtime behavior. Additionally, many approaches exist to analyze multiple departments in isolation, or the social network they form, but the cooperation between processes received little attention. In this paper we take a different approach by analyzing the runtime execution data to create a new visualization technique to uncover cooperation between departments by means of the Runtime Enterprise Architecture using process mining techniques. By means of a real-life case study at a large logistic organization, we apply the presented approach.

Keywords: Process mining, enterprise architecture, data analytics, business analytics, runtime enterprise architecture

1 Introduction

In larger organisations, departments work jointly to deliver the services or products of that organisation. Capturing this cooperation is part of the domain of Enterprise Architecture (EA) [28]. EA consists of principles, methods, and models to design and realize an enterprises organisational structure, business processes, information systems, and infrastructure. An important aspect within organisations is the cooperation between different departments in the overall processes of the organisation. ArchiMate [13] introduced the Business Process Co-operation Viewpoint (BPC) to model this explicitly. These models mainly focus on the design of cooperation: which processes and departments within an organisation are allowed to communicate. The runtime behavior, i.e., whether and when communication occurs, and the possible execution orders are typically left out.

Process mining [2] offers many opportunities to assist the enterprise architect in uncovering the runtime behavior of their EA. In process mining, many algorithms exist to discover processes (e.g. [5, 8, 10, 16]), to check for conformance (e.g. [3, 7, 21]), and to enhance process models (e.g. [11, 24]). Although the process is viewed from different perspectives [17], such as the case, process and resource perspective, the organisational perspective [22] has been given little attention. Approaches like PM^2 [23] and Process Diagnostics [6] focus on the overall process within an organisation, rather than focusing on how the different departments within the organisation contribute to deliver its service. Consequently, for larger organisations where multiple departments cooperate to deliver their services, current process mining techniques are hard to apply.

In this paper, we want to close the gap between the static descriptions created in EA and the runtime environment in which all these processes have been implemented.

As a running example, consider the insurance company InsComp, with three departments: the Policy Department, the Claim Department and the Financial Administration. InsComp delivers two services: the issuing of policies and the handling of claims, where the former is the responsibility of the Policy Department, and the latter of the Claim Department. The Claim Department sometimes asks the Policy Department to check a policy. Once a claim is approved, the Financial Administration is instructed to compensate the claim. As InsComp has a problem with the Claim Handling service, they want to obtain insights into the cooperation and functioning of the different departments.

To analyze such questions about the quality of the actual different departments and their cooperation, we would like to apply process mining techniques. We do this by introducing the Runtime Enterprise Architecture (REA) of an organisation, which uses the runtime operation data from the processes operated within the organisation. This allows us to create new visualizations to uncover the involvement of departments, their cooperation, and their relative achievements in the process.

In the remainder of this paper we make the following contributions:

– Incorporation of the runtime behavior of an organisation into the Business Process Co-operation Viewpoint of Enterprise Architecture (Sec. 2);
– Visualisation techniques to uncover the Runtime Business Process Co-operation View of an organisation (Sec. 3); and
– Showing the applicability and possibilities of the techniques through a case study in a large parcel distributor in the Netherlands (Sec. 4).

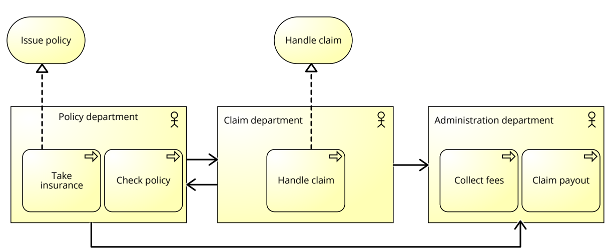


Fig. 1. Business Process Co-operation Viewpoint

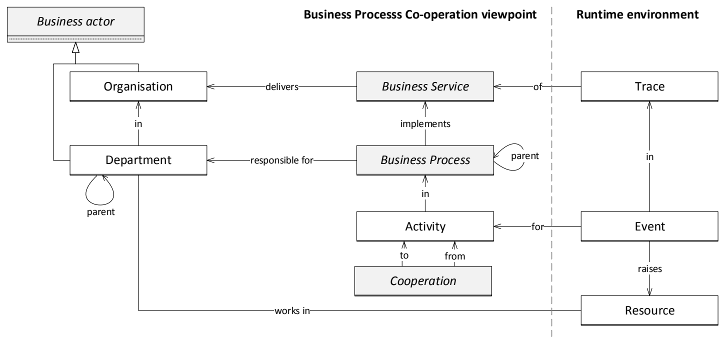


Fig. 2. Conceptual Model of Runtime Business Architectures

2 Runtime Enterprise Architectures

To model the different departments within an organisation and how these cooperate to deliver the services of an organisation, ArchiMate 3.0 [13] introduced the Business Process Co-operation Viewpoint (BPC) [15]. This viewpoint shows the relation between the business processes and their surroundings, and can be used to create a high-level design of business processes within their context and to provide insight into their dependencies [15]. The BPC viewpoint of our example organisation InsComp is shown in Fig. 1.

The BPC viewpoint reflects the allowed cooperations at design time. Whether in real life this blue print is always followed is a complete different question. With the logging capabilities of current Process-aware Information Systems (PAISs), we are able to record reality in the form of audit trails or event logs [25]. These event logs are input for process mining.

Key in process mining is that each event is related to a process instance of some businesss process. Within a large organisation with many different business processes, it is difficult to relate each process instance to a business process or business service. For this we define the *Runtime Enterprise Architecture* (REA) as the set of structures and metrics to capture and analyze the runtime behavior of that organisation based on its Enterprise Architecture. In this paper, we focus on the dynamic behavior of the BPC viewpoint, the Runtime Business Process Co-operation View (RBPC).

2.1 Meta-Model of the Runtime Business Process Co-operation View

The conceptual model that maps the relevant event log concepts to the concepts of the BPC viewpoint is shown in Fig. 2. On the left, the relevant elements of the BPC viewpoint are depicted. The gray elements are the default elements of ArchiMate. *Organisation* and *Department* are specializations of the ArchiMate element *Business Actor*. An *Organisation* has a hierarchical structure of *Departments*, and delivers some *Business Service*. A *Business Service* is implemented by one or more *Business Processes*. *Activities* in a *Business Process* form *Cooperation*. A *Cooperation* is always initiated by an *Activity* (relation *from*) and concluded by an *Activity* (relation *to*).

At runtime, *Business Services* are instantiated, resulting in *Traces*, that flow through the organisation. For a *Trace*, *Events* are raised by executing *Activities*. Possibly, the *Resource* is recorded as well. Notice that in many organisations, traces are identified by some global identifier that is used throughout the business service or organisation.

2.2 Runtime Business Process Co-operation View

At design time, different cooperations can be modelled in the BPC view. Let c be a cooperation. From the conceptual model, we can derive the following types of cooperations.

Intra-process A cooperation occurs within the same process, i.e., $in(from(c)) = in(to(c))$;

Inter-process A cooperation occurs between two different processes, i.e., $in(from(c)) \neq in(to(c))$;

Intra-departmental A cooperation within the same department (possibly between different processes), i.e., $responsible\_for(in(from(c))) = responsible\_for(in(to(c)))$

Inter-departmental A cooperation between different departments, i.e., $responsible\_for(in(from(c))) \neq responsible\_for(in(to(c)))$

Notice that a cooperation can be both inter-process and intra-deparmental at the same time, if the cooperation is between two business processes for which the same department is responsible.

3 Uncovering Cooperations

In order to obtain insight in the cooperations within an organisation, we first discuss how to discover cooperations. Next, we present a new visualization technique for cooperations, the Runtime Business Process Co-operation View that visualizes the runtime behavior of an organisation, rather than only focusing on the design time, as is current practice in EA.

3.1 Discovering Cooperations

Several techniques have been proposed in process mining to analyze both inter and intra organisations [1], such as social network analysis [22], artifact-centric techniques [20] and feature discovery [27]. Social Network Analysis (SNA) focuses on identifying nodes and their relationships [19]. A social network consists of nodes and a set of relationships or links. In [22], the authors use event logs to generate a social network of the resources within the event log. In process analysis this derived SNA can be used to identify resources in a network, and to show how these resources interact. Additionally, SNA can be used to study patterns within an organisations network and enabling organisations to use these patterns to create competitive advantages [14].

Whereas process mining relies on the assumption that each process instance belongs to the same business process, the artifact-centric approach assumes that the process instances are manipulated by artifacts [18], and tries to discover the processes and interactions of these artifacts [9, 20]. Artifact-centric mining is used to discover a process by using the artifacts that are present in the process and is therefore often used to create better process models for real life or physical processes [12].

Recent research focuses on the use of process discovery techniques to construct functional architectures [27] by relating software execution data to features. In this way, it is possible to discover the communication protocols between features from the behavioral profile [26].

Each of these process mining techniques can be used to enhance the existing BPC viewpoint by updating the *Cooperations* from the event log. Next step is to visualize and quantify the cooperations found at runtime.

3.2 Visualizing the Runtime Business Process Cooperation

At runtime, many different metrics are available about the business processes and their cooperations. The current visualization of the BPC viewpoint in ArchiMate only focuses on depicting the EA at design time. Consequently, we require new visualizations to provide useful insights in the organisation. For this, we introduce the RBPC, which is an interactive representation of process execution data. As an example, the RBPC of InsComp is depicted in Fig. 3. The view combines the chord diagram technique with a sunburst visualization, and is designed to give a high level overview and a detailed insight of the process in one visualization at the same time.

The inner circle of the RBPC represents the cooperations between the different departments, and is visualized using a chord diagram, which allows to combine many edges into flows without creating chaos. The length of each part is determined by the centrality of the node in the social network, which is a combination of the number of cooperations each part initiated and concluded, and the size of the flows represent the volumes traveling between the nodes. The color of the flow is determined by the node that initiates most cooperations.

The outer circle of the RBPC is a sunburst that can be used to visualize different metrics, such as the number of cooperations concluded, or the number of instances processed in that department. Each sunburst contains the subdepartments and processes within those departments, enclosed by an additional ring. The size of these elements indicates the percentage of units that are handled by the process or department, relative to the others. Additionally, we can color both the ring and the inner elements with other metrics, such as the overall duration of the cooperations, or the conformance of the different processes within the department. For example, in Fig. 8, the color of the elements represents the median duration time, whereas in Fig. 9, the color of the elements represents the fitness of the process model.

For the running example InsComp, the RBPC is depicted in Fig. 3. There are three departments, Claim, Policy and Financial Administration. From the diagram, we directly see that the Financial Department department plays a larger role in the organisation. A

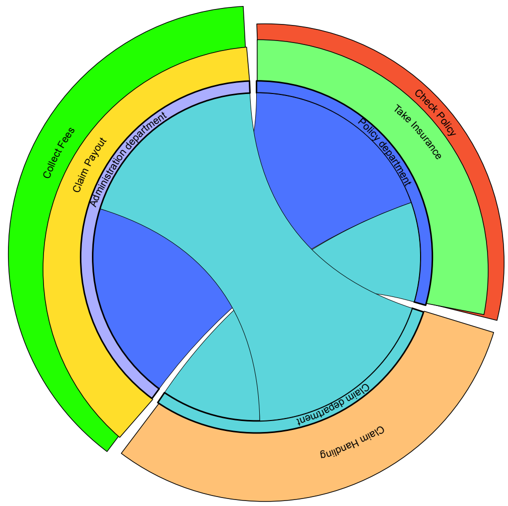


Fig. 3. A Runtime Business Process Co-operation View (RBPC) of our running example, InsComp

third of the cooperations with the Policy Department are with the Claim Department. Most cooperations of this department are with the Financial Administration.

The sunburst part focuses on the internal process execution of each department. Therefore, the number of sub processes per department were identified. For every process, the sunburst of that department is extended with an additional ring. For example, the Policy department consists of two processes, whereas the Claim Department has only one sub process. The colors of thee rings depict the total duration of the process. From this picture, we may conclude from the flows that a third of the communication of the Policy Department comes from the Claim Department, and from the coloring we may conclude that the “Check Policy” process in the Policy Department is a bottleneck, and that in the Financial Administration the duration for the “Claim Payout” process is above average, which would explain why Claim Handling service requires attention.

4 Validation of the Runtime Enterprise Architecture

We aim to validate the usage of the proposed Runtime Business Process Co-operation View with a non-trivial case study in a large logistics company. The selected case organisation is one of the largest mail and parcel distributors in the Netherlands, referred to as SendIT. In 2014, the organisation addressed 2,705 million mail items and 142 million parcels. SendIT has 18 distribution centers throughout the Netherlands for the distribution of parcels. Each center is responsible for their own part of the distribution process. Consequently, there are 18 instances of the same departmental processes. Each center has its own facilities to record the process execution data. Each of the instances can be analyzed and compared using this data. Currently, the organisation lacks proper visualizations of the performance of the different centers, and their cooperation in the different processes. In this case study, we use the data to compare the different distribution centers on process execution and performance, and to discover the different cooperations between departmental processes.

4.1 Distribution Process and Scan Trails

Each distribution center is responsible for three main processes: sorting shifts (B), sorting routes (J), and delivery (I). Each center represents a specific area for delivery. In different shifts, the centers sort the received parcels based on the postal code. Parcels that require transporting to a different area are each night transported to the respective centers that are responsible for that area. In the morning, each shift is sorted into multiple routes based on the scanned postal codes, for multiple delivery mans. When the parcels are divided over the delivery mans, they will deliver the parcels to the clients.

The parcels are scanned for each step in the processes. Every time a parcel is scanned, a new scanvalue is added to its respective barcode. A scanvalue has its own definition and possible consequences for the further distribution of this parcel. The first scan every parcel receives when entering a distribution center is the scanvalue B1: ‘Proof of Acceptance’. After this first scan, the parcel is placed onto a large conveyor, that sorts the parcels into different containers based on the address on the parcel. All parcels that have to be distributed in the same area are part of a certain shift. In the end all the parcels are divided over the correct areas, and thus over the correct shifts.

In the morning, the shifts are further divided into specific routes. Shifts are sorted consecutively. First, all parcels for a shift are placed on the conveyor. The parcels are scanned again (scanvalue J1: ‘Parcel sorted’). When a parcel is at the conveyor for its route, the sorter shoves it down. The parcel is then assigned the value J4: ‘Parcel sorted on route’, and placed in the van for that particular area. Once all the parcels for that shift are sorted, the delivery men are notified and can start the actual delivery. The system assigns after each shift the scanvalue J5 ‘out for delivery’ to all parcels handled in that shift. This process is repeated until all the shifts are sorted.

Each delivery man has a hand scanner, used upon delivering the parcel. Based on whether and where the parcel is delivered, different scanvalues are added to the parcel. For example, if the parcel is delivered at home, value I1, ‘delivered’ is used, whereas if the parcel is delivered at the neighbors, value I10 is used. If the parcel cannot be delivered, the delivery man assigns scanvalue J8 to the parcel. After returning at the

Table 1. Scan trail of a single package representing the most frequent happy flow

Barcode	Date	Time	Value	Reason	Description
1B1671337	10/02/2016	00:49:25	B	1	Proof of acceptance
1B1671337	10/02/2016	09:34:25	J	1	sorted
1B1671337	10/02/2016	09:34:26	J	40	sorted on route
1B1671337	10/02/2016	10:06:03	J	5	out for delivery
1B1671337	10/02/2016	12:28:45	I	1	Delivered

Table 2. Events per distribution center (DC). In total, the dataset contains 1.555.492 events divided over 136.575 scan trails

DC	# events
1	111.001
2	84.309
3	85.935
4	103.887
5	72.345
6	90.116

DC	# events
7	81.362
8	124.722
9	66.348
10	79.234
11	74.715
12	83.476

DC	# events
13	85.049
14	87.188
15	80.159
16	77.148
17	88.714
18	79.784

Table 3. Structure of an event in the event logs after conversion

Attribute	Example
Barcode (ID)	1B1671337
Timestamp	10/02/2016 00:49:25
Scan letter	B
Scan number	01
Combinedscan	B01
Location	LOC 3

center, all parcels that could not be delivered are scanned and added to the evening sorting process, so that they can be delivered the next day.

The scan trail of a parcel is a sequence of all its scanvalues and their occurence. Each scanvalue always consists of a character and a number, together with a timestamp. Several happy flows exist for these trails: flows where nothing went wrong and the parcel was delivered. The most frequent happy flow is depicted in Tbl. 1. For each parcel a trail can be exported from the PAISs at SendIT.

4.2 Data Selection and Extraction

SendIT handles roughly 30.000 parcels per distribution center per day, resulting in approxmatly 160M events per month. Consequently, we had no option than to take a random sample from this data set. The data selected for this case study covers the whole month February in 2016. For each center, a dataset was created with at most 500 parcels per day. For these parcels, the scan trails were extracted and combined into a large dataset for a distribution center. The total number of events per center is depicted in Tbl. 2. As a last step, all datasets were combined into a single dataset for analysis. This resulting dataset contains 136.575 scan trails.

Each dataset had to be prepared before it can be analyzed. An excerpt of the trail is depicted in Tbl. 1. For example, the date and time values had to be merged into a single timestamp, as this is required by the different process mining tools. The Value and Reason attributes in the scan trail are merged to create the activity name for each event. Both values were added to the event log, to be able to analyze the event log on different levels of abstraction, as the Value and Reason represent the business process, and the corresponding activity, respectively. The location is a three letter abbreviation representing the distribution center. The final structure of the dataset is shown in Tbl. 3.

4.3 Analysis

For the analysis of the datasets, the open source software ProM [25] is used. To exclude parcels that are not yet delivered, we filtered the dataset by removing all scan trails that do not contain an activity with an I-value. Next,multiple analyses have been executed with ProM to identify the structures and flows in the dataset. To create an overview of the entire process, the organisational process is visualized first. Next, a generic departmental process model is created from the most occurring traces in the dataset. Lastly, the subprocesses of the departmental processes are identified. By combining these models using Archimate, a static enterprise architecture is created.

Organizational Model The organisational model is created using the Social network miner plugin [4]. The result is depicted in Fig. 4. It is a complete graph, i.e., all distribution centers send parcels to each other.

Departmental models As a complete process model representing all 136.575 scan trails returns a spaghetti-like model, we decided to apply Occam's razor, and filtered a dataset containing the 10 most occurring traces of the 18 distribution centers. The ten most occurring traces cover together almost 68% of all scan trails. To be able to test for conformance, we created a process model using the Inductive Visual Miner [16], and then transformed it into a petri net. The constructed petri net representing the 10 most frequent scan trails is depicted in Fig. 5. The process starts with a B1 event, then the sorting process is started (J-valued events), after which the parcel is delivered (I-valued events). From the same dataset containing the 10 most frequent scan trails, we discovered for each of the processes a separate process model. The J process is depicted in Fig. 6.

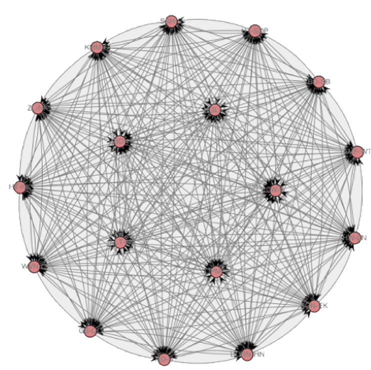


Fig. 4. Organisational model as mined with the Social Network Miner of ProM

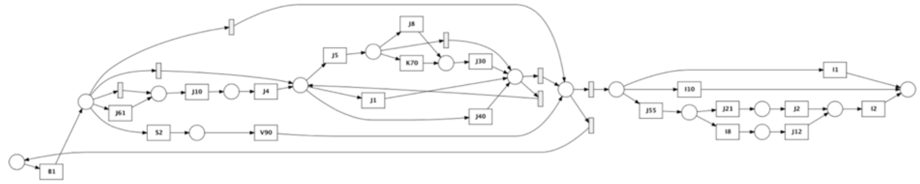


Fig. 5. 10 most occuring traces

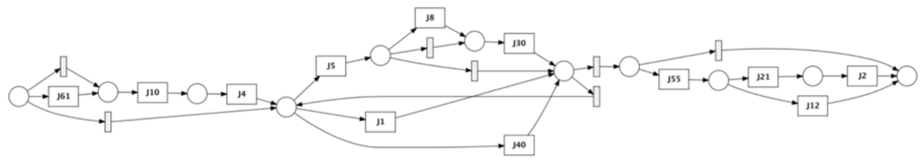


Fig. 6. Model of process J in isolation

To check the degree of conformance of the processes of the different distribution centers the logs are replayed through the petri net using the Replay for conformance checking plugin in ProM. This results in a fitness value per center, indicating how well a model represents a log. The resulting fitness values are shown in Tbl. 4. Overall, the process shows a high fitness. Additionally, we used Disco [1] to analyze the median duration time for each center. All values range between 14,3 and 23,3 hours, the average is 17,4 hours between first scan and delivery.

Enterprise architecture The Enterprise Architecture comprises the different processes, and abstracts from the detailed activities. To discover how the different processes cooperate, we decided to create an additional organisational model in which the hand-over of work is analysed between the different processes by taking the Value as resource. This resulted in the model depicted in Fig. 7(a). Based on this organisational model, it is possible to create the BPC viewpoint of the static EA of SendIT, shown in Fig. 7(b).

Runtime Business Process Co-operation View One of the main drivers of SendIT is to compare the runtime behavior of the different centers, and the amount of parcels that is transported between the centers. For this, we created two separate RBPC views for SendIT. Both RBPC views use the location changes of the parcels in the chord diagram, and the respective number of scan trails handled in the center for the length of the sunburst. For the coloring schema, the former is based on the median duration of scan trails, the latter is based on the fitness of the sub processes at each center.

To define the chord diagram of both RBPC views, we first analyzed the mined social network (Fig. 4), where the location changes have been defined as the hand-over of work

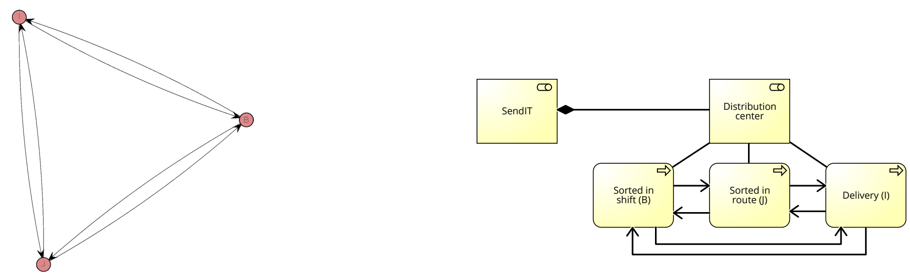
[1] https://fluxicon.com/disco/

Table 4. Fitness values and median durations in hours per distribution center (DC), calculated with the Replay for Conformance Plugin of ProM

DC	Fitness	Dur. (h)
1	0.911	15.8
2	0.956	16.0
3	0.960	14.3
4	0.916	18.6
5	0.964	17.9
6	0.870	15.9

DC	Fitness	Dur. (h)
7	0.957	15.8
8	0.943	16.8
9	0.954	15.1
10	0.941	19.1
11	0.954	19.6
12	0.955	23.3

DC	Fitness	Dur. (h)
13	0.968	19.7
14	0.954	18.7
15	0.952	17.6
16	0.949	15.1
17	0.948	16.0
18	0.949	17.6



(a) Social Network (b) Archimate BPC view

Fig. 7. The discovered BPC view of SendIT generated from the Social Network Miner of ProM

between the centers. As a first step, their respective frequencies have been analyzed. As each parcel is always at exactly one location, we counted the number of consecutive event pairs with different locations. As a next step, these numbers have been normalized using the total number of location changes. In this way, location changes become relative to each other, and all add up to 100%. Based on this data, the chord diagram is constructed.

To determine the length of each center in the RBPC views, i.e., the number of scan trails handled by the center, we analyzed for each center the ratio between parcels with a B scanvalue, i.e., the number of parcels that arrive, with the number of parcels that have an I scanvalue, i.e., the number of parcels delivered by the center. If this ratio is high, most parcels that arrive at the center are distributed in the region, i.e., the center handles many parcels, whereas if the ratio is low, most parcels are transported to different centers, thus the center handles few parcels.

Next, for each center we determine the size of the internal process by normalizing the amount of parcels each process handles with the ratio determined for that center. For example, if a center has 50% of B scanvalues, 25% of J scanvalues, and 25% of I scanvalues, parts J and I will be similar in size, and the size of B has the size of J and I combined. As each center has three processes, sorting shifts (B), sorting routes (J), and delivery (I), this results in three rings for each center in the RBPC views.

These two steps create the basis for the RBPC views. The first RBPC view uses the median duration, as depicted in Tbl. 4, for coloring its elements. The lowest median

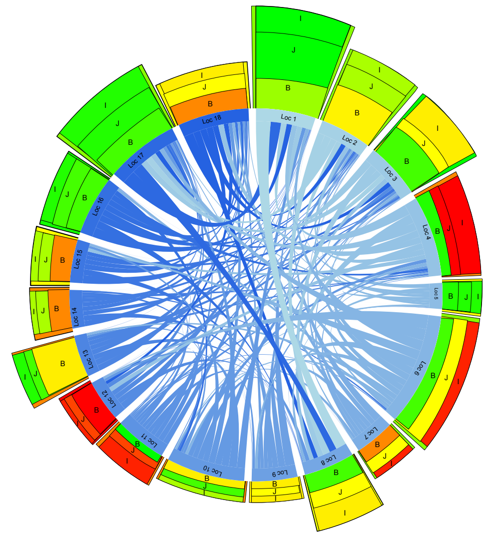


Fig. 8. RBPC view with the median duration metric

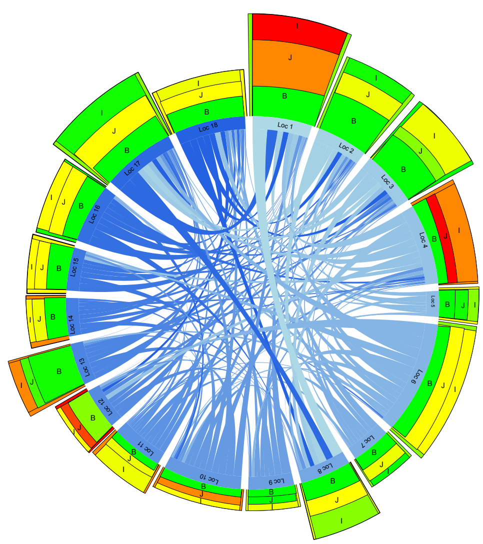


Fig. 9. RBPC view with the fitness metric

duration is colored green, the highest is colored red. Each center is assigned a gradient color relative to the higest and lowest median durations, resulting in the view depicted in Fig. 8.

Analyzing the two RBPC views, we directly observe that location 6 sends out many more parcels than that it handles, as its size is large, while its length is relatively low. Location 1 handles most parcels, as it has the largest length. Another observation in Fig. 8 is that at 5 centers (i.e., 28%) the delivery process have a longer than average median duration (red), and that at 3 centers (i.e., 17%) the process takes shorter than average. Only at one location, the sorting to shifts process (B) takes much longer than average. Similarly, analyzing Fig. 9, we observe that all centers conform the B process, whereas the other two processes have more deviations. Only location 1 and 4 show outlying fitness values for the I and J processes.

4.4 Expert Validation

To validate the results and different visualizations of the RBPC, we presented the results to two stakeholders of SendIT. The goal of the interviews was to identify the perceived usefulness of the design, and to obtain possible improvements and additional feedback. The goal of the validation was made clear to the two stakeholders, being one process manager and one process data analyst. The technical details of the design were discussed, as well as the possible business applications of the solution. Additionally, the feasibility within SendIT was part of the discussion.

The validation started with discussing the standard process mining models, which are also presented in this section. The stakeholders mentioned that these models were interesting to see, but are not really useful in day to day practice. There is no real chance for active management using only the petri nets. However, they could be useful to get some low level process insight.

The RBPC was perceived very useful. The idea and the ability to present real time process data to the Operations Management Division of the organisation was met with an enthusiastic response. This would lead to more proactive management of the distribution process throughout each day. Currently, most of the improvements actions that are undertaken are based on negative outliers of the past week. Using the RBPC on a real time basis would enable Operations management to pro-actively correct problems that occur in a distribution center.

Additionally, the RBPC could be used to bridge the gap between Operations and the organisations management. The management is not aware and not interested in low level process information. Their main objective is to fulfill the KPIs that were set for a certain period. Using the RBPC, it is possible to convert low level process data to high level process information. This information subsequently can be used to actively used when discussion the current execution and performance of the entire organisation.

Concluding, the RBPC was received positively. Some remarks were made how the RBPC could be adapted to better fit the organisation, but this is a matter of implementation that varies per organisation. The stakeholders mentioned that "in todays world the trick is to create something that can covert the abundance of data into information, so it can be used by someone who is not familiar with the data. This is what the RBPC does.".

5 Conclusions and Future work

Current Enterprise Architecture mainly focus on modeling an organisation design-time only. In this paper, we propose the Runtime Enterprise Architecture (REA) that enhances the EA of an organisation with runtime execution data. To visualize the cooperation between departments and their process within an organisation, we propose the Runtime Business Process Co-operation viewpoint that visualizes the runtime cooperation between departments and the relative volumes and quality of the different processes at the departments. The visualization combines the chord diagram for visualizing cooperations with the sunburst visualization for the volume of the processes. The coloring schema is used to depict the quality of the processes.

To illustrate the visualization, we applied it on one a large logistics organisation to analyze parcel transportation between departments. Initial validation with the organisation shows the perceived usefulness of the visualization technique. Although the proposed visualization technique itself is general, the case study organisation had no concurrency in their processes. Generalization of the validation results require further experimentation.

Many different viewpoints exist in EA modeling. In this paper we focused mainly on the Business Process Co-operation Viewpoint, but we envision the proposed techniques to be extended to different viewpoints as well. As the proof of the pudding is in the eating, we plan on fully automating the visualization technique in ProM to perform more in-depth case studies to explore further analysis and visualization possibilities of the technique.

References

1. W. M. P. van der Aalst. Intra- and inter-organizational process mining: Discovering processes within and between organizations. In *PoEM 2011*, volume 92 of *LNBIP*, pages 1–11. Springer, 2011.
2. W. M. P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.
3. W. M. P. van der Aalst, A. Adriansyah, and B. F. van Dongen. Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 2(2):182–192, 2012.
4. W. M. P. van der Aalst, H. A. Reijers, and M. Song. Discovering social networks from event logs. *Computer Supported Cooperative Work*, 14(6):549–593, 2005.
5. W. M. P. van der Aalst, A. J. M. M. Weijters, and L. Maruster. Workflow Mining: Discovering Process Models from Event Logs. *Knowledge & Data Engineering*, 16(9):1128–1142, 2004.
6. M. Bozkaya, J. Gabriels, and J. M. E. M. van der Werf. Process diagnostics: A method based on process mining. In *eKNOW 2009*, pages 22–27. IEEE Computer Society, 2009.
7. J. C. A. M. Buijs and H. A. Reijers. Comparing business process variants using models and event logs. In *EMMSAD 2014, Held at CAiSE 2014*, volume 175 of *LNBIP*, pages 154–168. Springer, 2014.
8. J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst. Discovering and navigating a collection of process models using multiple quality dimensions. In *Business Process Management Workshops*, volume 171 of *LNBIP*, pages 3–14. Springer, 2013.

9. D. Cohn and R. Hull. Business artifacts: A data-centric approach to modeling business operations and processes. *IEEE Data Eng. Bull*, 32(3):3–9, 2009.
10. B.F. van Dongen, A.K. Alves de Medeiros, and L. Wen. Process mining: Overview and outlook of petri net discovery algorithms. *Transactions on Petri Nets and Other Models of Concurrency II*, 5460:225–242, 2009.
11. D. Fahland and W. M. P. van der Aalst. Model repair - aligning process models to reality. *Inf. Syst.*, 47:220–243, 2015.
12. D. Fahland, M. de Leoni, B. F. van Dongen, and W.M.P. van der Aalst. Behavioral Conformance of Artifact-Centric Process Models. *Lecture Notes in Business Information Processing*, 87 LNBIP:37–49, 2011.
13. The Open Group. Archimate 3.0 specification, 2016. http://pubs.opengroup.org/architecture/archimate3-doc/.
14. Y. Kim, T. Y. Choi, T. Yan, and K. Dooley. Structural investigation of supply networks: A social network analysis approach. *J. Operations Management*, 29(3):194–211, 2011.
15. M. Lankhorst. *Enterprise Architecture at Work: Modeling, Communication and Analysis the Third Edition On*, volume 36. Springer, 2013.
16. S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst. Discovering block-structured process models from event logs - a constructive approach. In *PETRI NETS 2013*, volume 7927 of *LNCS*, pages 311–329. Springer, 2013.
17. F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst. Balanced multi-perspective checking of process conformance. *Computing*, 98(4):407–437, 2016.
18. E. H. J. Nooijen, B. F. van Dongen, and D. Fahland. Automatic Discovery of Data-Centric and Artifact-Centric Processes. In *Business Process Management Workshops*, LNBIP, pages 316–327. Springer, 2013.
19. E. Otte and R. Rousseau. Social Network Analysis: A Powerful Strategy, Also for the Information Sciences. *J. Information Science*, 28(6):441–453, 2002.
20. V. Popova, D. Fahland, and M. Dumas. Artifact Lifecycle Discovery. *arXiv preprint arXiv:1303.2554*, pages 1–27, 2013.
21. A. Rozinat and W. M. P. van der Aalst. Conformance checking of processes based on monitoring real behavior. *Inf. Syst.*, 33(1):64–95, 2008.
22. M. Song and W. M. P. van der Aalst. Towards comprehensive support for organizational mining. *Decision Support Systems*, 46(1):300–317, 2008.
23. M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst. PM ˆ2 : A process mining project methodology. In *CAiSE 2015*, volume 9097 of *LNCS*, pages 297–313. Springer, 2015.
24. B. Vázquez-Barreiros, S. J. van Zelst, J. C. A. M. Buijs, M. Lama, and M. Mucientes. Repairing alignments: Striking the right nerve. In *EMMSAD 2016, , Held at CAiSE 2016*, volume 248 of *LNBIP*, pages 266–281. Springer, 2016.
25. H. M. W. Verbeek, J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst. XES, XESame, and ProM 6. In *Information System Evolution*, volume 72, pages 60–75. Springer, 2011.
26. M. Weidlich and J. M. E. M. van der Werf. On profiles and footprints - relational semantics for petri nets. In *PETRI NETS 2012*, volume 7347 of *LNCS*, pages 148–167. Springer, 2012.
27. J. M. E. M. van der Werf and E. Kaats. Discovery of Functional Architectures From Event Logs. *International Workshop on Petri Nets and Software Engineering (PNSE'15)*, pages 227–243, 2015.
28. R. Winter and E. J. Sinz. Enterprise architecture. *Information Systems and e-Business Management*, 5(4):357–358, 2007.