0.0.1 A definition for Business Processes

Before elaborating on Process Mining, we have to establish the meaning of the term *process*. The term is widely-used and therefore has a rich semantic volume. A process generally refers to something that advances and changes over time[1]. Despite, legal or biological processes being valid interpretations, too, we focus on *business processes*.

An example is a loan application process in which an applicant may request a loan. The case would then be assessed and reviewed by multiple examiners and end in a final decision. The loan might end up in an approval or denial. The *business* part is misleading as these processes are not confined to commercial settings alone. For instance, a medical business process may cover a patients admission to a hospital, followed by a series of diagnostics and treatments and ending with the recovery or death of a patient. Another example from a Human Computer Interaction (HCI) perspective would be an order process for an online retail service like Amazon. The buyer might start the process by adding articles to the shopping cart and proceeding with specifying their bank account details. This order process would end with the submission or receival of the order.

All of these examples have a number of common characteristics. They have a clear starting point which is followed by numerous intermediary steps and end in one of the possible sets of outcomes. For this work, we mainly follow the understanding outlined in van der Aalst, Adriansyah, de Medeiros, Arcieri, Baier, Blickle, Bose, van den Brand, Brandtjen, Buijs, Burattin, Carmona, Castellanos, Claes, Cook, Costantini, Curbera, Damiani, de Leoni, Delias, van Dongen, Dumas, Dustdar, Fahland, Ferreira, Gaaloul, van Geffen, Goel, Günther, Guzzo, Harmon, ter Hofstede, Hoogland, Ingvaldsen, Kato, Kuhn, Kumar, La Rosa, Maggi, Malerba, Mans, Manuel, McCreesh, Mello, Mendling, Montali, Motahari-Nezhad, zur Muehlen, Munoz-Gama, Pontieri, Ribeiro, Rozinat, Seguel Pérez, Seguel Pérez, Sepúlveda, Sinur, Soffer, Song, Sperduti, Stilo, Stoel, Swenson, Talamo, Tan, Turner, Vanthienen, Varvaressos, Verbeek, Verdonk, Vigo, Wang, Weber, Weidlich, Weijters, Wen, Westergaard, and Wynn[4]. Each step, including start- and end-points, is a process event which was caused by an *activity*. Often, both terms, *event* and *activity*, are used interchangeably. However, there are subtle differences. We understand an event as something that happens at a specific point in time. The driving question is *when* the event happens. In contrast, an activity is related to the content of an event. Here, we ask *what* happens at a point in time. For instance, if we apply for a loan that requires an approval by one person and aferwards a second approval, we can call both activities **APPROVAL**. Although both activities are fundamentally the *same*, they

happen at different points in time. Henceforth, both events remain *different*. Mainly, because one can argue that both events have varying time dependent contexts. For instance, an approval at daytime might be caused by different reasons, than an event caused at night-time.

Each process event may contain additional information in the form of event attributes. If a collection of events *sequentially* relate to a single run through a process, we call them *process instance* or *trace*. These instances do not have to be completed. Meaning, the trace might end prematurely. In line with the aforementioned examples, these process instances could be understood as a single loan application, a medical case or a buy order. We can also attach process instance related information to each instance. Examples would be the applicants location, a patients age or the buyers budget. In its entirety, a business process can be summarised as a *graph*, a *flowchart* or another kind of visual representation. Figure 1's graphical representation is an example of such a *process map*[4].
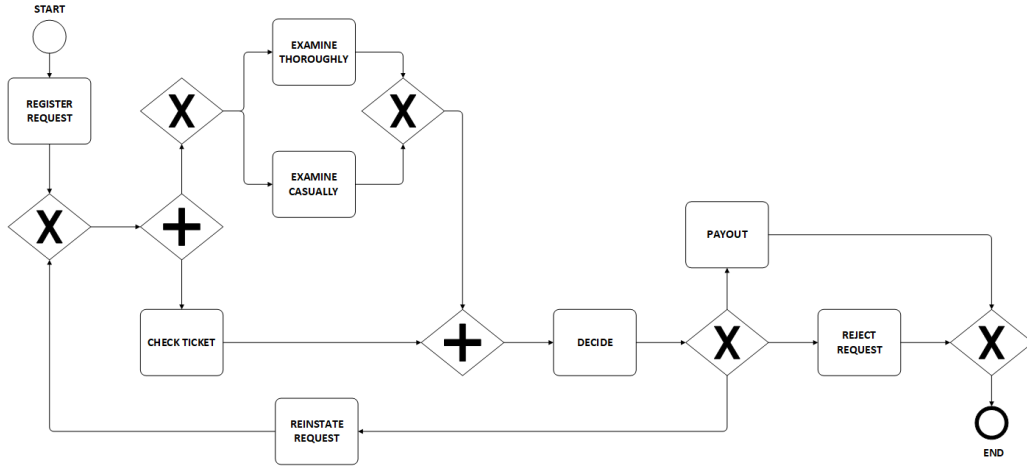

Figure 1: This graph shows an example of a Business Process Modell Notation (BPMN) process map.

In conclusion, in this thesis a *business process* refers to

> *A finite series of discrete events with one or more starting points, intermediary steps and end points. Each intermediate step has at least one precedent and at least one antecedent step.*

However, we have to address a number of issues with this definition.

First, it excludes infinite processes like solar system movements or continuous processes such as weather changes. There may be valid arguments to include processes with these characteristics, but they are not relevant for this thesis.

Second, in each example, we deliberately used words that accentuate modality such as *may, can* or *would*. It is important to understand that each process anchors its definition within an application context. Hence, what defines a business process is indisputably subjective. For instance, while an online marketplace like Amazon might be interested in the process starting from the customers first visit until the successful shipment, an Amazon vendor might only be interested in the delivery process of a product.

Third, the example provided in Figure 1 may not relate to the *real* underlying data generating process. As process *models* are inherently simplified, they may or may not be accurate. The *true* process is often unknown. Therefore, we distinguish between the *true process* and a *modelled process*. The *true process* is a hypothetical concept whose *true* structure remains unknown. In, contrast, a process *model* simplifies and approximates the characteristics of the *true process*.

0.0.2   What is Process Mining?

Having established a definition for a process, we next discuss *Process Mining*. This young discipline has many connections to other fields that focus on the modelling and analysis of processes such as Continuous Process Improvement (CPI) or Business Process Management (BPM)[4]. However, its data-centric approaches originate in Data Mining. The authors van der Aalst, Adriansyah, de Medeiros, Arcieri, Baier, Blickle, Bose, van den Brand, Brandtjen, Buijs, Burattin, Carmona, Castellanos, Claes, Cook, Costantini, Curbera, Damiani, de Leoni, Delias, van Dongen, Dumas, Dustdar, Fahland, Ferreira, Gaaloul, van Geffen, Goel, Günther, Guzzo, Harmon, ter Hofstede, Hoogland, Ingvaldsen, Kato, Kuhn, Kumar, La Rosa, Maggi, Malerba, Mans, Manuel, McCreesh, Mello, Mendling, Montali, Motahari-Nezhad, zur Muehlen, Munoz-Gama, Pontieri, Ribeiro, Rozinat, Seguel Pérez, Seguel Pérez, Sepúlveda, Sinur, Soffer, Song, Sperduti, Stilo, Stoel, Swenson, Talamo, Tan, Turner, Vanthienen, Varvaressos, Verbeek, Verdonk, Vigo, Wang, Weber, Weidlich, Weijters, Wen, Westergaard, and Wynn describe this field as a discipline "to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today's (information) systems"[4]. The discipline revolves around the analysis of event logs. A event log is a collection of process instances, which are retrieved from various sources like an Information System (IS) or database. Logs are often stored in data formats such as Comma Seperated Values (CSV) or eXtensible Event Stream (XES)[4].

0.0.3 The Challenges of Process Mining

As mentioned in **??**, process data modelling and analysis is a challenging task. van der Aalst, Adriansyah, de Medeiros, Arcieri, Baier, Blickle, Bose, van den Brand, Brandtjen, Buijs, Burattin, Carmona, Castellanos, Claes, Cook, Costantini, Curbera, Damiani, de Leoni, Delias, van Dongen, Dumas, Dustdar, Fahland, Ferreira, Gaaloul, van Geffen, Goel, Günther, Guzzo, Harmon, ter Hofstede, Hoogland, Ingvaldsen, Kato, Kuhn, Kumar, La Rosa, Maggi, Malerba, Mans, Manuel, McCreesh, Mello, Mendling, Montali, Motahari-Nezhad, zur Muehlen, Munoz-Gama, Pontieri, Ribeiro, Rozinat, Seguel Pérez, Seguel Pérez, Sepúlveda, Sinur, Soffer, Song, Sperduti, Stilo, Stoel, Swenson, Talamo, Tan, Turner, Vanthienen, Varvaressos, Verbeek, Verdonk, Vigo, Wang, Weber, Weidlich, Weijters, Wen, Westergaard, and Wynn mentions a number of issues that arise from processes[4].

The first issue arises from the quality of the data set. Process logs are seldomly collected with the primary goal of mining information and hence, often appear to be of subpar quality for information mining purposes. The information is often incomplete, due to a lack of context information, the ommision of logged process steps, or wrong levels of granularity[4].

This issue is exacerbated by the second major issue with process data. Mainly, its complexity. Not only does a process logs' complexity arise from the variety of data sources and differing levels of complexity, but also from the datas' characteristics. The data can often be viewed as multivariate sequence with discrete and continuous features and variable length. This characteristic alone creates problems explored in **??**. However, the data is also just a *sample* of the process. Hence, it may not reflect the real process in its entirety. In fact, mining techniques need to incorporate the *open world assumption* as the original process may generate unseen process instances[4].

A third issue which contributes to the datasets' incompleteness and complexity is a phenomenon called *concept drift*[4]. This phenomenon relates to the possibility of changes in the *true* process. The change may occur suddenly or gradually and can appear in isolation or periodically. An expression of such a drift may be a sudden inclusion of a new process step or domain changes of certain features. These changes are not uncommon and their likelihood increases with the temporal coverage and level of granularity of the dataset[4]. In other words, the more *time* the dataset covers and the higher its detail, the more likely a change might have occured over the time.

All three issues relate to the *representativeness* of the data with regards to the unknown *true* process that generated the data. However, they also represent open challenges that require research on their own. For our purpose, we have to assume that the data is representative and its underlying

process is static. These assumptions are widely applied in the body of process mining literature[2, 3].