

Artifact lifecycle discovery

Citation for published version (APA):

Popova, V., Fahland, D., & Dumas, M. (2013). *Artifact lifecycle discovery*. (arXiv.org; Vol. 1303.2554 [cs.SE]). s.n.

Document status and date:

Published: 01/01/2013

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Artifact Lifecycle Discovery

Viara Popova¹, Dirk Fahland², Marlon Dumas¹

¹ Institute of Computer Science,
University of Tartu, J. Liivi 2,
Tartu 50409, Estonia

{viara.popova, marlon.dumas}@ut.ee

² Eindhoven University of Technology
The Netherlands
d.fahland@tue.nl

March 12, 2013

Abstract

Artifact-centric modeling is a promising approach for modeling business processes based on the so-called business artifacts - key entities driving the company's operations and whose lifecycles define the overall business process. While artifact-centric modeling shows significant advantages, the overwhelming majority of existing process mining methods cannot be applied (directly) as they are tailored to discover monolithic process models. This paper addresses the problem by proposing a chain of methods that can be applied to discover artifact lifecycle models in Guard-Stage-Milestone notation. We decompose the problem in such a way that a wide range of existing (non-artifact-centric) process discovery and analysis methods can be reused in a flexible manner. The methods presented in this paper are implemented as software plug-ins for ProM, a generic open-source framework and architecture for implementing process mining tools.

Keywords: Artifact-Centric Modeling, Process Mining, Business Process Modeling

1 Introduction

Traditional business process modeling is centered around the process and other aspects such as data flow remain implicit, buried in the flow of activities. However, for a large number of processes, the flow of activities is inherently intertwined with the process' data flow, often to the extent that a pure control-flow model cannot capture the process dynamic correctly. A prime example is a build-to-order process where several customer orders are collected, and based on the ordered goods multiple material orders are created. Typically, one customer order leads to several material orders and one material order contains items from several different customer orders. These n-to-m relations between customer orders and material orders influence process dynamics, for instance if a customer cancels her order. Such complex dynamics cannot be represented in a classical activity-flow centric process model.

Artifact-centric modeling is a promising approach for modeling business processes based on the so-called business artifacts [5, 19] – key entities driving the company's operations and whose lifecycles define the overall business process. An artifact type contains an information model with all data relevant for the entities of that type as well as a lifecycle model which specifies how the entity can progress responding to events and undergoing transformations from its creation until it is archived.

Most existing work on business artifacts has focused on the use of lifecycle models based on variants of finite state machines. Recently, a new approach was introduced – the Guard-Stage-

Milestone (GSM) meta-model [14, 15] for artifact lifecycles. GSM is more declarative than the finite state machine variants, and supports hierarchy and parallelism within a single artifact instance.

Some of the advantages of GSM [14, 15] are in the intuitive nature of the used constructs which reflect the way stakeholders think about their business. Furthermore, its hierarchical structure allows for a high-level, abstract view on the operations while still being executable. It supports a wide range of process types, from the highly prescriptive to the highly descriptive. It also provides a natural, modular structuring for specifying the overall behavior and constraints of a model of business operations in terms of ECA-like rules.

Process mining includes techniques for discovery and analysis of business process models (such as conformance checking, repair, performance analysis, social networking and so on) from event logs describing actual executions of the process. While artifact-centric modeling in general and GSM in particular show significant advantages, the overwhelming majority of existing process mining methods cannot be applied (directly) in such a setting. The prime reason is that existing process mining techniques are tailored to classical monolithic processes where each process execution can be described just by the flow of activities. When applied to processes over objects in n-to-m relations (expressible in artifacts), classical process mining techniques yield incomprehensible results due to numerous side effects [8].

This paper addresses the problem by proposing a chain of methods that can be applied to discover artifact lifecycle models in GSM notation. We decompose the problem in such a way that a wide range of existing (non-artifact-centric) process discovery and analysis methods can be reused in the artifact-centric setting in a flexible manner. Our approach is described briefly in the following paragraphs.

Typically, a system that executes a business process in a process-centric setting records all events of one execution in an isolated case; all cases together form a log. The cases are isolated from each other: each event occurs in exactly one case, all events of the case together describe how the execution evolved.

Traditional methods for automatic process discovery assume that different process executions are recorded in separate cases. This is not suitable in the context of artifact-centric systems. Indeed, artifact-centric systems allow high level of parallelism and complex relationships between multiple instances of artifacts [8]. This can result in overlapping cases and one case can record multiple instances of multiple artifacts. Therefore, we do not assume that the logs given as input are structured in terms of cases. Instead, all events may be recorded together, without any case-based grouping, as a raw event log.

Such a raw log contains all observed events where each event includes a timestamp reflecting when the event occurred and a collection of attribute-value pairs, representing data that is read, written or deleted due to the occurrence of the event. Fig. 4 gives an example of how such a raw log might look like. The specific format might differ depending on the system that generates it.

Taking such a log as a starting point, we propose a tool chain of methods that can produce a model in GSM notation which reflects the behavior demonstrated in the event log. We decompose the problem in three main steps which would allow us to take advantage of the vast amount of existing research in process mining and reuse some of the existing tools for process discovery.

Fig. 1 shows the proposed overall architecture of the artifact lifecycle discovery process. First, based on the data incorporated in the events, the artifact decomposition of the problem is determined, i.e. how many artifacts can be found and which events belong to which instances of which artifact. Using this information, the log can be decomposed to generate so-called artifact-centric logs where each trace contains the event for one instance of one artifact and each log groups the traces for one artifact.

The artifact-centric logs can then be used to discover the lifecycle of each artifact. For this, any existing method can be used most of which generate models in Petri Net notation. Therefore, as a final step, we apply a method for translating Petri Net models into GSM notation.

The methods presented in this paper are implemented as software plug-ins for ProM [31], a generic open-source framework and architecture for implementing process mining tools. The implementation is part of the **ArtifactModeling** package which is available from www.processmining.org.

The paper is organized as follows. Section 2 presents a working example used for illustration

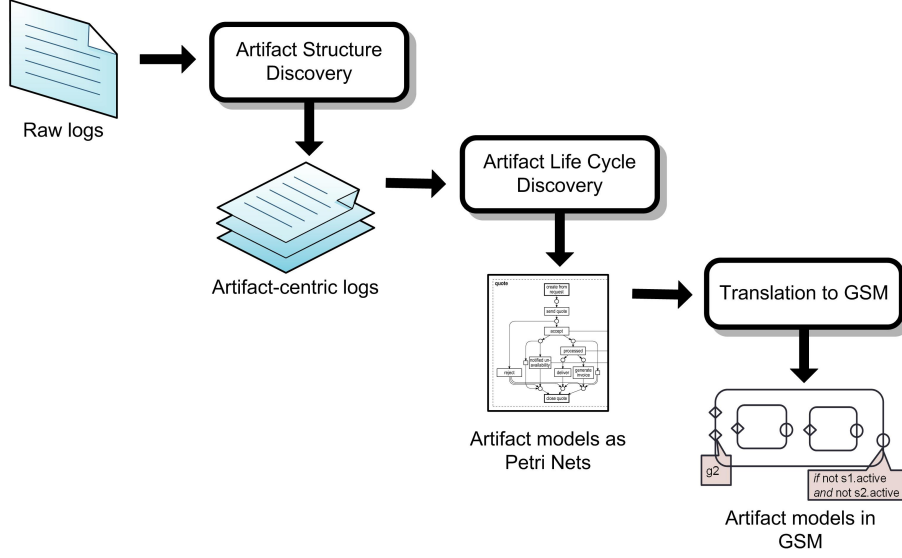


Figure 1: The overall architecture of the artifact lifecycle discovery process.

of the proposed methods and Section 3 reviews the necessary background knowledge. Section 4 presents the artifact structure discovery step of the tool chain. Section 5 discusses the artifact lifecycle discovery step and Section 6 presents the method for translating Petri net models to GSM. Finally, Section 7 concludes the paper with a discussion and directions for future work.

2 The Build-to-Order Scenario

As a motivating example used to illustrate the proposed methods, we consider a build-to-order process as follows. The process starts when the manufacturer receives a purchase order from a customer for a product that needs to be manufactured. This product typically requires multiple components or materials which need to be sourced from suppliers. To keep track of this process, the manufacturer first creates a so-called work order which includes multiple line items – one for each required component. Multiple suppliers can supply the same materials thus the manufacturer needs to select suppliers first, then place a number of material orders to the selected ones.

Suppliers can accept or reject the orders. If an order is rejected by the supplier then a new supplier is found for these components. If accepted, the items are delivered and, in parallel, an invoice is sent to the manufacturer. If the items are of sufficient quality then they are assembled into the product. For simplicity we do not include here the process of returning the items, renegotiating with the supplier and so on. Instead it is assumed that the material order will be marked as failed and a new material order will be created for the items to a different supplier. When all material orders for the same purchase order are received and assembled, the product is delivered to the customer and an invoice is sent for it.

Fig. 2 shows the underlying data model for the build-to-order example which indicates that we can distinguish two artifact types: Purchase order and Material order and one Purchase order corresponds to one or more Material orders.

Fig. 3 shows one way of modeling the lifecycle of the Material order artifact type using Petri net notation. The details of the Petri net notation are explained in the next section. Fig. 4 gives an example of how a raw log recording actual executions of a build-to-order process might look like.

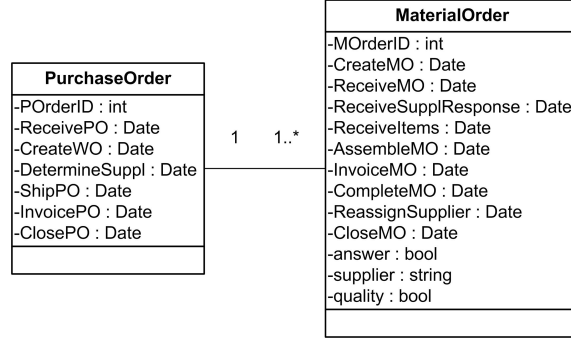


Figure 2: The underlying data model of the build-to-order process.

3 Background

We first give the necessary background in order to present the artifact lifecycle discovery methods by a very brief introduction to the relevant modeling approaches.

3.1 Petri nets

Petri nets [18] are an established notation for modeling and analyzing workflow processes. Its formal basis allow to perform formal analysis w.r.t. many static and dynamic properties. Petri nets are expressive while still being executable which makes them appropriate for application in realistic scenarios. They have been used in a wide variety of contexts and a great number of the developed process mining techniques assume or generate Petri nets.

A Petri net is a directed bipartite graph with two types of nodes called *places* (represented by circles) and *transitions* (represented by rectangles) connected with arcs. Intuitively, the transitions correspond to activities while the places are conditions necessary for the activity to be executed. Transitions which correspond to business-relevant activities observable in the actual execution of the process will be called visible transitions, otherwise they are invisible transitions. A labeled Petri net is a net with a labeling function that assigns a label (name) for each place and transition. Invisible transitions are labeled by a special label τ .

An arc can only connect a place to a transition or a transition to a place. A place p is called a pre-place of a transition t iff there exists a directed arc from p to t . A place p is called a post-place of transition t iff there exists a directed arc from t to p . Similarly we define a pre-transition and a post-transition to a place.

At any time a place contains zero or more tokens, represented graphically as black dots. The current state of the Petri net is the distribution of tokens over the places of the net. A transition t is enabled iff each pre-place p of t contains at least one token. An enabled transition may fire. If transition t fires, then t consumes one token from each pre-place p of t and produces one token in each post-place p of t .

A Petri net N is a *workflow net* if it has a distinguished initial place, that is the only place with no incoming arcs, a distinguished final place, that is the only place with no outgoing arcs, and if every transition of the net is on a path from initial to final place. The initial place is also the only place with a token in the initial marking.

N is *free-choice* iff there is a place p that is a pre-place of two transitions t and s of N (t and s compete for tokens on p), then p is the only pre-place of t and of s . In a free-choice net, a conflict between two enabled transitions t and s can be resolved locally on a single place p . N is *sound* iff every run of N starting in the initial marking can always be extended to a run that ends in the final marking (where only the final place of N is marked), and if for each transition t of N there is a run where t occurs.

In order to model interactions between artifacts in artifact-centric systems, we can use **Pro-clets** [27] notation where each proklet represents one artifact type as a Petri net and constructs

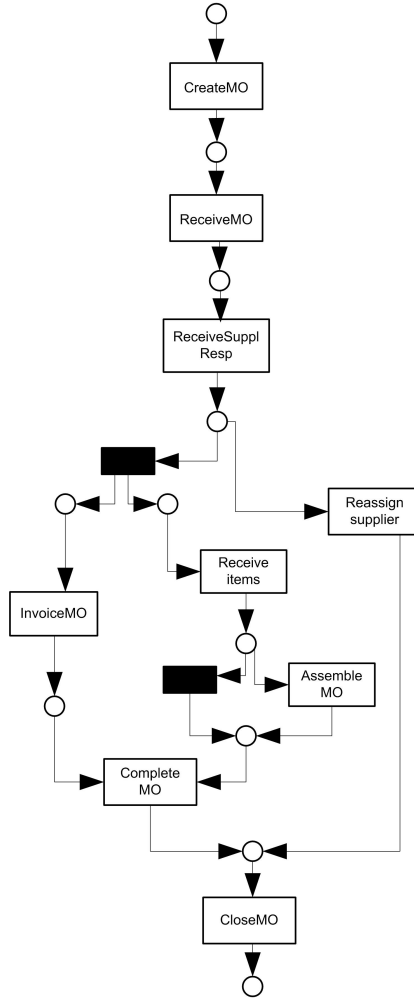


Figure 3: A Petri net model for the lifecycle of the Material Order artifact, part of the build-to-order example

such as ports and channels can be used to represent different types of interactions between the artifacts. In this paper we concentrate on the artifact lifecycles. The difference to classical process discovery is that classical discovery considers just one monolithic process model (Petri net), whereas we consider sets of related Petri nets.

Fig. 3 shows one way of modeling the lifecycle of the Material order artifact type from the build-to-order example using Petri net notation.

3.2 Guard-Stage-Milestone meta-model

The Guard-Stage-Milestone meta-model [14, 15] provides a more declarative approach for modeling artifact lifecycles which allows a natural way for representing hierarchy and parallelism within the same instance of an artifact and between instances of different artifacts.

The key GSM elements for representing the artifact lifecycle are stages, guards and milestones which are defined as follows.

Milestones correspond to business-relevant operational objectives, and are achieved (and possibly invalidated) based on triggering events and/or conditions over the information models of active artifact instances. Stages correspond to clusters of activities performed for, with or by an artifact instance intended to achieve one of the milestones belonging to the stage.

11-24,17:12	ReceivePO	items=(it0), POrderID=1
11-24,17:13	CreateMO	supplier=supp6, items=(it0), POrderID=1, MOrderID=1
11-24,19:56	ReceiveMO	supplier=supp6, items=(it0), POrderID=1, MOrderID=1
11-24,19:57	ReceiveSupplResp	supplier=supp6, items=(it0), POrderID=1, MOrderID=1, answer=accept
11-25,07:20	ReceiveItems	supplier=supp6, items=(it0), POrderID=1, MOrderID=1
11-25,08:31	Assemble	items=(it0), POrderID=1, MOrderID=1
11-25,08:53	ReceivePO	items=(it0,it1,it2,it3), POrderID=2
11-25,12:11	ShipPO	POOrderID=1
11-26,09:30	InvoicePO	POOrderID=1
11-26,09:31	CreateMO	supplier=supp1, items=(it1,it2,it3), POrderID=2, MOrderID=2
11-28,08:12	ReceiveMO	supplier=supp1, items=(it1,it2,it3), POrderID=2, MOrderID=2
11-28,12:22	CreateMO	supplier=supp4, items=(it0), POrderID=2, MOrderID=3
12-03,14:34	ClosePO	POOrderID=1
12-03,14:54	ReceiveMO	supplier=supp4, items=(it0), POrderID=2, MOrderID=3
12-03,14:55	ReceiveSupplResp	supplier=supp1, items=(it1,it2,it3), POrderID=2, MOrderID=2, answer=accept
12-04,15:02	ReceivePO	items=(it1,it2), POrderID=3
12-04,15:20	ReceiveSupplResp	supplier=supp4, items=(it0), POrderID=2, MOrderID=3, answer=accept
12-04,15:33	CreateMO	supplier=supp2, items=(it2), POrderID=3, MOrderID=4
12-04,15:56	ReceiveMO	supplier=supp2, items=(it2), POrderID=3, MOrderID=4
12-04,16:30	CreateMO	supplier=supp5, items=(it1), POrderID=3, MOrderID=5
12-05,09:32	ReceiveMO	supplier=supp5, items=(it1), POrderID=3, MOrderID=5
12-05,09:34	ReceiveItems	supplier=supp4, items=(it0), POrderID=2, MOrderID=3
12-05,11:33	ReceiveItems	supplier=supp1, items=(it1,it2,it3), POrderID=2, MOrderID=2
12-05,11:37	Assemble	items=(it0), POrderID=2, MOrderID=3
12-05,11:50	ReceiveSupplResp	supplier=supp2, items=(it2), POrderID=3, MOrderID=4, answer=reject
12-05,13:03	ReceiveSupplResp	supplier=supp5, items=(it1), POrderID=3, MOrderID=5, answer=accept
12-06,05:23	Assemble	items=(it1,it2,it3), POrderID=2, MOrderID=2
12-06,05:25	ReassignSupplier	items=(it2), POrderID=3, MOrderID=4
12-06,07:14	ReceiveItems	supplier=supp5, items=(it1), POrderID=3, MOrderID=5
12-06,07:15	Assemble	items=(it1), POrderID=3, MOrderID=5
12-06,07:25	InvoicePO	POOrderID=2
12-06,09:34	ShipPO	POOrderID=2
12-12,20:41	CreateMO	supplier=supp5, items=(it2), POrderID=3, MOrderID=6
12-12,20:50	ReceiveMO	supplier=supp5, items=(it2), POrderID=3, MOrderID=6
12-13,03:20	ReceiveSupplResp	supplier=supp5, items=(it2), POrderID=3, MOrderID=6, answer=accept
12-13,03:21	ReceiveItems	supplier=supp5, items=(it2), POrderID=3, MOrderID=6
12-13,04:30	ClosePO	POOrderID=2
12-13,08:36	Assemble	items=(it2), POrderID=3, MOrderID=6
12-13,08:37	InvoicePO	POOrderID=3
12-13,08:38	ShipPO	POOrderID=3
12-13,08:39	ClosePO	POOrderID=3

Figure 4: An example of a raw log.

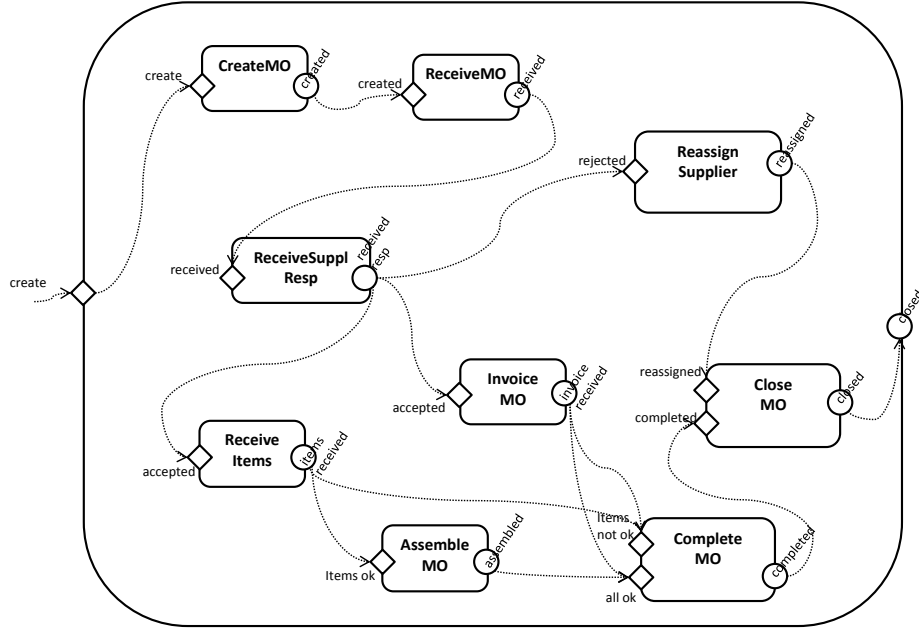


Figure 5: A GSM model for the lifecycle of the Material Order artifact, part of the build-to-order example

Guards control when stages are activated, and, as with milestones, are based on triggering events and/or conditions. A stage can have one or more guards and one or more milestones. It becomes active (or open) when a guard becomes true and inactive (or closed) when a milestone becomes true.

Furthermore, sentries are used in guards and milestones, to control when stages open and when milestones are achieved or invalidated. Sentries represent the triggering event type and/or a condition of the guards and milestones. The events may be external or internal, and both the internal events and the conditions may refer to the artifact instance under consideration, and to other artifact instances in the artifact system.

And finally, stages can contain substages which can be open when the parent stage is active. If a stage does not contain any substages, it is called atomic.

Fig. 5 shows the Material order artifact type as a GSM model. The stages are represented as rectangular shapes. A diamond on the left side of the stage represents a guard with its name and a circle at the right side of the stage represents a milestone. The dotted lines between stages are not part of the model and were only added to give an indication of the control flow that is implicit through the guards and milestone sentries which are expressed in GSM in terms of logics.

4 Artifact Structure Discovery

The first problem to cope with when discovering models of artifact-centric processes is to identify which artifacts exist in the system. Only then a lifecycle model can be identified for each artifact.

In order to discover the artifact structure and subsequently generate the artifact-centric logs, we explore the implicit information contained in the data belonging to the events. As a first step we apply data mining and analysis methods to the raw log data to discover correlation information

between the events which allows to build the underlying Entity-Relationship (ER) model. This includes methods for discovering functional and inclusion dependencies which allow us to discover which event types belong to the same entity and how entities are related to each other.

Using the discovered ER model, we perform analysis which then suggests to the user which entities should be chosen as artifacts and provides support in the selection process.

Finally, for the entities designated as artifacts, we can extract artifact instance-specific traces which can be used to discover the lifecycle of each artifact. The collection of such traces is called *artifact-centric logs*.

The rest of this section presents each of these steps in more detail, as also shown in Fig. 6.

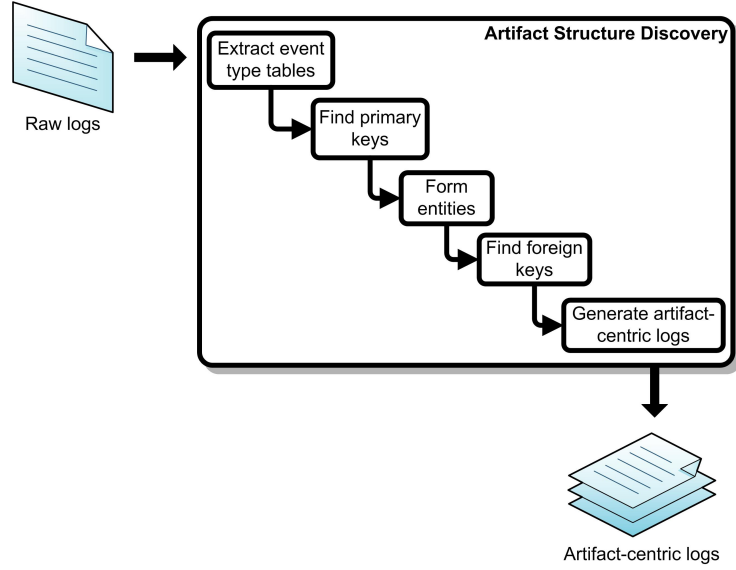


Figure 6: Artifact structure discovery.

4.1 Entity Discovery

Each event in a raw log belongs to an event type, for example the event type `CreateMO` can have one or more instances in the log with specific timestamps and for specific orders being created. Each event contains a timestamp, one or more data attributes and belongs to exactly one event type. Formally, we define a raw log as described in the following paragraphs.

Definition 1 (Event). Let $\{A_1, A_2, \dots, A_n\}$ be a set of attribute names and $\{D_1, D_2, \dots, D_n\}$ - a set of attribute domains where D_i is the set of possible values of A_i for $1 \leq i \leq n$. Let $\Sigma = \{a_1, a_2, \dots, a_m\}$ be a set of event types. Event e is a tuple $e = (a, \tau, v_1, v_2, \dots, v_k)$ where

1. $a \in \Sigma$ is the event type to which e belongs,
2. $\tau \in \Omega$ is the timestamp of the event where Ω is the set of all timestamps,
3. for all $1 \leq i \leq k$ v_i is an attribute-value pair $v_i = (A_i, d_i)$ where A_i is an attribute name and $d_i \in D_i$ is an attribute value.

All events of an event type a are called event instances of a .

Note that the definition allows for multiple attribute-value pairs of the same event to refer to the the same attribute name. In such cases we talk about multi-valued attributes. For example event type `ReceivePO` has a multi-valued attribute `items`.

Definition 2 (Raw log). A raw log L is a finite sequence of events $L = e_1 e_2 \dots e_n$. L induces the total order $<$ on its events with $e_i < e_j$ iff $i < j$.

The order of events in L usually respects the temporal order of their timestamps, i.e. if event e precedes event e' temporally then $e < e'$.

To present the methods proposed in this Section, we adopt notation from *Relational Algebra* [24]. A *table* $T \subseteq D_1 \times \dots \times D_m$ is a relation over domains D_i and has a *schema* $\mathcal{S}(T) = (A_1, \dots, A_m)$ defining for each *column* $1 \leq i \leq m$ an *attribute name* A_i . For each *entry* $t = (d_1, \dots, d_m) \in T$ and each column $1 \leq i \leq m$, let $t.A_i := d_i$. We write $\mathcal{A}(T) := \{A_1, \dots, A_m\}$ for the attributes of T , and for a set \mathcal{T} of tables, $\mathcal{A}(\mathcal{T}) := \bigcup_{T \in \mathcal{T}} \mathcal{A}(T)$. The domain of each attribute can contain a special value \perp which indicates the null value, i.e., the attribute is allowed to have null values for some entries. The set of timestamps Ω does not contain the null value.

All instances of an event type form a data table where each row corresponds to an event instance observed in the logs and consists of the values of all data attributes for that event. In order to transform this table into second normal form (2NF) we need to separate the multi-valued attributes and treat them differently - this will be discussed later. In the following, a single-valued attribute A for event type a is an attribute for which, for each event e of type a in raw log \mathcal{L} , e contains at most one attribute-value pair (A_i, d_i) where $A_i = A$.

Formally, this can be expressed as follows.

Definition 3 (Event type table). *Let a be an event type in the raw log \mathcal{L} and let $E = \{e_1, e_2, \dots, e_n\}$ be the set of events of type a , i.e. $e_i = (a, \tau_i, v_{i_1}, v_{i_2}, \dots, v_{i_m})$ where $m \geq 1$ for all $1 \leq i \leq n$ and $v_{i_j} = (A_j, d_{i_j})$ and A_j is a single-valued attribute for e_i . An event type table for a in \mathcal{L} is a table $T \subseteq \Omega \times D_1 \cup \{\perp\} \times \dots \times D_m \cup \{\perp\}$ s.t. there exists an entry $t = (\tau, d_1, \dots, d_m) \in T$ iff there exists an event $e \in E$ where $e = (a, \tau, (A_1, d_1), (A_2, d_2), \dots, (A_k, d_k))$ and $d_i \in D_i \cup \{\perp\}$ for $1 \leq i \leq m$. The first attribute of the table is called the timestamp attribute of the event type and the rest are the data attributes of the event type.*

Note that the events of the same type might in general have a different number of attributes and the schema of the event type table will consist of the union of all single-valued attribute names that appear in events of this type in the raw log. Therefore there might be null values for some attributes of some events.

Fig. 7 shows two of the event type tables for the raw log in Fig. 4, namely for event types ReceivePO and CreateMO.

The set of event type tables for a given raw log forms a database which implicitly represents information about a number of *entities*. Such entities can be detected by discovering their identifiers in the data tables of the event types. For example if the attribute POrderID is the identifier of both event types ReceivePO and ShipProduct then we can assume that these two events refer to the same entity, in this case the entity PurchaseOrder.

These identifiers will be *keys* in the data tables and can be detected using algorithms for discovering functional dependencies between data attributes. By selecting a primary key, the table is transformed into 2NF.

Definition 4 (Key). *A set of data attributes $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ is a key in an event type table T iff:*

1. *for every data attribute A' in T and every pair of entries $t, t' \in T$ if $t.A' \neq t'.A'$ then $(t.A_1, t.A_2, \dots, t.A_n) \neq (t'.A_1, t'.A_2, \dots, t'.A_n)$, and*
2. *no subset of \mathcal{A} is also a key.*

Note that the timestamp attribute of a table cannot be part of a key.

For example $\{\text{POOrderID}\}$ is a key in the ReceivePO table in Fig. 7 and $\{\text{MOrderID}\}$ is a key in the table CreateMO.

A table can have multiple keys from which only one is selected as the primary key. The multi-valued attributes can then be represented as additional data tables in the following way. For each multi-valued attribute A in event type a with primary key $\{A_1, \dots, A_n\}$ a new table T is constructed, $T \subseteq D_1 \times \dots \times D_n \times D$ such that: D is the domain of A , D_i is the domain

POrderID		ReceivePO	
1		11-24,17:12	
2		11-25,08:53	
3		12-04,15:03	

MOrderID	POrderID	supplier	CreateMO
1	1	supp6	11-24,17:13
2	2	supp1	11-26,09:31
3	2	supp4	11-28,12:22
4	3	supp2	12-04,15:33
5	3	supp5	12-04,16:30
6	3	supp5	12-12,20:41

Figure 7: Event type tables for ReceivePO and CreateMO discovered from the raw log in Fig. 4

POrderID	items
1	it0
2	it0
2	it1
2	it2
2	it3
3	it1
3	it2

Figure 8: The additional table for multi-valued attribute items of event type ReceivePO from the raw log in Fig. 4

of A_i for $1 \leq i \leq n$ and for each tuple $t \in T$, $t = (t.A_1, \dots, t.A_n, t.A)$ there exists event $e = (a, \tau, (A_1, t.A_1), \dots, (A_n, t.A_n), (A, t.A), \dots)$.

An example of such an additional table is the table in Fig. 8 which represents the multi-valued attribute items of event type ReceivePO. An event type table together with all the additional tables associated with it is called an event type cluster.

Event type tables share keys if their primary keys have the same attribute names. An entity is a set of event type tables that share primary keys. The shared primary key is called an identifier for the entity. Every value of the identifier defines an instance of the entity.

Definition 5 (Entity, identifier, instance). Let $\mathcal{T} = \{T_1, T_2, \dots, T_m\}$ be the set of tables of a set of event type clusters in the raw log \mathcal{L} . Let $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ be the (shared) primary key for the tables in \mathcal{T} . An entity E for \mathcal{T} is defined as $E = (\mathcal{T}, \mathcal{A})$ and \mathcal{A} is referred to as an instance identifier for E . An instance s of E is defined as $s = ((A_1, d_1), \dots, (A_n, d_n), (A_{n+1}, d_{n+1}), \dots, (A_l, d_l))$ where an attribute-value pair (A_i, d_i) is in s , $n+1 \leq i \leq l$, iff there exists an entry $t \in T$, $T \in \mathcal{T}$, s.t. $t.A_i = d_i$ and for all $1 \leq k \leq n$ $t.A_k = d_k$.

In this context, we are not discovering and analyzing all entities in the sense of ER models. The specific entities that we discover in the execution logs have behavior, and therefore a non-trivial lifecycle, i.e., go through different states. The changes of state are represented by events in the logs which in turn are reflected in timestamp attributes.

For the raw log in Fig. 4, we find two entities - one with a shared primary key {POrderID} and the other one with {MOrderID}.

The primary key becomes the instance identifier for the corresponding entity. For example an instance of the entity with instance identifier POrderID = 1 will be $s = ((\text{ReceivePO}, \tau_1), (\text{ShipPO}, \tau_2), (\text{InvoicePO}, \tau_3), (\text{ClosePO}, \tau_4))$ where $\tau_1=11-24,17:12$, $\tau_2=11-25,12:11$, $\tau_3=11-26,09:30$ and $\tau_4 = 12-03,14:34$. Names for the entities can be assigned based on the names of the attributes in their instance identifiers or more meaningful names can be given by the user if needed.

The event type tables of the same entity can be combined into one table where attributes with the same name become one attribute. The only exception is when data attributes of the same name which are not part of the primary key have different values for the same value of the primary key. Then they should be represented by different attributes so that both values are present in the joined table. This does not occur in the example of Fig. 4.

An example of an entity table for entity PurchaseOrder is given in Fig. 9. In the following we refer to the combined entity table by the name of the entity.

POrderID	ReceivePO	ShipPO	InvoicePO	ClosePO
1	11-24,17:12	11-25,12:11	11-26,09:30	12-03,14:34
2	11-25,08:53	12-06,09:34	12-06,07:25	12-13,04:30
3	12-04,15:02	12-13,08:38	12-13,08:37	12-13,08:39

Figure 9: The combined table for the entity PurchaseOrder

Once the entities have been discovered then any relationships among them need to be identified. This can be done using algorithms for detecting inclusion dependencies, which discover candidate foreign keys between data tables. From the set of candidate foreign keys the user selects the foreign keys which are meaningful in the specific application.

A database $\mathcal{D} = (\mathcal{T}, K)$ is a set \mathcal{T} of tables with corresponding schemata $\mathcal{S}(T)$, $T \in \mathcal{T}$ s.t. their attributes are pairwise disjoint, and a *key relation* $K \subseteq (\mathcal{A}(\mathcal{T}) \times \mathcal{A}(\mathcal{T}))^{\mathbb{N}}$.

The key relation K expresses foreign-primary key relationships between the tables \mathcal{T} : we say that $((A_1, A'_1), \dots, (A_k, A'_k)) \in K$ relates $T \in \mathcal{T}$ to $T' \in \mathcal{T}$ iff the attributes $A_1, \dots, A_k \in \mathcal{A}(T)$ together are a *foreign key* of T pointing to the *primary key* $A'_1, \dots, A'_k \in \mathcal{A}(T')$ of T' .

In our context we are interested in foreign-primary key relationships between entities. A foreign key in entity E_1 of entity E_2 is an attribute in E_1 which is a foreign key to the instance identifier in E_2 .

A foreign key in entity E_1 which refers to entity E_2 indicates a relationship between E_1 and E_2 . The foreign keys provide us with the necessary correlation information of how specific instances of one entity relate to specific instances of another entity.

The additional tables associated to event type tables contain foreign keys to the primary key of the event type table - such foreign keys do not need to be discovered as they are already known and included by design.

Closer examination of the foreign keys values allows us to find the multiplicities of the discovered relationships.

The collection of entities and their relationships in a raw log \mathcal{L} forms an ER model.

For instance, (MaterialOrder.POrderID, PurchaseOrder.POrderID) is a primary-foreign key relation from the table of entity MaterialOrder to the table of entity PurchaseOrder.

Applying the proposed method to the example raw log we discover an ER model similar to the one in Fig. 2 with the exception of a few of the attributes (PurchaseOrder.CreateWO, PurchaseOrder.DetermineSuppl, MaterialOrder.InvoiceMO, MaterialOrder.CompleteMO and MaterialOrder.CloseMO) since they were not present in the log.

4.2 From Entities to Artifacts

After discovering the underlying ER model, one additional step is needed in order to discover which entities will become artifacts. Intuitively, an artifact can consist of one or more entities from which one is the main entity which determines which events belong to the same instance, i.e. the primary key of the main entity becomes the artifact identifier and all events with the same value of the artifact identifier belong to the same instance of the artifact.

As with any modeling process, there is always a subjective element in deciding what is important and needs to be represented more prominently in the model. Such decision is also domain-specific and depends on the goals and purpose of the model. In the context of artifact-centric modeling, this applies to the decision which entities are important enough to be represented as separate artifacts. However certain guidelines do exist and they were used here to provide heuristics for pruning the options that are not appropriate and for assisting in the selection process.

First of all, an artifact needs to have a lifecycle which means that at least one event type has to be associated with it.

Secondly, m-to-n relations between entities signify that they should belong to different artifacts. For example if a purchase order can be realized by multiple material orders and a material order combines items from different purchase orders then combining them in one artifact is not a good design solution - an instance of such an artifact will overlap with multiple other instances of the same artifact.

Finally, we use the intuition that an entity whose instances are created earlier than the related instances (through foreign key relations) of another entity has a higher probability of being more important. If an entity is not preceded in this way by another one then it should be chosen to become a separate artifact in the artifact-centric model. This for example represents the case where instances of one entity trigger the creation of (possibly multiple) instances of another entity. In the build-to-order example this is the case for the PurchaseOrder since one purchase order triggers the creation of one or more material orders and material orders cannot exist if no purchase order exists.

These considerations will be defined more precisely in the rest of this Section.

Let \mathcal{E} be the set of discovered entities. Let $E, E' \in \mathcal{E}$ be entities and let the tuple $((A_1, A'_1), (A_2, A'_2), \dots, (A_n, A'_n)) \in K$ be a primary-foreign key relationship which defines a relationship between E and E' .

Given such a relationship, we say that an instance s of E identifies a set of instances $\{s'_1, s'_2, \dots, s'_m\}$ of E' if $s.A_i = s'_j.A'_i$ for all $1 \leq i \leq n$, $1 \leq j \leq m$. Here $s.A$ denotes the value of attribute A for the instance s . If, for all instances s in E and for all primary-foreign key relations between E and E' , s identifies at most one instance of E' then we say that E uniquely identifies E' .

The function $H : \mathcal{E} \rightarrow \mathcal{P}(\mathcal{E})$ denotes the logical horizon of an entity [11] where an entity E_1 is in the logical horizon of another entity E_2 if E_2 uniquely (and transitively) identifies E_1 . For example the entity MaterialOrder uniquely identifies the entity PurchaseOrder while PurchaseOrder does not uniquely identify MaterialOrder. Therefore $\text{PurchaseOrder} \in H(\text{MaterialOrder})$ and $\text{MaterialOrder} \notin H(\text{PurchaseOrder})$.

For each instance of entity $E \in \mathcal{E}$, one attribute signifies the timestamp of the creation of this instance - the earliest event that refers to this instance, i.e. the earliest timestamp in the instance. Based on this, for each two instances of the same or different entities we can find out which was created before the other one.

Definition 6 (precedence, top-level entity). *An entity $E_1 \in \mathcal{E}$ precedes an entity $E_2 \in \mathcal{E}$, denoted by $E_1 \prec E_2$, iff $E_2 \in H(E_1)$ and for each instance of E_1 , its creation is always earlier than the creation of the corresponding instance of E_2 . An entity is a top-level entity if it is not preceded by another entity.*

In the build-to-order example, PurchaseOrder is a top-level entity if the raw logs are sufficiently close to the description of the process. Therefore it will become a separate artifact. MaterialOrder can also become an artifact if the user considers it sufficiently important. For example an entity which has a lot of event types (i.e. a more elaborate lifecycle) might be a better choice for an artifact than a smaller entity - still domain-specific aspects and the context of the modeling process need to be considered as well. In this particular example it seems a good choice to make MaterialOrder an artifact as well.

After artifacts are selected, the remaining entities can be joined with artifacts with which they have a key-foreign key relation. If multiple such artifacts exist, then input from the user can be provided to select the most appropriate choice. The resulting grouping of the entities is called an artifact view which is defined more precisely as follows:

Definition 7 (Artifact View). *Let $\mathcal{D} = (\mathcal{T}, K)$ be a database of event type tables. Let $\{E_1, \dots, E_n\}$ be the entities over \mathcal{D} . A set $\text{Art} \subseteq \{E_1, \dots, E_n\}$ of entities is an artifact iff:*

1. *Art contains exactly one designated main entity $\text{main}(\text{Art})$ whose identifier is designated as the artifact instance identifier,*
2. *Art contains at most one top-level entity $\text{top}(\text{Art})$ and, if it exists, $\text{top}(\text{Art}) = \text{main}(\text{Art})$.*

A set $\{\text{Art}_1, \dots, \text{Art}_k\}$ of artifacts is an artifact view on \mathcal{D} iff there is an artifact for each entity in the set of top-level entities $\{E_1, \dots, E_m\}$, $m \leq k$.

4.3 Artifact-Centric Logs

In order to be able to discover an artifact's lifecycle, we need an explicit representation of the executions of its instances based on the events in the raw log. For this purpose, here we introduce the notion of an *artifact-centric log*. As a result of the raw log analysis steps discussed in the previous subsections, the data is now represented in a database in 2NF. A set of entities are represented in the database where each entity has a corresponding table whose primary key is the identifier of the entity and the table contains all the event types belonging to the entity. A set of artifacts are also defined where each artifact consists of one or more entities. Each artifact has as identifier the identifier of one of its entities while the rest are connected through foreign key relations to this entity. In this way, for each artifact we can determine which event types belong to it (the event types of its entities) and which specific events belong to the same instance of the artifact (determined by the same value of the artifact identifier).

Given is a database \mathcal{D} , its set of artifacts and their instance identifiers $\mathcal{I} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n\}$. In the following, $\mathcal{I}(\mathcal{T})$ denotes the set of instance identifiers present in a set of tables \mathcal{T} (and their underlying ER model). Each instance identifier $\mathcal{A} = (A_1, A_2, \dots, A_m)$ has a domain $D_1 \times \dots \times D_m$. For simplicity we denote by $\mathcal{I}(\mathcal{A}) = \{id_1, \dots, id_k\}$ the set of unique values of the instance identifier \mathcal{A} present in the database, $id_i \in D_1 \times \dots \times D_m$. This corresponds to the set of unique instances of the entity for which \mathcal{A} is the instance identifier.

Definition 8 (Instance-aware events). *Let $\Sigma = \{a_1, a_2, \dots, a_n\}$ be a finite set of event types and \mathcal{I} the set of values of instance identifiers. An instance-aware event e is a tuple $e = (a, \tau, id)$ where:*

1. $a \in \Sigma$ is the event type,
2. $\tau \in \Omega$ is the timestamp of the event,
3. $id \in \mathcal{I}$ is the instance for which e occurred.

Let $\mathcal{E}(\Sigma, \mathcal{I})$ denote the set of all instance-aware events over Σ and \mathcal{I} .

Definition 9 (Artifact case, lifecycle log). *An artifact case $\rho = \langle e_1, \dots, e_r \rangle \in \mathcal{E}(\Sigma, \mathcal{I})$ is a finite sequence of instance-aware events that*

1. *occur all in the same instance defined by a value of an instance identifier $\mathcal{A} \in \mathcal{I}$, i.e., for all $e_i, e_j \in \rho$ holds $id_i = id_j$, and*
2. *are ordered by their stamps, i.e., for all $e_i, e_j \in \rho$, $i < j$ implies $\tau_i \leq \tau_j$.*

An artifact lifecycle log for a single artifact is a finite set L of artifact cases. An artifact lifecycle log for an artifact system is a set $\{L_1, \dots, L_n\}$ of lifecycle logs of single artifacts.

We can now address the problem of extracting from a given database D (containing event information from a raw log) and for a given artifact view V , an artifact-centric log that contains a life-cycle log for each artifact in V .

Technically, we need to correlate each event in a table belonging to the artifact to the right instance of that artifact. Once this is done, all events of an artifact are grouped into cases (by the values of the artifact's identifier), and ordered by time stamp. As the instance identifier for an event can be stored in another table the event's time stamp, we need to join the tables of the artifact to establish the right correlation. This requires some notion from relational algebra that we recall first.

Relational algebra defines several operators [24] on tables. In the following, we use *projection*, *selection*, and the canonical crossproduct. For a table T and attributes $\{A_1, \dots, A_k\} \subseteq \mathcal{A}(T)$, the projection $Proj_{A_1, \dots, A_k} T$ restricts each entry $t \in T$ to the columns of the given attributes A_1, \dots, A_k . Selection is a unary operation $Sel_\varphi(T)$ where φ is a boolean formula over atomic propositions $A = c$ and $A = A'$ where $A, A' \in \mathcal{A}(T)$ and c a constant; the result contains entry $t \in Sel_\varphi(T)$ iff $t \in T$ and t satisfies φ (as usual). We assume that each operation correspondingly produces the schema $\mathcal{S}(T')$ of the resulting table T' .

$$\Sigma_{\text{PurchaseOrder}} = \{\text{ReceivePO}, \text{CreateMO}, \text{ReceiveMO}, \text{ReceiveSupplResponse}, \text{ReceiveItems}, \text{Assemble}, \text{ReassignSupplier}, \text{ShipPO}, \text{InvoicePO}, \text{ClosePO}\},$$

$$\text{Inst}(\Sigma_{\text{PurchaseOrder}}) = \text{POOrderID}$$

event type $a \in \Sigma_{\text{PurchaseOrder}}$	$TS(a)$	$Path_i(a)$
ReceivePO	ReceivePO	$\{(\text{PurchaseOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
CreateMO	CreateMO	$\{(\text{MaterialOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
ReceiveMO	ReceiveMO	$\{(\text{MaterialOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
ReceiveSupplResponse	ReceiveSupplResponse	$\{(\text{MaterialOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
ReceiveItems	ReceiveItems	$\{(\text{MaterialOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
Assemble	Assemble	$\{(\text{MaterialOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
ShipPO	ShipPO	$\{(\text{PurchaseOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
InvoicePO	InvoicePO	$\{(\text{PurchaseOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
ClosePO	ClosePO	$\{(\text{PurchaseOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$
ReassignSupplier	ReassignSupplier	$\{(\text{MaterialOrder.POrderID}, \text{PurchaseOrder.POrderID})\}$

Table 1: The artifact view for artifact PurchaseOrder

We define the partial function $Path : 2^{\mathcal{A}(\mathcal{T})} \times 2^{\mathcal{A}(\mathcal{T})} \not\rightarrow K^*$ that returns for two sets of attributes the sequence of key relations needed to connect the attributes. Technically, $Path(\mathcal{A}_s, \mathcal{A}_e) = (\mathcal{A}_1, \mathcal{A}'_1) \dots (\mathcal{A}_k, \mathcal{A}'_k)$ where:

1. for all $i = 1, \dots, k$, $(\mathcal{A}_i, \mathcal{A}'_i) \in K$ or $(\mathcal{A}'_i, \mathcal{A}_i) \in K$, and
2. there exist tables $T_1, \dots, T_k \in \mathcal{T}$ such that $\mathcal{A}_s, \mathcal{A}_1 \in \mathcal{A}(T_1), \mathcal{A}'_k, \mathcal{A}_e \in \mathcal{A}(T_k)$ and for all $i = 1, \dots, k-1$, $\mathcal{A}'_i \in \mathcal{A}(T_i)$ and $\mathcal{A}_{i+1} \in \mathcal{A}(T_{i+1})$.

Let $T(Path(\mathcal{A}_s, \mathcal{A}_e)) = \{T_1, \dots, T_k\}$ denote the tables of this path.

To relate the values of \mathcal{A}_s and \mathcal{A}_e to each other, we need to join the tables of the path that connects both attribute sets. Let $Path(\mathcal{A}_s, \mathcal{A}_e)$ be the path from \mathcal{A}_s to \mathcal{A}_e and let $\{T_1, \dots, T_k\} = T(Path(\mathcal{A}_s, \mathcal{A}_e))$ be the involved tables. Then $Join(Path(\mathcal{A}_s, \mathcal{A}_e)) = Sel_\varphi(T_1 \times \dots \times T_k)$ where $\varphi = \bigwedge_{(A, A') \in Path(\mathcal{A}_s, \mathcal{A}_e)} (A = A')$ selects from the cross product $T_1 \times \dots \times T_k$ only those entries which coincide on all key relations.

$Path$ defines for each event type associated with a specific artifact the path to the instance identifier attribute associated to this artifact. The definition does not impose any restrictions on these paths however in practice it often makes sense to choose the shortest path between the tables. Classical graph theory algorithms such as Dijkstra algorithm can be used for finding all shortest paths.

For the sake of illustrating these definitions, let us assume that MaterialOrder was not chosen as a separate artifact and was joined with PurchaseOrder through the foreign key relation provided by the attributes MaterialOrder.POrderID and PurchaseOrder.POrderID. Table 1 presents the artifact view for the artifact PurchaseOrder for this variation of our running example.

Log Extraction. After specifying an artifact view, an artifact log can be extracted fully automatically from a given database \mathcal{D} .

Let TS_i be the set of all timestamp attributes in \mathcal{T}_i , i.e. $TS \in TS_i$ iff there exists table $T \in \mathcal{T}_i$ and $TS \in \mathcal{A}(T)$ with domain of T being Ω .

Definition 10 (Log Extraction). *Let $\mathcal{D} = (\mathcal{T}, K)$ be a database and let $\{E_1, \dots, E_n\}$ be the entities over \mathcal{D} . Let $\{Art_1, \dots, Art_k\}$ be an artifact view on \mathcal{D} .*

For each artifact Art_i , the artifact life-cycle log of Art_i is extracted as follows.

1. Let $(\hat{\mathcal{T}}_i, \hat{\mathcal{A}}_i) = \text{main}(Art_i)$ be the main entity of Art_i with identifier $\hat{\mathcal{A}}_i$. Let $\mathcal{T}_i = \bigcup_{(T, A) \in Art_i} \mathcal{T}$ be the set of all tables of Art_i .
2. For each timestamp attribute $TS \in TS_i$ we define the table $T_{TS}^+ = \text{Proj}_{\hat{\mathcal{A}}_i, TS} Join(Path(\{TS\}, \hat{\mathcal{A}}_i))$.
3. Each entry $t = (id, ts) \in T_{TS}^+$ defines an instance aware event $e = (TS, ts, id)$ of type with timestamp TS in instance id .
4. The set of all instance-aware events of artifact i is $\mathcal{E}_i = \{(TS, ts, id) \mid \exists TS \in TS_i, (ts, id) \in T_{TS}^+\}$, let $\mathcal{I}_i = \{id \mid (TS, ts, id) \in \mathcal{E}_i\}$ be the instance identifiers of Art_i .

$Join(\{PurchaseOrder, MaterialOrder\}, K)$					
PurchaseOrder .POrderID	MaterialOrder .POrderID	MaterialOrder .supplier	MaterialOrder .ReceiveMO	MaterialOrder .MOrderID	...
1	1	supp6	11-24,19:56	1	...
2	2	supp1	11-28,08:12	2	...
2	2	supp4	12-03,14:54	3	...
3	3	supp2	12-04,15:56	4	...
3	3	supp5	12-05,09:32	5	...
3	3	supp5	12-12,20:50	6	...

Table 2: Intermediate table when extracting events of artifact order.

5. For each $id \in \mathcal{I}_i$, the artifact case ρ_{id} contains all events $(TS, ts, id) \in \mathcal{E}_i$ ordered by their timestamps.
6. The artifact log $L_i = \{\rho_{id} \mid id \in \mathcal{I}_i\}$ contains all artifact cases of Art_i .

We illustrate the log extraction using our running example from Sect. 2. We consider the artifact view on PurchaseOrder as specified in Tab. 1. We explain event extraction on event type ReceiveMO.

1. First join the tables PurchaseOrder and MaterialOrder on (PurchaseOrder.POrderID, MaterialOrder.MOrderID); Tab. 2 shows a part of the resulting table.
2. Projection onto the instance identifier $Inst(PurchaseOrder) = POrderID$ and the timestamp attribute ReceiveMO yields six entries (1, 11-24,19:56), (2, 11-28,08:12), (2, 12-03,14:54) and so on.

Similarly, events for the other event types of artifact PurchaseOrder can be extracted to construct full cases.

In general we extract one log for every artifact and they can be considered independently to discover the lifecycle of each artifact. This will be discussed in the next Section.

5 Artifact Lifecycle Discovery

Using the artifact-centric logs, we apply process mining techniques in order to discover the lifecycle of each artifact independently. A great number of algorithms for process discovery exist with varying representational and search bias [26, 25]. The representational bias refers to the chosen formalism for representing the process model. Most approaches use some form of a directed graph, for example Petri Nets [28, 3, 29, 32], Finite State Machines [6], Causal Nets [33], Process Trees [4], and so on. A few use other representations such as temporal logic [21] or user-defined constraint templates [17].

The search bias refers to the algorithm used to traverse the solution space and the criteria used to select the final answer (i.e., the generated process model). Many approaches have been proposed including Markov Models [6], Genetic Algorithms [7, 28, 4], Neural Networks [6], Integer Linear Programming [34], custom algorithms [29, 3, 32, 33] and so on.

The primary goal of the process discovery approaches is to generate models that accurately represent the behavior of the system as evidenced by the logs. A number of criteria were defined as well as measures for assessing to what degree the model reflects the desired behavior, as discussed in [26]. The four prominent measures are fitness (to which degree a model can replay each trace in the log), generalization (to which degree a model allows replaying traces that are similar to the traces in the log and considered as part of the process), precision (to which degree a model allows replaying traces that are not contained in the log, but different from the traces in the log and not considered part of the process), and structural simplicity of the model. These four measures are partly contradictory, e.g. the simplest model has a very low precision, a perfectly fitting model can require very complex structure and thus a low simplicity. For this reason, different algorithms consider different subsets of these criteria.

For complex processes, the importance of the readability (i.e. simplicity) measure for the process model increases. Different approaches to increasing the readability of the generated models have been proposed including model simplification [9], abstraction [3], fuzzy models [30, 12], trace clustering to generate simpler process variants [2], block-structured models [4] and so on.

As an illustration, we briefly describe two of the successful algorithms with their advantages and disadvantages.

The ILP Miner [34] uses an Integer Linear Programming approach based on the language-based theory of regions. It generates models that are guaranteed to be consistent with the logs. However it can produce models that are not very structured and less readable (i.e. spaghetti models). As an example, Figure 10 shows a Petri net model mined from logs for the **PurchaseOrder** artifact in a variation of the build-to-order example.

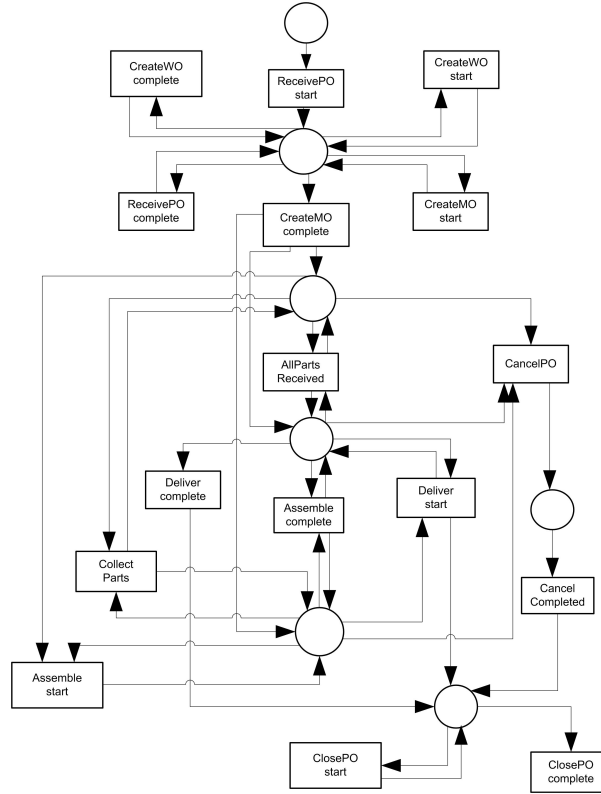


Figure 10: Petri Net mined using the ILP Miner.

The Tree Miner [4] uses a Genetic Algorithm for generating a model where the fitness function is balanced between the four criteria for model quality: replay fitness, precision, simplicity and generalization. The internal representation of the model is in the form of a process tree which has activities as leaves and the internal nodes represent operators such as sequence, parallel execution, exclusive choice, non-exclusive choice and loop execution. This guarantees that the generated models will be block-structured and sound. However the models are not guaranteed to have perfect fitness and therefore might not be consistent with the logs. Another disadvantage is that the algorithm takes significantly more time than for example the ILP miner.

Figure 11 shows a Petri net model mined from the same logs of the **PurchaseOrder** artifact using the tree miner. We can clearly see that the model is more structured and readable than the one in Figure 10. The downside is that this model deviates somewhat from the logs - for example a couple of the activities that appear in the logs are not present in the model. Running the algorithm multiple times until a better fitness is achieved could potentially result in a better model though this will additionally increase the necessary time.

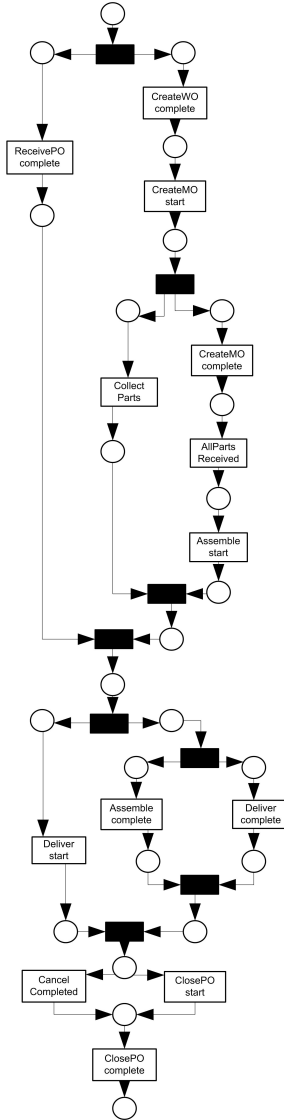


Figure 11: Petri Net mined using the Tree Miner.

In addition to the wide range of process discovery algorithms, a number of other relevant approaches have been proposed for tasks such as conformance checking [1, 23], model repair [10] and so on. Conformance checking methods can be used in this setting to check if the generated model conforms to the logs and, if that is not the case, model repair methods can be applied to repair the misconformances.

The above overview proves the benefits of choosing an approach that allows to reuse existing work and allow flexibility and compositionality of the tool chain. Since the majority of process discovery and analysis approaches generate Petri nets or models that can trivially be converted to Petri nets, we choose Petri nets as intermediate representation of the single artifact models generated from artifact-centric logs. The generated Petri nets are then translated into Guard-Stage-Milestone models which represent the same behavior. Next Section presents the proposed translation algorithm in detail.

6 Petri Nets to GSM models

We now concentrate on the last step in our artifact lifecycle discovery tool chain which deals with translating Petri net models to GSM.

As discussed in Section 3, in the GSM model each (atomic) stage has a guard and a milestone. The stage becomes active if and when the guard sentry evaluates to true. The guard sentry depends on a condition and/or an event. Events can be internal or external and can reflect the changes in the state of the instance or other instances, for example opening or closing of stages, achieving of milestones, etc. When an atomic stage opens, the activity associated with it will be executed. This can happen automatically (for example when the activity is the assignment of a new value(s) to variable(s) in the instance's information model) or can require actions by human agents. In the latter case the actual task execution is external. The finishing of the execution of the task also generates an event which is denoted here as follows: for an atomic stage A , the event `ATaskExecuted()` is generated when the task execution finishes. The milestone of the stage is achieved or invalidated also based on a sentry that depends on an event and/or condition. Achieving a milestone M of some stage generates an event `MAchieved()` that can be used in the guard of another stage. This way, the execution of stages can be ordered.

A straightforward approach to translating Petri nets to GSM models would proceed as follows. The visible transitions of the Petri net represent activities which are part of the business process. Therefore it is logical to represent them as atomic stages where the activity corresponds to the task associated with the stage. The control flow of the Petri net can then be encoded using the guards and milestones of these stages.

More specifically, in the Petri net N , the order of execution of transitions is expressed by the places and arcs of N . To impose the same execution order on the stages of M , we encode the execution order of N in the guards and milestones of the stages of M . In particular, if transition t_2 is a successor of transition t_1 , then the guard g_{t_2} has to express the opening of stage s_{t_2} in terms of the milestone m_{t_1} of stage s_{t_1} .

It is possible to encode the places of the Petri net (and their marking) in variables which will be part of the information model of the artifact. These variables will be assigned true or false simulating the presence or absence of tokens in the places. This will be a relatively intuitive approach for designers skilled in the Petri net notation. However we argue that this would make the model less intuitive to the user and the relations between the tasks and stages become implicit and not easy to trace. Here we take a different approach which will be discussed first at a more general level and in the next subsections in more detail.

The intuition behind this approach is that the immediate ordering relations between transitions in the Petri net are extracted, translated into conditions and combined using appropriate logical operators (for AND- and XOR-splits and joins) into sentries which are then assigned to the guards. The milestones are assigned sentries that depend on the execution of the task associated with the stage - a milestone is achieved as soon as the task is executed and is invalidated when the stage is re-opened.

As an example, consider the transition `ReceiveMO` from the build-to-order model in Fig. 3. It can only be executed after the transition `CreateMO` has been executed and there is a token in the connecting place. This can be represented as a part of a GSM model in the following way. Both transitions are represented by atomic stages. The guard of the stage `ReceiveMO` has a sentry with expression "`on CreateMOMilestoneAchieved()`" where `CreateMOMilestoneAchieved` is the name of the event generated by the system when the milestone of stage `CreateMO` is achieved. Therefore the sentry will become true when the event of achieving the milestone of stage `CreateMO` occurs and the stage will be open.

The milestone of stage `ReceiveMO` has a sentry "`on ReceiveMOTaskExecuted()`" where `ReceiveMOTaskExecuted` is the name of the event generated by the system when the task associated with atomic stage `ReceiveMO` completes. Therefore the sentry will become true when the associated task is executed, the milestone will be achieved and the stage - closed. Similarly the milestone of `CreateMO` has a sentry "`on CreateMOTaskExecuted()`".

While this example is very straightforward, a number of factors can complicate the sentries.

For example, we need to consider the possibility of revisiting a stage multiple times - this can be the case when the corresponding transition in the Petri net is part of a loop. At the same time, the transition might depend on the execution of multiple pre-transitions together and this cannot be represented using events - conditions need to be used instead. The conditions should express the fact that new executions of the pre-transitions have occurred. This means that the last execution of each relevant pre-transition occurred after the last execution of the transition in focus but also after every “alternative” transition, i.e., transition that is an alternative choice.

For example consider the transition **CompleteMO** in Fig. 3 which can fire for example if both **AssembleMO** and **InvoiceMO** have fired (which will result in tokens in both pre-places). While this is not part of the model, imagine the hypothetical situation that **CompleteMO**, **AssembleMO** and **InvoiceMO** were part of a loop and could be executed multiple times. Since a sentry cannot contain multiple events, the guard of **CompleteMO** has to be expressed by conditions instead. The naïve solution “if **AssembleMOTask.hasBeenExecuted** and **InvoiceMOTask.hasBeenExecuted**” which checks if the two tasks have been executed in the past is not correct, since it becomes true the first time the activities **AssembleMO** and **InvoiceMO** were executed and cannot reflect any new executions after that. We need a different expression to represent that new executions have occurred that have not yet triggered an execution of **CompleteMO**. This will be discussed in detail in the next section.

Another factor that needs to be considered is the presence of invisible transitions, i.e., transitions without associated activity in the real world. For such invisible transitions no stage will be generated as they have no meaning at the business level that is meant to be reflected in the GSM model. Therefore, in order to compose the guard sentries, only visible pre-transitions should be considered. Thus we need to backtrack in the Petri net until we reach a visible transition and “collect” the relevant conditions of the branches we traverse. As an example, consider the transition **ReceiveItems** in Fig. 3. It can only fire when the invisible pre-transition represented by a black rectangle fires. We backtrack to find the pre-places of the invisible transition and their pre-transitions. Here we determine that the only such pre-transition is **ReceiveSupplResponse** and this branch has an associated condition - we can only take this branch if the supplier rejects an order and a new supplier has to be determined.

With all these considerations in mind, the resulting guard sentry can become more complex and partly lose its advantage of being able to give intuition about how the execution of one task influences the execution of others. In order to simplify the sentry expressions, we apply methods for decomposing the expression into multiple shorter and more intuitive sentries which are then assigned to separate guards of the same stage. Each guard of a stage forms an *alternative* way of opening a stage, i.e., only one guard has to evaluate to true. The composition and decomposition of guard sentries will be described more precisely in the next section.

Let t_o be the “origin”, i.e., the (visible) transition for which we compose a guard. At a more abstract level the proposed method for generating guard sentries for the stage of t_o proceeds as follows (as also shown in Fig. 12):

Step 1: Find the relevant branch conditions and the pre-transitions whose execution will (help) trigger the execution of t_o .

Step 2: Decompose into groups that can be represented by separate guards.

Step 3: For each group, determine the appropriate format of the sentry and generate its expression.

We describe in detail each of these steps in the next subsections.

6.1 Guard Sentries Generation

Our approach for achieving step 1 is inspired by the research presented in [20] for translating BPMN models and UML activity Diagrams into BPEL. It generates so-called precondition sets for all activities which encode possible ways of enabling an activity. Next, all the precondition sets with their associated activities, are transformed into a set of Event-Condition-Action (ECA) rules.

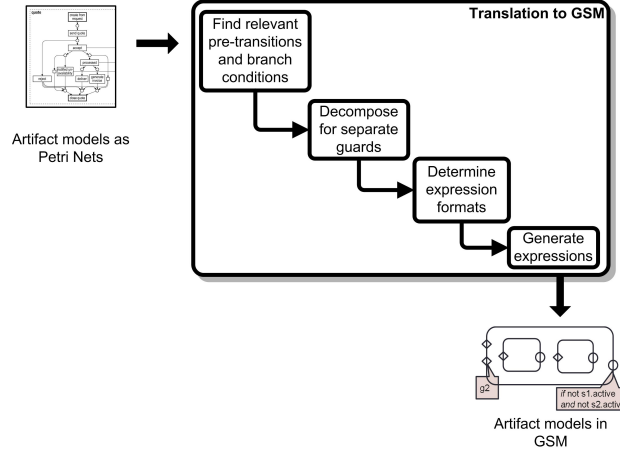


Figure 12: Translation of Petri Nets to GSM models.

Before giving the precise definitions of the approach proposed here, we first illustrate the intuition behind it by a couple of examples. Consider the transition **CompleteMO** in Fig. 3. In order for it to be enabled and subsequently fire, there need to be tokens in both of its pre-places. Therefore the precondition for enabling **CompleteMO** is a conjunction of two expressions, each of which related to one pre-place and representing the fact that there is a token in this pre-place. This token could come from exactly one of the pre-transitions of this place.

As a second example, consider the transition **CloseMO**. It has one pre-place which has two pre-transitions. The token needed to enable **CloseMO** could come from either **CompleteMO** or **ReassignSupplier**. Therefore the precondition here is a disjunction of two expressions each related to the firing of one pre-transition.

Thus the general form of the composed guard sentry expression is a conjunction of disjunctions of expressions. These expressions, however, can themselves be conjunctions of disjunctions. This happens when a pre-transition is invisible (not observable in reality) and we need to consider recursively its pre-places and pre-transitions.

The building blocks of the composed expression are expressions each of which corresponds to the firing of one visible transition t that can (help) trigger the firing of the transition in focus t_o (the “origin”). We denote each of these building blocks by $tokenAvailable(t, t_o)$ for a transition t with respect to t_o and they will be discussed in the next section. They represent the intuition that t has produced a token that can enable t_o if all other necessary tokens (produced by other pre-transitions) are available. More precisely, $tokenAvailable(t, t_o)$ is true if t has produced a token that has not yet been consumed by t_o or by any other transition that could be enabled by it and is false otherwise.

The presence of a token in a pre-place is not a guarantee that a transition will fire. In the case of **AssembleMO**, a token in its pre-place enables two transitions, **AssembleMO** and an invisible transition, but only one will fire. In order to resolve the non-determinism w.r.t. which transition fires and consumes the token, conditions are associated with each outgoing arc of the place. Here **AssembleMO** will fire if the received items are of sufficient quality. These conditions are domain-specific and, in the following, we assume that these conditions are given - they can be provided by the user or mined from the logs using existing tools such as the decision miner from [22]. Therefore, the general form of the composed expression should have these conditions added to the conjunction.

Using this approach for the transition **CompleteMO** we can find a sentry expression of the following type (only given informally here): “**CompleteMO** will open if **InvoiceMO** is executed AND (**AssembleMO** was executed OR (**Receiveltems** was executed AND quality is insufficient))”. In the rest of this Section the format will be defined precisely in a way that allows automatic generation given a Petri net model.

For the translation, we assume that the artifact lifecycle is modeled as a sound, free-choice workflow net N . For example the genetic algorithm presented in Section 5 always returns a lifecycle model that is a sound, free-choice workflow net.

By $enabled(t_o)$ we denote the composed expression (of “pre-conditions”) of the guard sentry for a stage/transition t_o . It is true if all necessary tokens for firing t_o are available and the needed branch conditions are true. As a guard sentry it will ensure that, as soon as this is the case, the stage will open.

By N being a free-choice net, the enabling condition of transition t is essentially a positive boolean formula of conjunctions and disjunctions over the *visible* predecessors of t : for t to be enabled, there has to be a token in each pre-place $p \in pre(t)$ of t , and a token in p is produced by *one* pre-transition $s \in pre(p)$ of p . In addition, the occurrence of transition t itself can be subject to a condition represented by a variable $cond_t$ in the information model of the artifact. We assume that in general $cond_t$ can change its value during the lifecycle of the instance.

Also, $init$ denotes the specific expression used to represent the event of the creation of an artifact instance, e.g. “onCreate()”.

We can then define the predicate $enabled(t)$, assuming an “origin” t_o , by a recursive definition as follows:

$$enabled(t) = (\bigwedge_{p \in pre(t)} markable(p, t)) \wedge cond_t$$

stating that transition t is enabled iff its guarding condition is satisfied and each pre-place p can be marked. The predicate $markable(p, t)$ is defined as:

$$markable(p, t) = \begin{cases} init, & p \text{ initially marked} \\ \bigvee_{r \in pre(p)} occurred(r, t), & \text{otherwise} \end{cases}$$

stating that place p is either initially marked or can become marked by an occurrence of any of its directly preceding transitions. A directly preceding transition r is either visible or invisible. The predicate $occurred(r, t)$ is then defined as follows:

$$occurred(r, t) = \begin{cases} tokenAvailable(r, t_o), & r \text{ visible} \\ enabled(r), & \text{otherwise} \end{cases}$$

Here, as mentioned earlier, $tokenAvailable(t_p, t_o)$ is the specific expression that will be added to the sentry condition for each relevant visible transition t_p with respect to the “origin” t_o . The condition has to be precise enough to resolve any non-determinism between the guards of the subsequent stages connected to this sentry. The format of these conditions will be discussed in the next section.

The recursion in the above definition stops either when reaching an initially marked place or when reaching all visible predecessors. Thus, the predicate is only well-defined if the net N contains no cycle of invisible transitions.

The expression for $enabled(t_o)$ can be represented in a tree structure in a straightforward way. The internal nodes of the tree represent logical operators (“and” or “or”) which are applied on their child branches. The leaves represent either transitions that need to fire (which will be represented in the guard sentry by an expression $tokenAvailable(t_p, t_o)$ for the specific transition t_p in the leaf) or decision point conditions that need to be true in order for the “origin” transition t_o to be able to fire. In the following we use the words tree and expression interchangeably since, in this context, they represent the same information.

An example of such a tree is given in Fig. 13 constructed for the transition **CompleteMO**. Looking at the model in Fig. 3 we can see that **CompleteMO** can only fire if there is a token in each of its pre-places. One of these tokens is generated by firing the transition **InvoiceMO**. The other token can arrive from two possible transitions - **AssembleMO** or the invisible transition represented as a black rectangle. We traverse back from the invisible transition and find out that it can only fire if the transition **ReceiveItems** fires and the condition associated with the connecting arc is true (the received items have sufficient quality). This analysis results in the tree

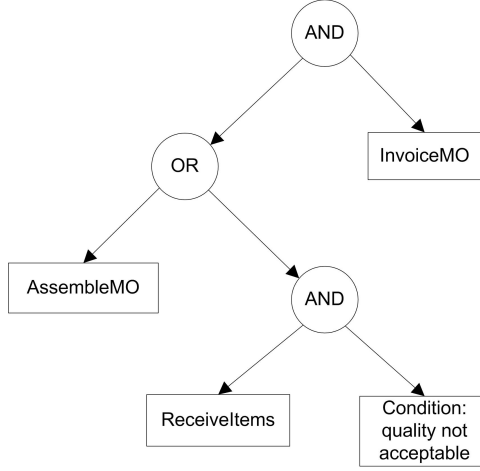


Figure 13: An example of an expression tree which will be used to generate the guard(s) for stage CompleteMO.

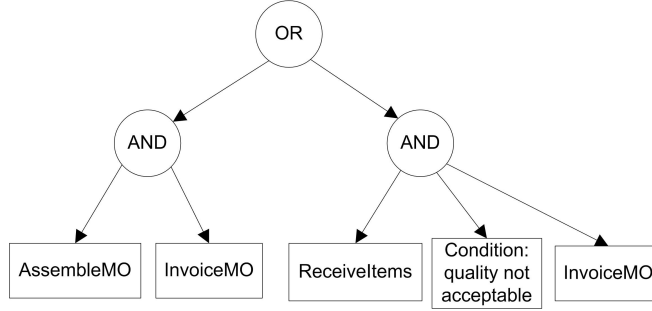


Figure 14: The expression tree for stage CompleteMO in DNF.

in Fig. 13. The leaves of the tree are named by the corresponding transition or condition and, in fact, represent the specific expression for that transition/condition. However we delay the exact formulation of the expressions until the tree is built and analyzed, as will be described in the next section.

As mentioned earlier, an intermediate step of the algorithms decomposes $enabled(t_o)$ into several expressions which are then used to generate separate guards of the stage. Since $enabled(t_o)$ is a logical formula, we can convert it into Disjunctive Normal Form (DNF) and assign each conjunction to a separate guard sentry.

After converting the example tree from Fig. 13 into DNF, we now have the tree in Figure 14. Each child of the root node will generate one separate guard - here we have two guards. Intuitively the first guard tells us that the stage will open if the items were assembled and invoice received. Similarly, the second guard tells us that the stage will open if the items and invoice were received but the quality was insufficient to perform the assembling step.

As a final step, the $tokenAvailable(t_p, t_o)$ for the leaves of the tree are assigned as discussed in the next section.

6.2 Formats for Pre-condition Expressions

In this section we look into the expressions $tokenAvailable(t_p, t_o)$ in more details and define their format. The assignment of a specific format to the expressions is delayed until the end, after $enabled(t_o)$ is composed and, if needed, decomposed into separate sentries $\{enabled_1(t_o), \dots, enabled_n(t_o)\}$ representing all alternative ways the stage can be open. Only then

it can be decided which format each expression should take. We consider two possible formats for the expression of $tokenAvailable(t_p, t_o)$ depending on the context as discussed below.

6.2.1 A simple format for pre-condition expressions

The most simple case is when $enabled_i(t_o)$ contains only one transition t_p with its expression $tokenAvailable(t_p, t_o)$ and $init$ is not present in $enabled_i(t_o)$. Then $tokenAvailable(t_p, t_o)$ can be replaced by the event corresponding to the finished execution of the activity of t_p . It can be expressed using the event of achieving the milestone of the stage of t_p or, alternatively, the closing of that stage among other options.

For example, for $t_o = \text{ReceiveSupplResponse}$ the expression tree contains only one leaf corresponding to the transition $t_p = \text{ReceiveMO}$, i.e., the only way to enable t_o is by a token produced by t_p and this token cannot be consumed by another transition. Then the expression for t_p and t_o will be $enabled(t_o) = tokenAvailable(t_p, t_o) = \text{"on ReceiveMOMilestoneAchieved()}"$ where $\text{ReceiveMOMilestoneAchieved()}$ is the event generated by the system when the milestone of stage ReceiveMO is achieved, therefore the task associated with the stage was completed and the stage is closed.

If this is not the case, i.e., multiple transitions are present, then a more complex version of the $tokenAvailable(t_p, t_o)$ expression needs to be included since we cannot use more than one event in the sentry. This form of the expression is discussed in the following sub-section.

6.2.2 A complex format for pre-condition expressions

A token produced by the visible transition t_1 is available for t_2 as long as neither t_2 , nor any transition that is alternative to t_2 consumed that token. In a free-choice Petri net, a transition t is alternative to t_2 if they both have a common preceding place p that is itself preceded by t_1 . We say a node x precedes a node y in net N , written $x \rightarrow y$ iff there is path from x to y along the arcs of N . If this path only involves τ -labeled transitions (and arbitrarily labeled places), we write $x \xrightarrow{\tau} y$.

With this in mind, we now define the following set of transitions that succeed t_p and are alternative to t_o , i.e., visible transitions that are connected to a place on the path from t_p to t_o :

$$Alt(t_p, t_o) = \{t \mid \exists \text{ place } p : t_p \xrightarrow{\tau} p \xrightarrow{\tau} t_o \wedge p \xrightarrow{\tau} t\}.$$

$Alt(t_p, t_o)$ are the set of transitions that “compete” with t_o for the token produced by t_p . Therefore, in order to represent the situation when a token is present in the pre-place of t_o and the stage t_o should be opened, we need to consider whether any of the “alternative” transitions have occurred (and “stolen” the token). Note that, according to this definition, t_o will also belong to the set.

Let us consider again $t_o = \text{CompleteMO}$ and the expression tree in Fig. 14. Here we cannot use the simple format of the expressions for each leaf since in each AND-subtree there is more than one transition. For the right-most AND-subtree, let us consider the leaf $t_p = \text{ReceiveItems}$, $Alt(t_p, t_o) = \{\text{CompleteMO}, \text{AssembleMO}\}$. Looking at Fig. 3, we can see that the transition AssembleMO is indeed an “alternative” to the invisible transition in the path from t_p to t_o in the sense that it can “steal” the token produced by the transition ReceiveItems in the connecting place.

As a second building block we define the expression $executedAfter(t_p, t_s)$ which is true when there is a new execution of t_p which occurs after the last execution of t_s (meaning that it is relevant for triggering the opening of the stage of t_o) and false otherwise. In terms of the Petri net it will be true when t_s has produced a token that can potentially enable t_o , if it does not get “stolen” by another transition in the $Alt(t_p, t_o)$ set.

How $executedAfter(t_p, t_s)$ will be expressed in the specific implementation can vary. Here we show how this can be done using the state of a milestone (achieved or not) and the time a milestone was last toggled. By $m_p.hasBeenAchieved$ we denote a Boolean variable in the information model of the artifact which is true if the milestone m_p of stage t_p is in state “achieved” and false otherwise. For every milestone m_p present in a stage of the artifact there is also a variable in the information

Stage	Guard
CreateMO	onCreate()
ReceiveMO	on CreateMOMilestoneAchieved()
ReceiveSupplResponse	on ReceiveMOMilestoneAchieved()
ReassignSupplier	on ReceiveSupplResponseMilestoneAchieved() if answer = reject
InvoiceMO	on ReceiveSupplResponseMilestoneAchieved() if answer = accept
ReceiveItems	on ReceiveSupplResponseMilestoneAchieved() if answer = accept
AssembleMO	on ReceiveItemsMilestoneAchieved() if quality = acceptable
CompleteMO	if InvoiceMOMilestone.hasBeenAchieved = true and AssembleMOMilestone.hasBeenAchieved = true and InvoiceMOMilestone.lastToggled > CompleteMOMilestone.lastToggled and AssembleMOMilestone.lastToggled > CompleteMOMilestone.lastToggled
CompleteMO	if InvoiceMOMilestone.hasBeenAchieved = true and ReceiveItemsMilestone.hasBeenAchieved = true and InvoiceMOMilestone.lastToggled > CompleteMOMilestone.lastToggled and ReceiveItemsMilestone.lastToggled > CompleteMOMilestone.lastToggled and ReceiveItemsMilestone.lastToggled > AssembleMOMilestone.lastToggled and quality = notacceptable
CloseMO	on CompleteMOMilestoneAchieved()
CloseMO	on ReassignSupplierMilestoneAchieved()

Figure 15: The guards generated for the GSM model in Fig. 5

model $m_p.lastToggled$ which gives the latest time stamp when m_p was achieved or invalidated (i.e. toggled its state).

Therefore the expression looks as follows:

$$executedAfter(t_p, t_s) = m_p.hasBeenAchieved \wedge (m_p.lastToggled > m_s.lastToggled).$$

In other words, the milestone m_p of t_p is achieved and it was last toggled after the milestone m_s of t_s . Here we rely on the fact that the milestone of a stage will be invalidated as soon as the stage is reopened. This is ensured by including an invalidating sentry for each milestone.

If for all members of $Alt(t_p, t_o)$ $executedAfter(t_p, t_s)$ is true then the token is still available to enable t_o . We express that as follows:

$$tokenAvailable(t_p, t_o) = \bigwedge_{t_s \in Alt(t_p, t_o)} executedAfter(t_p, t_s).$$

Of course, as discussed in the previous section, other tokens produced in different branches of the Petri net might also be needed for enabling t_o as well as the relevant branch conditions need to be true in order for t_o to fire.

As an example, consider transitions $t_o = \text{CompleteMO}$, $t_p = \text{ReceiveItems}$ and $t_s = \text{AssembleMO}$,

$$\begin{aligned} tokenAvailable(t_p, t_o) &= executedAfter(t_p, t_s) \wedge executedAfter(t_p, t_o) = \\ &= m_p.hasBeenAchieved \wedge (m_p.lastToggled > m_s.lastToggled) \wedge \\ &\quad \wedge (m_p.lastToggled > m_o.lastToggled), \end{aligned}$$

in other words, **ReceiveItems** was executed after the last execution of **AssembleMO** and after the last execution of **CompleteMO**, i.e., the token in the connecting place has not been consumed yet.

This format is more general than the simple format and can be used in all cases. However it is less intuitive and therefore should be replaced by the simple format wherever possible. A third format of intermediate complexity can also be used which requires additional analysis of the Petri net for example for discovering the presence and location of cycles. This format is not presented here in order to simplify the discussion.

Using the proposed approach for translating the Petri net in Fig. 3 to GSM, we generate the model in Fig. 5 with guards as listed in Fig. 15

7 Conclusions

This paper presented a chain of methods for discovering artifact lifecycles which decomposes the problem in such a way that a wide range of existing process discovery methods and tools can be

reused. The proposed tool chain allows for great flexibility in choosing the process mining methods best suited for the specific business process.

The presentation concentrates mostly on the artifact lifecycles. Additionally, the artifact information model can be built from the logs by extracting the data attributes for each event type of the artifact. Existing tools such as [22] can be used to mine data-dependent conditions for the guards based on the discovered information model.

The methods in this paper generate a flat model where no hierarchy of stages is used. Future work will also consider methods for stage aggregation. One possible solution is to use existing algorithms for process abstraction (e.g. [3, 13]) for business process models and translate the discovered process hierarchy to GSM stage hierarchy. For example the Refined Process Structure Tree [16] can be a first step to discovering such a hierarchy.

Future work will also develop methods that allow to discover the interactions between artifacts and thus multi-artifact GSM models can be generated such that instances of different artifacts can synchronize their lifecycles in various ways.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no 257593 (ACSI).

References

- [1] Adriansyah, A., van Dongen, B.F., van der Aalst, W.M.P.: Towards Robust Conformance Checking. In: Muehlen, M.z., Su, J. (eds.) Business Process Management Workshops. LNBI, vol. 66, 122-133. Springer, Heidelberg (2011).
- [2] Bose, R.J.C., van der Aalst, W.: Trace Alignment in Process Mining: Opportunities for Process Diagnostics. In Business Process Management, vol. 6336, Springer Berlin / Heidelberg, 227-242 (2010).
- [3] Bose, R.P.J.C., Verbeek H.M.W., van der Aalst, W.M.P.: Discovering Hierarchical Process Models using ProM. In: Proc. of the CAiSE Forum 2011, London, UK (CEUR Workshop Proceedings, Vol. 734, 33-40)
- [4] Buijs, J., van Dongen, B., van der Aalst, W.: A genetic algorithm for discovering process trees. In: Proceedings of the 2012 IEEE World Congress on Computational Intelligence (2012).
- [5] Cohn, D., Hull, R.: Business artifacts: A data-centric approach to modeling business operations and processes. IEEE Data Eng. Bull., 32, 3-9 (2009).
- [6] Cook, J. E., Wolf, A. L.: Automating Process Discovery through Event-Data Analysis. In: Proceedings of the 17th international conference on Software engineering, New York, NY, USA, 73-82 (1995).
- [7] De Medeiros, A., Weijters, A., van der Aalst, W.: Genetic Process Mining: an Experimental Evaluation. Data Mining and Knowledge Discovery, vol. 14, no. 2, 245-304 (2007).
- [8] Fahland, D., De Leoni, M., Van Dongen, B. F., van der Aalst, W. M. P.: Many-to-many: Some observations on interactions in artifact choreographies. In: Proc. of 3rd Central-European Workshop on Services and their Composition(ZEUS), 9–15. CEUR-WS.org (2011).
- [9] Fahland, D., van der Aalst, W.M.P.: Simplifying Mined Process Models: An Approach Based on Unfoldings. In: Proc. of BPM, 362–378 (2011).

- [10] Fahland, D., van der Aalst, W.M.P.: Repairing process models to reflect reality. In: Business Process Management 2012, volume 7481 of Lecture Notes in Computer Science, 229-245. Springer (2012).
- [11] Feldman, P., Miller, D.: Entity model clustering: Structuring a data model by abstraction. The Computer Journal, 29(4), 348–360 (1986).
- [12] Günther, C., van der Aalst, W.: Fuzzy Mining Adaptive Process Simplification Based on Multi-perspective Metrics. In: Business Process Management, vol. 4714, Springer Berlin / Heidelberg, 328-343, (2007).
- [13] Günther, C., van der Aalst, W.: Mining Activity Clusters from Low-Level Event Logs, BETA Working Paper Series, WP 165, Eindhoven University of Technology, Eindhoven (2006).
- [14] Hull, R. et al. Introducing the guard-stage-milestone approach for specifying business entity lifecycles. In: Proc. of 7th Intl. Workshop on Web Services and Formal Methods (WS-FM 2010), LNCS 6551. Springer-Verlag (2010).
- [15] Hull, R. et al: Business Artifacts with Guard-Stage-Milestone Lifecycles: Managing Artifact Interactions with Conditions and Events. In: DEBS 2011, 51–62 (2011).
- [16] Johnson, R., Pearson, D., Pingali, K.: The Program Structure Tree: Computing Control Regions in Linear Time. In: Proc. of the ACM SIGPLAN 1994 conference on Programming language design and implementation, 171–185, ACM (1994).
- [17] Maggi, F., Mooij, A., van der Aalst, W.: User-guided discovery of declarative process models. In: IEEE Symposium on Computational Intelligence and Data Mining, 192–199 (2011).
- [18] Murata, T.: Petri nets: Properties, Analysis and Applications. In: Proc. of the IEEE, 541–580 (1989).
- [19] Nigam, A., Caswell, N.S.: Business artifacts: An approach to operational specification. IBM Systems Journal, 42(3), 428-445 (2003).
- [20] Ouyang, C., Dumas, M., Breutel, S. ter Hofstede, A.H.M.: Translating Standard Process Models to BPEL. In: CAiSE'06, 417–432 (2006).
- [21] Pesic, M., van der Aalst, W.: A Declarative Approach for Flexible Business Processes Management. In: Business Process Management Workshops. Springer Berlin / Heidelberg, 169–180 (2006).
- [22] Rozinat, A., van der Aalst, W.M.P.: Decision Mining in ProM. In: S. Dustdar, J.L. Fiadeiro, and A. Sheth, editors, BPM 2006, LNCS 4102 , 420-425. Springer-Verlag (2006).
- [23] Rozinat, A., Van der Aalst, W.M.P.: Conformance Checking of Processes Based on Monitoring Real Behavior. Information Systems 33, 64-95 (2008).
- [24] Silberschatz, A., Korth, H. F., Sudarshan, S.: Database System Concepts, 4th Edition. McGraw-Hill Book Company (2001).
- [25] Tiwari, A., Turner, C.J., Majeed, B.: A Review of Business Process Mining: State-of-the-Art and Future Trends. Business Process Management Journal, vol. 14, no. 1, 5-22 (2008).
- [26] van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin (2011).
- [27] van der Aalst, W., Barthelmeß, P., Ellis, C., Wainer, J.: Proclets: A Framework for Lightweight Interacting Workflow Processes. Int. J. Cooperative Inf. Syst., 10(4), 443–481 (2001)

- [28] van der Aalst, W., De Medeiros, A., Weijters, A.: Genetic Process Mining. In Applications and Theory of Petri Nets 2005, vol. 3536, Springer Berlin / Heidelberg, p. 985 (2005).
- [29] van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering, 16(9), 1128–1142 (2004).
- [30] van Dongen, B. F.: Process Mining: Fuzzy Clustering and Performance Visualization, In: BPM Workshops, 158-169, (2009).
- [31] Verbeek, H., Buijs, J. C., van Dongen, B. F., van der Aalst, W. M. P.: Prom: The process mining toolkit. In: Proc. of BPM 2010 Demonstration Track, CEUR Workshop Proceedings, vol. 615, 2010.
- [32] Weijters, A.J.M.M., Van der Aalst, W.M.P.: Rediscovering Workflow Models from Event-Based Data using Little Thumb. Integrated Computer-Aided Engineering, 10(2), 151–162 (2003).
- [33] Weijters, A.J.M.M., Ribeiro, J.T.S.: Flexible Heuristics Miner (FHM). In: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011, IEEE Symposium Series on Computational Intelligence 2011, 310–317 (2011).
- [34] Van der Werf, J.M.E.M., Van Dongen, B.F., Hurkens, C.A.J., Serebrenik, A.: In: K.M. van Hee, R. Valk (Eds.), Applications and Theory of Petri Nets, LNCS vol. 5062, 368–387, Springer (2008).