A Data Model to Apply Process Mining in End-to-End Order Processing Processes of Manufacturing Companies

G. Schuh[1], A. Gützlaff[1], S. Cremer[1], S. Schmitz[1], A. Ayati[1]
[1]Laboratory for Machine Tools and Production Engineering (WZL), RWTH Aachen University, Germany
(s.schmitz@wzl.rwth-aachen.de)

Abstract – **To master ongoing market competitiveness, manufacturing companies try to increase process efficiency through process improvements. Mapping the end-to-end order processing is particularly important, as one needs to consider all order-fulfilling core processes to evaluate process performance. However, process mapping in manufacturing companies is mostly applied in partial processes and not on the end-to-end order processing. Process mining provides a data-based description of processes and their performance and thus objectively and effortlessly maps real as-is processes. The data basis for process mining is an event log. Nevertheless, the generation of an event log in end-to-end order processing is not as trivial as in partial processes, as different data from different information systems must be merged. This paper discusses the development of a data model through an Action Design Research (ADR) method, derived and validated across ADR-cycles. The data model presents, which data can be extracted from integrated database sources to create the required event log for process mining in end-to-end order processing.**

Keywords – **Process Mining, Process Performance, Data Model, End-to-End Order Processing Process**

I. INTRODUCTION

Today, manufacturing companies emphasize the importance of process performance (PP) to continually improve the quality of their products and services to stay ahead of the competition [1]. PP describes how processes perform effectively and efficiently and serve as a precondition for analyzing and improving business processes [2]. Thereby, assessing PP is an essential means in order to identify strengths and weaknesses as well as managing the interacting business processes to realize the companies' goals of sustainable process improvements.

For manufacturing companies, the improvement of the end-to-end order processing process (ETEOPP) from initial customer order to customer order fulfillment is significant, as underperformance results in deterioration of companies' success. The ETEOPP comprises all technical-operative core processes of a company such as sales, procurement and manufacturing. Many empirical findings specify that to improve the ETEOPP, the mapping of the ETEOPP and its interrelations is crucial in order to determine useful improvements for PP increases. By mapping, the order of work activities, their inputs and outputs, as well as the operating processes necessary for PP are determined [1]. Traditional methods, such as workshops or interviews, can be used for ETEOPP mapping but only with limitations, wherefore process performance improvements fail in up to 70% of the cases [3, 4].

Process mining (PM) can be used to improve process mapping and thus increase PP by effortless, fact-based, objective and dynamic process mapping. PM is a process analysis method that aims to discover, monitor and improve actual processes based on extracted event logs already available in companies' information systems [5]. An event log can be defined as a two-dimensional, column-structured table. Each event within an event log refers to a processing activity, is related to a particular process instance (i.e., orders in terms of case-IDs in the given context) and contains a case-ID, activity and activity timestamp as the minimum required data. A wide range of information systems are utilized for PM, such as enterprise resource planning (ERP) systems or customer relationship management (CRM) systems, to extract needful data from actually executed processes. However, until now, manufacturing companies applied PM mostly in administrative or partial processes, for which the required event logs are stored in one single information system and not in varying sources all along the whole ETEOPP. To link the required data across varying database infrastructures and along with merging of data extracted from databases, a data model for PM in ETEOPP is considered. By doing so, the data model can demonstrate, which sort of process data should be extracted from the databases to generate the event log.

Aiming to enable PM applicable in the ETEOPP, this paper addresses the research question: *How does a data model look like to merge required data from information systems in an event log to apply PM in ETEOPP for process performance increases?* To address this question, an Action Design Research (ADR) was conducted to develop a data model. First, the data model involves the extraction of event logs from different databases for each core process initially, and second, the merging of various process-related event logs into one unified event log that can be used by PM techniques to map the ETEOPP.

The remainder of this paper is structured as follows: First, the related work is examined and the challenges of using PM in ETEOPP are described. Chapter 3 describes the investigated information systems across the ETEOPP. The research methodology is described in chapter 4, from which the data model is derived in chapter 5.

II. RELATED WORK

A preparatory step for PM is the extraction of event logs, which are recorded by the databases. Authors

978-1-5386-7220-4/20/$31.00 ©2020 IEEE

proposed a variety of databases (e.g., ERP and CRM) that are utilized for generating event logs to record the real execution of processes within enterprises [6]. However, obtaining the right event logs from different databases is a challenging issue. The PM manifesto [7] presents some guiding principles that can help to extract event data and to query data to derive process models. In [8], an approach is presented that PM discovery enables when multiple case identifiers correlate to each other in one-to-one, one-to-many and many-to-many relations. Despite the relations between databases and event log acquisition, no substantial approach has been found defining the relations of multiple case identifiers through the data models.

Few investigations addressed process discovery of production processes by unique databases. For instance, [8] proposed PM by using run-time information in terms of process events in the assembly workshop. Based on tracked event logs, the authors derived a process model to define the actual machine performances and their relationships in order to evaluate assembly processes. [9] proposed a PM framework to demonstrate the practical applicability of PM within manufacturing processes. For process analysis, event logs extracted from a manufacturing execution system (MES) to discover the manufacturing process model. [10] investigated more deeply on PM approaches for evaluating production processes. They focused on performance evaluation of block manufacturing processes in a shipbuilding company by PM method. Besides, [11] developed a manufacturing-data-based PM approach to replace manual planning activities by value-adding activities in prototype production and to create and evaluate assembly planning documents automatically. However, most of the available approaches focus on certain parts of the manufacturing processes rather than the whole ETEOPP. Furthermore, an explicit data model, which links the required event log data for PM applications within different databases, is not given in the approaches.

In previous work [1], PM was applied to the ETEOPP of a manufacturing company for the first time to increase PP by merging event logs from ERP-system, MES and CRM-system. However, a data model that is compatible with further ETEOPP use cases was not presented. A data model's importance has been highlighted in research papers to represent the data flow mapping of the applications [12]. Nevertheless, no approach is found that uses a data model to represent the data flow for PM in ETEOPP. Such a data model demonstrates the combination of event logs from different information systems to one log to apply PM.

Ordering of activities for event logs needs to be related to the cases [13]. [14] proposed the term of flattening reality into event logs that pursued to sort orders based on unique order numbers, which belongs to a customer in the given context. Two articles focused intensely on the correlation and merging of event logs. [15] suggested the structural relations of logs for flow analysis by merging several logs. They highlighted the importance of an end-to-end process view to understand the relationships between the merging logs. [16] implied the principle of object-centric logs rather than process-centric logs. Using object-centric logs is helpful when recorded data are categorized and grouped in segregated classes. By this, correlated case events are merged based on the objects' groups in the classes. In doing so, one-to-many and many-to-many relations can perfectly be captured. Nevertheless, there was no extensive research found on providing the approaches for the correlation of event logs with a focus on ETEOPP.

III. INVESTIGATED INFORMATION SYSTEMS WITHIN ETEOPP

Fig. 1 shows an exemplary ETEOPP. The purpose is to deal with different databases along the process and merged the required process data stored in those databases into one single event log. All the data formats, such as database objects with traces and the correlation between them, must be self-contained in a unique form to make it easy to share or exchange the data between other databases [16].

In general, in manufacturing companies, to initiate a customer order, all necessary customer data are stored in one customer-related database, for instance, in a CRM-system. Once the order initiated, the proceeding process starts, where invoices and payments are performed. Afterwards, customer data is forwarded from customer service to the development step by transferring the data from the CRM system to product lifecycle management (PLM) system. Engineering activities start in development process steps, which PLM system assigned to connect to this process step. If required, the data from customer service transfer to the development process. Afterwards, the data from CRM system and PLM system are merged. Once the data merged, a new case-ID is generated, which used as a unique reference number for product order identification. Through the value chain from procurement to distribution processes, the new case□ID stays as a fixed attribute in all data classes.

Subsequently, procurement, inbound logistics and material handlings processes are tracked via ERP systems. Once order-related material is procured, manufacturing processes start. In manufacturing processes, MES is used in addition to ERP systems. Via MES, processes are controlled and resources are planned to increase the operational performance [17]. After manufacturing processes, assembly processes start with product assemblage, product commissioning and product support. Sometimes, assembly is considered as one segregated core process in production companies. At the end of the assembly, the testing phase starts. While testing, it is proven whether the products met the expected requirements in all specifications [18]. ERP-systems

Fig. 1. Exemplary end-to-end order processing process (ETEOPP) of manufacturing companies.

record all the changes and activities within the assembly. Finally, for the distribution process step, including the packing and loading of products, as well as shipping to the customers, the data of this process step supposed to be stored in ERP-systems.

IV. RESEARCH METHODOLOGY

A systematic approach based on an ADR is adopted to derive and validate the data model through multiple stages. ADR is described as a research method "to generate prescriptive design knowledge through learning from the intervention of building and evaluating an artefact in an organizational setting to address a problem" [19]. Cyclical and reiterative stages are demonstrated in the ADR method that are (1) formation of problem (2) building, intervention and evaluation and (3) reflection and learning as well as the formalization of learning [19]. According to Fig. 2, five ADR cycles were conducted over four months to derive a reference data model to enable PM in ETEOPP.

The interdisciplinary ADR team includes nine participants with required domain-knowledge on industrial practice, process mining and business process management as well as one small and series machinery company to validate the process model in industrial practice. Each of the cycles is outlined further below.

The first cycle starts by proving the significance of applying PM in ETEOPP for performance improvements and the conceptualization of a first data model. The decision proposed the ETEOPP business units in the data model that should consider as classes, the information systems as databases and the data requirements as attributes. As a result, the relevance of a data model for PM in ETEOPP was proofed and the initial model confirmed, which was detailed in further cycles.

In the second cycle, all requirements of the technical-operative core processes that need to be covered were investigated. Through the negotiations with the ADR team, it was agreed to obtain information on ETEOPP from research partnership. For each core process, it was necessary to define the associated sub-processes. The standard reference DIN 8580 was used for the generation of the manufacturing sub-processes since the data model should not fit a specific production company and should be as generic as possible. Thus, the data model was further detailed, including core processes (classes) and related sub-processes (subclasses).

In the third cycle, the data model has been specified by including the necessary information systems required for data extraction to apply PM in ETEOPP. Through the discussion with the ADR team, it was decided to focus on the information systems that are used in the different core processes. Through literature research, the predefined information systems have been identified that are aligned with the research scope. In particular, CRM, PLM, MES and ERP were subsequently confirmed as sufficient by the whole ADR team. Moreover, the link between information systems and ETEOPP was established. At the end of this cycle, the information systems to be considered in the data model and their assignment on the respective core process units were determined.

Cycle four aimed to make the data model more robust, wherefore the focus was to extract the required data obtained by information systems from the business processes. Based on process mining expert investigations, minimum data requirements for the application of PM were defined as case-ID, activity and timestamps. To further evaluate process performance by PM, the data requirements to measure process time, process cost and process quality were discussed within the ADR team. By doing so, process cost is determined by multiplying the processing time of each process activity by cost drivers and the process quality is calculated based on process loops of the resulted process model, wherefore no additional data were required in the data model. For data extraction, the required relationships between the information systems and the sub-processes in the data model were depicted.

In the fifth cycle, the proposed data model was finalized. The problem formulation concerned to unify correlated event logs, which is explicitly required to map ETEOPP by PM. Through discussion, an additional fixed unique case-ID was established, which can be used across the ETEOPP. Additionally, the approach to establish a new case-ID was validated in practice by the industrial partner and it was proved that the serial number of the designed product could be used for this purpose.

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6+ (Future)
1. Problem formulation	Lack of investigation on the applicability of PM for improving the process performance of the ETEOPP of production companies.	Based on previous cycle, all required technical-operative core processes need to be covered.	Based on previous cycle, different sorts of required information systems need to be identified and the linkage between information systems and ETEOPP must be defined.	Based on previous cycle, the data to be obtained must be known to apply PM in ETEOPP. Besides that , the required data to describe the process performance questioned.	It must conduct an application case study to evaluate the applicability of PM in real ETEOPP of additional production company.	It must conduct an application case study to evaluate the applicability of PM in real ETEOPP of additional production company.
2. Building, intervention and evaluation	1. Discussion with ADR team on the necessity of applying PM in ETEOPP 2. Research study and ADR team validation of data model elements 3. An initial data model is investigated	1. Initializing ETEOPP construction from research partnerships 2. Discussion with the ADR team on the ETEOPP 3. Initializing sub-processes of the ETEOPP 4. Discussion with the ADR team on the sub-processes for the ETEOPP	1. Discussion with research associate about selecting the information systems based on core processes. 2. Predefined information systems have been identified through literature research 3. ADR team suggested few information systems 4. Research associate suggested to establish the link between information systems and ETEOPP	1. Based on our knowledge and literature research, the required data for PM was identified 2. Data scientist proof the required data 3. ADR team agreed on, some of the process performance evaluation can be obtained by the specified minimum required data.	1. Discussion with research associate about making the correlation between event logs that created through ETEOPP 2. Agreed with ADR team on making a fix unique case-ID for each log.	1. Detailed validation with small series of machinery company 2. Mapping of the ETEOPP by getting advices from industrial manager and project manager
3. Reflection and learning	Investigation of PM in ETEOPP was proofed. The initial model confirmed to investigate.	The model included with six core processes (classes) and eighteen sub-processes (sub-classes).	Data model includes CRM, ERP, PLM and MES, which considered as the required information systems. The linkage between information systems and ETEOPP identified.	Minimum data requirements are shown as the attributes in the model.	Finding out the necessitate changes require to comprehensively valid the data model.	Finding out the necessitate changes require to comprehensively valid the data model.

Fig. 2. Overview of ADR cycles.

V. DATA MODEL FOR PROCESS MINING IN ETEOPP

A UML (Unified Modeling Language) data model is derived to represent the data for applying PM in ETEOPP. From a performance viewpoint, the parameters in this data model are considered for preparing an appropriate model for process discovery. Specifically, the parameters are the number of classes and subclasses, the number of attributes (i.e., data), multiplicities of classes and distribution of attributes to the classes. According to Fig. 3, the classes and subclasses of each process step clustered based on the appropriate core process steps.

As stated, a class in this data model expresses a core process step, whereas subclasses relate to sub-processes. Subclasses are in contact with their related classes in a discrete cluster. Attributes of subclasses of the same cluster are aggregated to attributes of the specified class. Therefore, the data of each class consists of the data set of its connected subclasses. The databases in the model are linked to classes in one or many connections. Standardized attributes for each class or subclass include the following attributes: case-ID, activity and timestamp. Additional attributes for each class or subclass consider context-related data for order identification purposes and to identify the activity explicitly.

Additionally, the required data for PP analysis according to performance indicators are considered. The first substantial PPI is process time that requires the start and ends timestamps as data. Thereby it depends on the information system, whether both timestamps or either start/end timestamps are logged [20]. However, in this paper, assumed that each process log includes starts and ends timestamps. Second, process costs can be determined based on the processing time by multiplying each process time with corresponding cost drivers. Third, process quality is analyzed by the resulting process model as the ratio of the scheduled orders with process loops and the total completed orders [21]. Therefore, no additional attributes for process cost and process time are required as data in the process model.

According to Fig. 3, the data for the first class, which is ‘Customer Service’, must be taken from the CRM-system to define the class and subclass elements. In the data model, two subclasses allocated for customer service, which is intended to be identified and used as the track of customer requests. In these subclasses, in addition to the minimum available attributes, further attributes are included as customer name, product name and quantity number, which are required for initiating the customer’s order. The data stored in the CRM-systems can be transferred to the PLM-system that connected to the class ‘Development’. The purpose is to transfer the customer requests data to the development process stage to tailor the designs and development according to customer demands. To be able to analyze and control the ETEOPP, considered establishing one unique reference number (new case-ID) from the procurement process until the end of the production process. The unique reference number has been done through creating an ‘element relation’, which includes gathering the attributes from two processes of ‘customer service’ and ‘development’.

Hence, the data from customer service and development stages aimed to merge, and from their data combination, a new case-ID is establishing, which purposed to transferred to ERP-system for further progressing processes. Afterwards, a new case-ID set as an additional attribute for the rest of the classes for order product identification.

VI. DISCUSSION AND FUTURE RESEARCH

PM provides a proven method to fact-based map real processes and to derive lasting potentials for process improvements. The presented data model enables manufacturing companies to create process transparency of the whole ETEOPP by using PM. Through the data model, the proceeding of extracting the process-related data shows how PM can be used for operational support in order to provide knowledge about the actual running ETEOPP. By following the attributes, the data model addresses the data that needs to be measured and

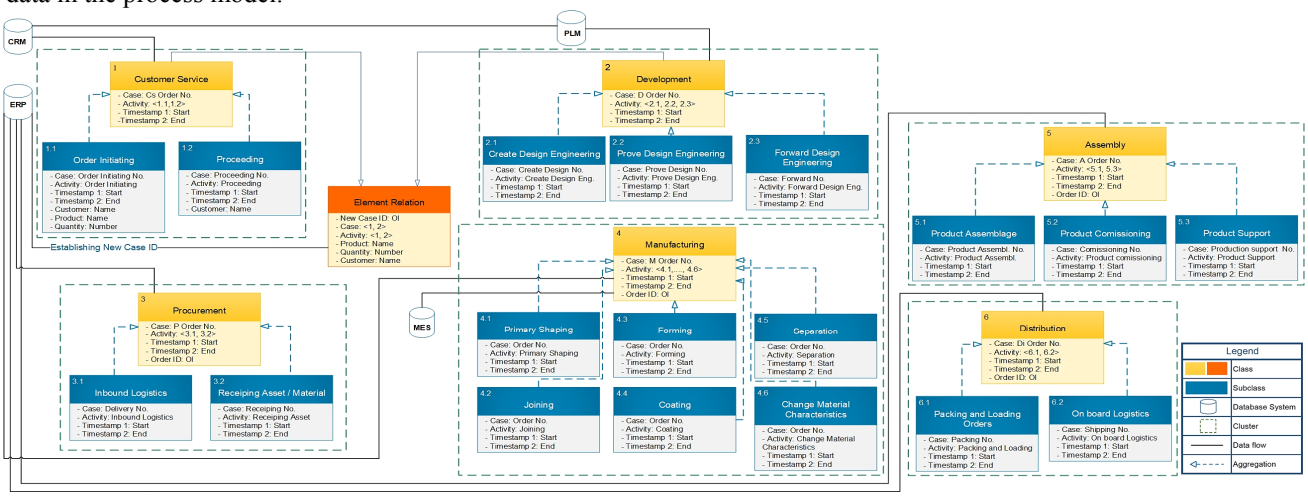

Fig. 3. Data model to apply PM in ETEOPP.

aggregated by the company information systems (such as ERP or MES) to create real process transparency for fact-based PP improvements. It also addresses the challenge of merging different business-unit-specific orders (e.g., customer order, development order, production order) by element relations to adopt PM to cross-functional processes. By providing the processes and sub-processes, the data model has the right level of granularity to be generically adopted to different producing industries. The data model can be transferred to every manufacturing company that wants to apply PM in ETEOPP and is a guideline for data gathering, extraction and merging, which is the first constraint for every PM application to improve business processes.

Besides the aforementioned business benefits, the developed data model covers limitations. Although the ETEOPP was derived by literature-based reference processes, process experts and small machinery company, it should be validated by further manufacturing companies. The ETEOPP and corresponding sub-processes might vary due to different company-specific order penetration points (e.g., engineer-to-order, make-to-order, assemble-to-order, make-to-stock) that change the data attributes of activities. Additionally, further validation proofs whether in practice, all required in the respective sub-processes of the ETEOPP have similar semantics and are therefore unambiguously reflected in the data model. It is also possible to find out in more detail which attributes can be added to the global event log to meet company-specific requirements. Hence, more work is needed to describe additional process performance indicators by PM based on the data model. As a result, the proposed data model should be empirically evaluated by other use cases in the future. Based on the data model, a methodology for the application of PM in ETEOPP should be further researched, which serves as a guideline for companies to increase process performance.

ACKNOWLEDGMENT

Funded by the Deutsche Foschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production – 390621612

REFERENCES

[1] G. Schuh, A. Gützlaff, S. Schmitz, WMP. van der Aalst, "Data-based description of process performance in end-to-end order processing". CIRP Annals, 69(1), 2020.

[2] M. Dumas, M. La Rosa, J. Mendling, HA. Reijers, Fundamentals of Business Process Management. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018.

[3] D. Knoll, G. Reinhart, M. Prüglmeier "Enabling value stream mapping for internal logistics using multidimensional process mining". Exp. Syst. with Applications, 124, pp. 130–142, 2019.

[4] S. Park, YS. Kang "A Study of Process Mining-based Business Process Innovation". Procedia Computer Science 91, pp. 734–743, 2016.

[5] G. Schuh, A. Gützlaff, S. Cremer, M. Schopen "Understanding Process Mining for Data-Driven Optimization of Order Processing." Conference on Learning Factories, Procedia Manufacturing, 2020.

[6] M. Jans, "From relational Database to valuable event logs: From Process Mining Purposes: A Procedure." 2014.

[7] W. van der Aalst, A. Adriansyah, AKA. Medeiros et al. "Process Mining Manifesto." In: F. Daniel, K. Barkaoui, S. Dustdar. Business Process Management Workshops, vol 99. Springer, Berlin, pp 169–194, 2012.

[8] WD. Sunindyo, T. Moser, D. Winkler et al. "Process Analysis and Organizational Mining in Production Automation Systems Engineering." 2010.

[9] SY. Son, B. Yahya, B. Nurgroho et al. "Process Mining for Manufacturing Process Analysis: A case Study", 2015.

[10] J. Park, D. Lee, J. Zhu "An integrated approach for ship block manufacturing process performance evaluation: Case from a Korean shipbuilding company." Intern. Journal of Production Economics 156: pp. 214–222, 2014.

[11] G. Schuh, JP. Prote, A. Gützlaff, S. Cremer, S. Schmitz "Process Mining in Prototype Production." ZWF Vol. 114, No.11, pp. 707-710, 2019.

[12] BF. van Dongen, WMP. van der Aalst, A Meta Model for Process Mining Data, 2014.

[13] S. Suriadi, R. Andrews, AHM. ter Hofstede et al. "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs." Information Systems 64: pp. 132–150, 2017.

[14] WMP van der Aalst, Process Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[15] L. Raichelson, P. Soffer, E. Verbeek E "Merging event logs: Combining granularity levels for process flow analysis." Information Systems 71: pp. 211–227, 2017.

[16] G. Li, E. Murillas, R. Medeiros de Carvalho, WMP van der Aalst "Extracting Object-Centric Event Logs to Support Process Mining on Databases." CAiSE 2018, pp. 182-199, 2018.

[17] AC. Deuel "The benefits of a manufacturing execution system for plantwide automation." ISA Transactions 33(2): 113–124, 1994.

[18] TME. Zaal "Integrated design and engineering: As a business improvement process", 1st ed. Maj Engineering, 2009.

[19] AM. Petersson, J. Lundberg "Applying Action Design Research (ADR) to Develop Concept Generation and Selection Methods". Procedia CIRP 50: pp. 222–227, 2016

[20] RPJ. Chandra Bose, RS. Mans, W. van der Aalst "Wanna Improve Process Mining Results?" IEEE Symposium on Computational Intelligence and Data Mining, 2013.

[21] B. Sarkar, LE. Cárdenas-Barrón, M. Sarkar et al. „An economic production quantity model with random defective rate, rework process and backorders for a single stage production system." Journal of Manufacturing Systems 33(3): pp. 423–435, 2014.