Process Data Science for Workflow Optimization in Digital Pathology: A status report

Patrick Stünkel[1], Sabine Leh[1,2] and Friedemann Leh[1]

[1]Department of Pathology, Haukeland University Hospital, Bergen, Norway
[2]Department of Research and Development, Haukeland University Hospital, Bergen, Norway

Abstract

Pathology is the study of causes and effects of diseases. It is an integral part of medical diagnostics based on the microscopic analysis of tissue, cells, or body fluids. Like other medical disciplines, pathology is currently undergoing a "digital transformation", i.e., witnessing a transition from the assessment of physical tissue slides under a microscope towards analysing digital images of the same tissue slides on a computer screen. The recent advent of powerful machine learning methods and tools for digital image analysis opened the door to novel ways of conducting pathological diagnostics. Still, in order to yield a digital image of a specimen, the specimen has to pass through an elaborate multi-stage preparation process in the laboratory. We argue that in order to achieve a holistic framework of digital pathology, one must not only consider digital image analysis techniques, but also consider means for analysing the process as a whole. Concretely, we propose to analyse the event log data of the laboratory information system in order to understand flow patterns of specimens, find bottlenecks, predict the amounts of incoming samples, and plan resource allocations in an optimal manner. This is highly relevant to meet the ever-increasing number and complexity of specimens, that are handled by pathology departments around the world. The data science method working with event data is called process mining. Process mining is a relatively young but growing research discipline that seeks to bridge the gap between classic data science and business process management. It enables the discovery of control flow structures, data flow patterns, resource utilization, process performance, and more. This paper represents a report on the current state of a work-in-progress project on process mining at a large regional hospital in Western Norway. The main contribution of this report is a list of concrete challenges that we encountered when conducting process mining project in pathology, some of which, we believe, have received less attention in the literature so far. Concretely, we find current process mining techniques not perfectly suited to be directly applied to the pathology laboratory process.

Keywords

Pathology, Process Mining, Workflow Modelling, Event Log Data, Process Analysis, Data Management

1. Introduction

"*Good health and well-being*" is the third of the United Nations' 17 sustainable development goals. Health care is an integral public service that every government around the globe has to provide for its population. A growing and increasingly older population combined with the

The International Health Data Workshop (HEDA-2022), June 19–24, 2022, Bergen, Norway

Patrick.Stunkel@helse-bergen.no (P. Stünkel); Sabine.Leh@helse-bergen.no (S. Leh); Friedemann.Leh@helse-bergen.no (F. Leh)
https://past.corrlang.io/ (P. Stünkel); https://www.uib.no/en/persons/Sabine.Leh (S. Leh)
0000-0002-0877-7063 (P. Stünkel); 0000-0002-0764-2941 (S. Leh); 0000-0002-5986-7114 (F. Leh)

limited availability of trained medical personnel exacerbates the delivery of such health care services, e.g., in its 2013 report the OECD states that health care stands for roughly 10% of the gross domestic products of its member states, and it is expected that this number will grow even further in the future [1]. *Information and communication technology (ICT)* is seen as an opportunity to leverage the aforementioned issue by supporting health care professionals in their daily work and by offloading repetitive, and thus automatable, tasks onto machines in order to utilize the limited human resources more efficiently [2]. A traditional application of ICT, are *information systems* [3, 4], which make the "right" information, at the "right time", available to the "right people". A well-known example of health care information systems are *electronic health record* systems [5]. Another application of ICT in healthcare lies in the area of *computer aided diagnostics*. The latter is mainly facilitated by recent breakthroughs in the field of *artificial intelligence/machine learning (AI/ML)* in the context of medical image analysis [6, 7].

Pathology is a diagnostic medical discipline that, through microscopy of tissues, cells, and fluids, often in combination with molecular diagnostics, determines the presence of diseases as well as morphological and molecular abnormalities. With the increasing availability of so-called *whole slide scanners* and *image viewing software*, also pathology becomes more and more *digitized* [8], i.e., the examination is not performed under a microscope anymore but on a computer screen. The latter enables the application AI/ML methods for automatic image analysis [9, 10]. Still, in order to arrive at a diagnostic result, specimens have to undergo an elaborate preparation process before. While "classical data science" methods are commonplace in image analysis (classification, clustering, etc.), "process data science" is less prevalent. The latter is also known as *process mining* [11], i.e., the discovery of process models from event log data. There are several reports on successful applications of process mining in healthcare, see [12] for a comprehensive survey, but none for pathology in particular.

The goal of this paper is to present an ongoing project at the pathology department of a regional hospital at the west coast of Norway, namely *Haukeland universitetssjukehus*, in the following abbreviated as *HUS*. In this project, the present authors are applying process mining techniques to the preparation workflow of specimens in the pathology laboratory in order to analyse cycle times, detect possible bottlenecks, and, in the long run, optimize the flow times of the samples. The project is still in an early stage, but already from first experiences, we can report on some issues that have received less attention in the process mining literature. Thus, the goal of this paper is to shed more light on the possibilities of process mining in pathology, the intricacies that arise on the organizational, methodical, technical, and social level when conducting such a project, as well as to present our approach on addressing these problems.

The paper is structured as follows: Section 2 introduces the problem and solution domain of this project, namely pathology and process mining. Afterwards, section 3 presents the project itself, its context, and goals. Section 4 presents the challenges, which we have been facing until now, those we are facing right now, and those we are expecting to encounter in the future. Section 5 presents our approach to one of our main challenges, i.e., managing sensitive and heterogeneous data. Eventually, section 6 concludes this paper.

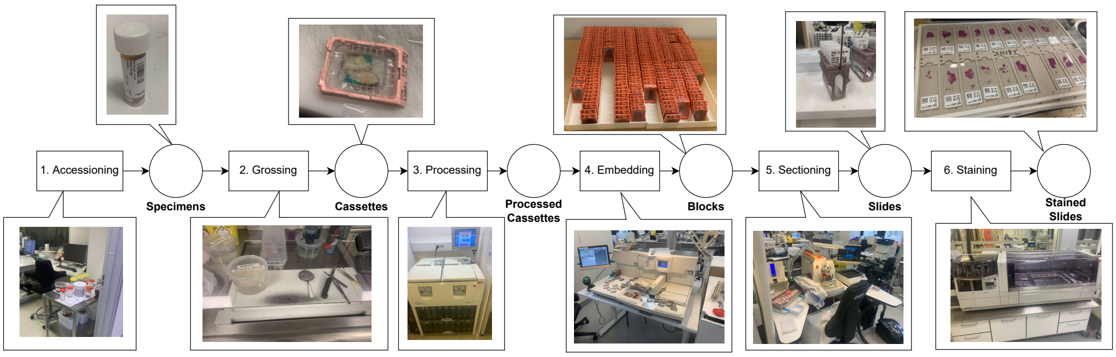


Figure 1: Specimen Workflow in the Pathology Laboratory

2. Background

2.1. Problem Domain: Pathology

The term *pathology* comprises the two Greek words *"pathos"* (suffering) and *"logos"* (study), hence, literally translates as the "study of diseases". Today, it is understood in a more narrow sense as the "study of causes and effects of diseases". A pathologist takes the role of a consultant towards another clinician, who is exerting primary care to a patient. The primary clinician takes a specimen from the patient, e.g., a tissue sample, and sends it to the pathologist, who examines the specimen and writes a report, most often with a conclusive diagnosis, which will help the clinician on deciding the further treatment, e.g. whether surgery or chemotherapy has to be scheduled. The historical development of pathology is closely related to the historical development of medicine itself and is characterized by several technological breakthroughs. Starting with cultural changes in Europe during the 16th century, *autopsy* (i.e., the examination of human corpses) became possible, elucidating the understanding of the human body, its organs, and the effects of diseases. With the use of the microscope to study body tissues during the 19th century, *histology* a.k.a. microscopic anatomy was established as a discipline. Most recently, methods and techniques within *immunohistochemistry* and *molecular biology* enabled further means to understand and diagnose diseases on the cellular and molecular level. When talking about pathology, one distinguishes between the sub-disciplines *autopsy*, *histology*, *cytology* (analysis of cell specimen), and *molecular pathology*. Here, we will focus on *histopathology*.

In order to yield a *histological slide*, which can be analysed under a microscope, the specimen has to undergo a preparation process. This process is abstractly visualized in Fig.1 in the form of a *petri net* [13]. When a specimen arrives at the pathology laboratory, it is first assigned to a case (*"Accessioning"*), i.e. various metadata (patient data, information about the sample type, clinical inquiries) are aggregated in the *laboratory information system (LIS)*, a priority is assigned, and the specimens are labelled with a lab-internal identifier. In most modern labs, this identifier has the form of an industrial barcode, which leverages electronic tracing throughout the process. When the specimen has been immersed in a fixative solution (e.g., formalin) for a sufficient amount of time, it can be delivered to the next stage of the process: *"Grossing"*.

Here, the tissue is examined on a macroscopic level (i.e., "with the naked eye") for abnormal findings and marked. In case of larger specimens, slices with findings of interest are selected from the specimen. Tissues are placed in a *cassette* and delivered to *"Processing"*. This step is performed by a specialized machine that automates dehydration, clearing and infiltration of the tissue with paraffin wax. Afterwards, the processed tissue is taken to *"Embedding"*. This means that it is placed in molten paraffin wax to form a so-called *block*. The cooled-down paraffin block is mounted on a *Microtome*, which allows cutting very thin slices (~ 3-4μm) from the tissue-paraffin-block. The slices are placed on a glass slide and delivered to the *"Staining"* process step. Here, the slide is put through different chemicals, which amplify contrasts and highlight certain biological structures, e.g. *hematoxylin* stains cell nuclei blue and *eosin* stains cell bodies (cytoplasm) red. Finally, a protective cover-slip is mounted on top of the stained tissue slice and the slide is ready to be analysed by a pathologist.

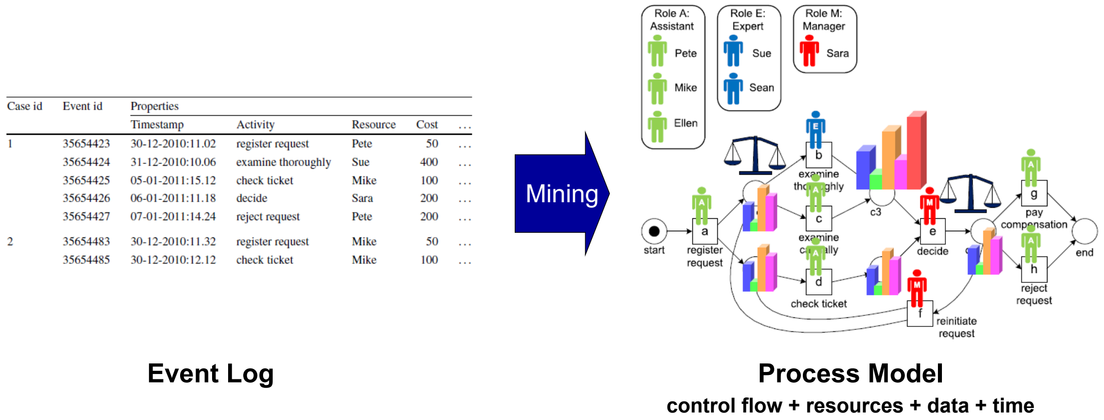
2.2. Solution Domain: Process Mining

Process Mining is a scientific approach that bridges *business process management (BPM)* and *data science*. The former is an interdisciplinary field with roots in Taylor's theory of *"Scientific Management"* [14] and gained significant attention during the 90s when enterprise resource planning software and process-aware information systems were introduced in many organizations [15, 16]. BPM advocates organizing a business around the services that are delivered and the processes that are executed. The associated academic discipline is concerned with all aspects of identifying, analysing, and (re-)designing such business processes. Data science is another interdisciplinary field that brings together statistics, computer science and other related disciplines [17]. Its increasing popularity and significance is mainly due to the abundant availability of "big" data, allowing businesses to gain new insights [18].

While "classic" data science focuses on the derivation of prediction variables (structural features) from a set of given predictor variables, process mining is about the discovery of process models (dynamical features) from event data. Process mining started as a project proposal in the late 90s at Technical University of Eindhoven and has since then grown into its own discipline, with an active community holding conferences and workshops[1]. In terms of publications, [19] and [20] are considered to be the seminal papers in this line of research, while the textbook [11] provides the most recent comprehensive overview over the field.

The principal idea of process mining is sketched in Fig. 2: the base data set is called an *event log*: a collection of events, where each event at least must contain (i) a *case identifier* (to group a set of events w.r.t. to a case), (ii) a *timestamp* (to order the execution of activities within a case), and (iii) the *name of an activity* (to identity the activities within a case). The first step after obtaining an event log is to identify the *control flow* structure of the process model, i.e., the order in which activities may be executed, this is called "play-in" [11]. With a control flow model and an event log at hand, one may do a "replay" [11]. This means to simulate the execution of the event log on the control flow model, which enables *conformance checking* [21], i.e., verifying whether there are deviations between the process model and the event log. Moreover, one is able to discover additional perspectives of a process model. These perspectives

[1] https://icpmconference.org/



Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...

Figure 2: Process Mining schematically, adapted from [11]

are called *data*, *resource*, and *time* [11]. The data perspective highlights how certain properties of a process instance affect the paths that the case takes in the control flow structure. The resource perspective highlights the resources that are required for the execution of certain activities. And, the time perspective looks into the execution times of activities and cases. Hence, process mining does not only encompass the discovery of control flow, but also enables the detection of bottlenecks (performance analysis) and hidden dependencies (data flow analysis).

3. The Project's Goal

According to [22], a (process) data science project may seek answers to one or more of the following questions:

- "What happened?"
- "Why did it happen?"
- "What will happen?"
- "What is the best that can happen?"

These questions can be associated with four activities *report*, *analyse*, *predict*, and *plan*, which are depicted in the upper left quadrant of Fig. 3. These activities also correspond to the sub-goals of our project at HUS. The overarching objective of the project is to reduce the overall cycle time in the pathology department, i.e., the time from receiving a specimen to sending a diagnostic report back. This is a highly relevant concern, because of four key challenges the pathology department at HUS is facing: long cycle times, increasing number of specimens, increasing number of analyses per specimen, and a more or less constant number of resources. In order to understand the first question "What happened?", a precise model of the current process is required. The right-hand side of Fig. 3 shows three different types of models. *Descriptive* models (e.g., process diagrams, statistical indicators, or plots) are simplified representations

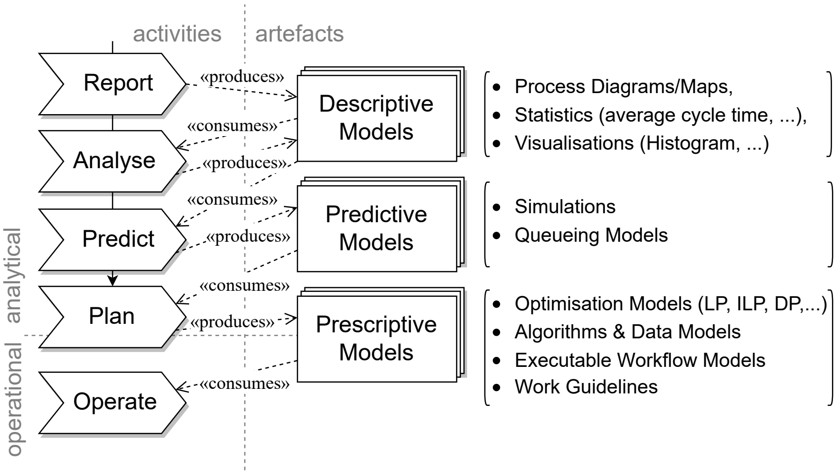


Figure 3: Process Activities and Artefacts

of reality and are a result of the *reporting* activity. *Predictive* models (e.g., simulations) allow making forecasts about the future and are produced in the *prediction* activity based on existing descriptive models. They play an important role for achieving *prescriptive* models. The latter are also called *specifications*. They steer how the actual *operational* tasks are performed. There are many examples of prescriptive models: they range from more abstract work guidelines (e.g., in what order blocks shall be cut on a microtome) over daily plans (e.g., worker assignments to process steps) to the level of concrete machine instructions (e.g., a computer program that routes specimens into different pathways). The ultimate objective of this project is to create such prescriptive models in order to reduce the cycle times in the laboratory.

4. Challenges ...

For the first phase of our project (i.e., reporting), process mining has been selected as a methodology. In this section, we want to report on our own experiences after one and a half years after starting a process mining project in pathology.

In [11], v.d. Aalst presents the so-called L^*-model of process mining. It describes the architecture, stages, and activities of a process mining project. It is motivated by CRISP-DM [23], a cross-industry reference model for conducting data science projects. Fig. 4 contains a graphical depiction of the L^* model, taken from [11], augmented with situations where we experienced resp. expect to experience concrete issues. Our project currently finds itself in stage two of this model. Thus, this section mainly focus on the first three issues.

4.1. ...until now, ...

In the preliminary stage of a L^*-project, one has to justify the purpose of the project and to apply for data access. Since we are conducting our project within the health care domain, there are especially strict requirements concerning access to data: the project had to apply for exemption

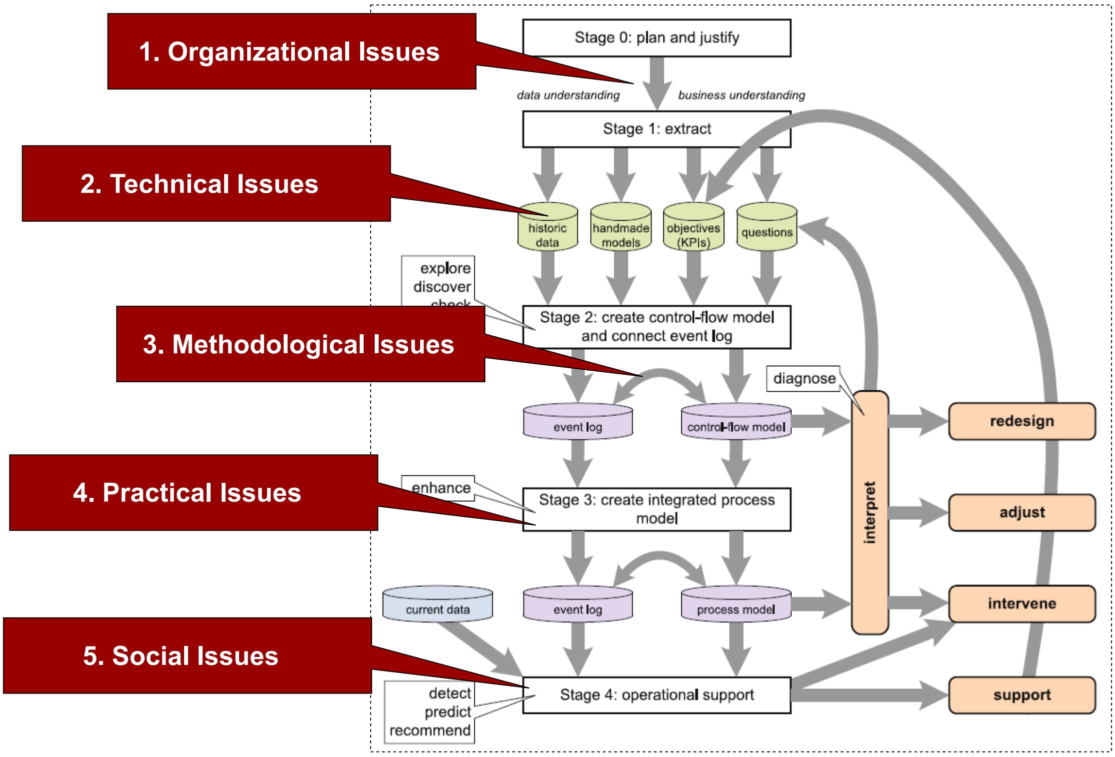


Figure 4: Challenges throughout a Process Mining Project

from the duty of *confidentiality*, to do a *data protection impact assessment*, to carry out a *risk analysis* and to establish a *data management plan*.

Issue #1 (Organizational)

There are cyclic dependencies when writing data access applications.

When writing these applications, we experienced that in order to provide the required documentation about *what* kind of data we need to extract and *how* we are planning to safeguard privacy concerns, extensive knowledge about the database of the laboratory information system was required. To overcome this "chicken-and-egg" problem, it was essential to identify key personalities that have both clearance for accessing the database (because of their regular job description) and a sufficient understanding of the objectives of the process mining project. From our experience, this can be a challenging endeavour because these personalities are often occupied with their operational work. A complementary approach is to have the process data scientists sign respective *non-disclosure agreements (NDAs)*. This requires to already have a legal framework in place for this. Otherwise, juridical personnel has to be involved in the project. In our case, the solution was to employ the primary technical investigator at the hospital.

After applications are approved, the first stage of the process mining project (extraction) is entered. Here, the goal is to obtain event logs, which can be processed by process mining algorithms.

Issue #2 (Technical)

The source information system, generally, does not always offer a viable event log structure.

The concept of an event log has been introduced in Sect. 2.2. In our case, the LIS logs all types of analyses performed on specimens, including histological slide preparation. The main challenge, however, is that the LIS database does directly provide the relevant events. The latter have to be extracted by combining records from several tables. In addition, not every process step is always tracked. For example, in our lab, there is no explicit registration of when the staining of a slide begins (there is only a notification when it is finished). However, it is possible to infer the start timestamp when knowing the staining programme that was executed. In a different situation, i.e., to identify when the grossing or microscopic analysis is started and finished, a separate “user access log” table has to be consulted to retrieve this information. Another issue is that the granularity of the logged events varies greatly, e.g., the system logs some internal function calls, which are not relevant for our analysis. Furthermore, event names in the database are cryptic at times and not unambiguous, which requires combining multiple fields and context information to map event records to real lab actions. Moreover, sometimes case meta-information and resource-specific event attributes are missing (e.g. at what workstation an activity was performed). Bose et al. [24] discuss such “*data quality*” issues and group them into three categories: (i) the event log does not contain events that really happened, (ii) the event log contains more events than in reality, and (iii) the real events are concealed in the log. All three categories apply in our case.

Not cleansing the log would result in unwanted results during the process discovery phase. By conducting several small iterations, where we extracted a small excerpt of raw events from the database, mapped them to an event log, and performed process discovery on it, we could quickly see that a “naive” approach leads to inappropriate results. In our case, it was possible to assess the “quality” of the event log through the resulting control flow model because we have a clear understanding about how the general process should look like, see Sect. 2.1.

Thus, we had to deviate from the principle of “keeping the event data as raw as possible” [25] and to design a transformation from the LIS database structure to an appropriate event log. Designing this transformation, however, required extensive knowledge of both the information system and the domain. To bridge the gap between the domain and IT experts, we made positive experiences by having regular meetings where both sides could exchange their knowledge and by having the IT experts getting direct “hands-on” experience in laboratory. Seeing how the lab technicians work with the LIS, helped immensely in understanding how the system digitally represents the physical actions in reality.

4.2. ...,right now,...

The second stage of L^* describes the transition from an event log to a control-flow process model. This is facilitated by a *process discovery algorithm*.

Issue #3 (Methodological)

The existing process mining algorithms are not perfectly suited for the specimen preparation workflow in the pathology laboratory.

There is a plethora of process discovery algorithms, see [11] for an overview. All of these algorithms are based on the notion of *atomic token*-based workflow modelling languages, i.e., a case is represented as an atomic token that flows through a net structure representing the control flow. The token may become duplicated if activities are performed in parallel but, in general, the case is not decomposed during the execution of the process. In pathology, however, there is a hierarchy of different token types flowing through the lab at the different process steps: a *diagnostic request* (i.e., case) may contain multiple *specimens*, which can become multiple *cassettes/blocks*, which again may result in multiple *slides*. The fact that a pathologist can order additional analyses in between (i.e., creating new blocks and/or slides) requires considering all these artefacts on different levels of granularity at the same time.

When we experimented with the various process discovery algorithms implemented in the open-source tool *ProM*[2], most algorithms produced unwanted results: in most cases they simply returned a control flow where all activities principally could happen in parallel. The *fuzzy miner* [26] algorithm produced yet the "best" result compared to others, in a sense that it discovered the general structure of Fig. 1. Still, the algorithm was not able to discover the correct causal dependencies between less frequent process steps and when decreasing the abstraction level, *"spurious cycles"* appeared on all activities. The latter phenomenon can be explained by the fact that the process steps happen in parallel while operating with different level of granularity.

Different granularity levels are discussed in [11, Chap.5.5], where the aforementioned atomic token abstraction is described as *"flattening"*. The chapter mentions the idea of *"proclets"* [27], i.e., disassembling the overall process into several process operating at different levels of granularity, and refers to a research project (ACSI project) that promotes the use of such proclets. However, the referenced website does not seem to be active any more today.

In our case, we are more or less aware how the control flow must look like. Thus, process discovery algorithms are actually less interesting for us and we can resort to creating a precise process model by hand. The latter is a confirming sign that we are dealing with a so-called *"Lasagna"*-process [11], i.e., a process model with a simple and well-understood control flow. We discovered that *coloured petri nets (CPNs)* [28] are an appropriate formalism for our case, since they naturally model the idea of different types of tokens flowing through the net. Hence, our immediate next objective is to design the pathology lab process in the form of a coloured petri net and to extend the notion of event-log replay on petri nets with the notion of different token types. This is necessary to obtain the performance information of the individual process steps and different types of specimens.

4.3. ...and later

According to the L^*-model, our project is currently in stage two. Yet, we want to give an outlook on the issues, that we are expecting to arise in the coming stages. The third stage is the

[2] http://promtools.org/doku.php

creation of an integrated model, i.e. a process model combining the notions of control flow, data, resources and time. This will, for the first time, allow giving feedback to the original process. The L^*-model discusses several options for this, namely *redesigning* (changing the whole process model), *adjusting* (changing the process configuration, for example, resource allocations), or *intervening* (performing concrete actions during the execution of process instance).

Issue #4 (Practical)

It is not exactly clear how process mining observations can be translated into actions.

We are currently uncertain of how we eventually can transfer the analytical results to operational results. For instance, there are some physical limitations to what degree a “redesign” of the process is possible. The literature mentions approaches on how to transit from process mining to simulation [29]. But, it does not mention specific methodologies for getting to means of operational support, the final stage of L^*.

Issue #5 (Social)

It is not clear how to best anticipate and mitigate social ramifications.

The final objective is to reduce the overall cycle times via intelligent planning of resources and routing of specimens. When automatically assigning tasks to individual workers, both individual skills, individual preferences for particular tasks and the laboratory’s current need for specific activities matter. There is a theoretical possibility to assess the performance data of individual workers. Thus, our project has to safeguard that this contingency remains unfeasible. Currently, we are hashing all usernames with a random and hidden salt. When designing reporting solutions, we have to make sure that performance data is only presented aggregated over multiple cases, such that it is not possible to identify individuals from context information of a single case. In all of this, it is paramount to include all stakeholders in the project to make them aware of the technical possibilities and the data stored in the system. Even though this issue remains in the more distant future, it is important to be aware of it already.

5. Executable Data Management: A Model-based Approach

In Sect. 4 we have seen that the raw event data poses several challenges. First, there is a (organizational) challenge in gaining access to it, which necessitates to document what is extracted and how sensitive data is protected. Second, there is a (technical) challenge when it comes to mapping the raw data into an event log so that it can be used for process mining.

It turns out that *metadata* plays a crucial role when addressing these challenges. They serve both as documentation as well as specification for extraction and transformation. Since it is required to put them under *version control* to enable auditing, revision, and iterative development, one might as well consider utilizing these documents more “directly”. Hence, we decided to adopt a model-based paradigm [30] and consider these artefacts not only as mere means of documentation (descriptive) but also as means to configure the extraction and transformation

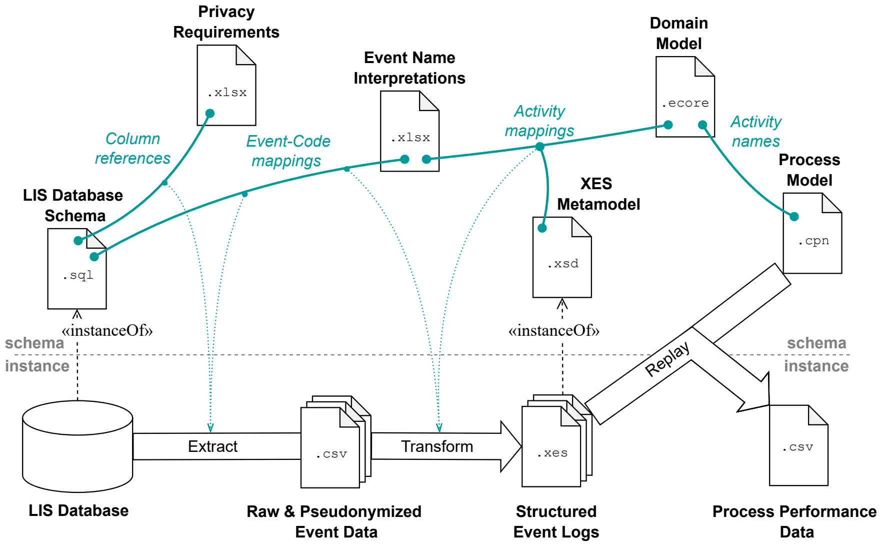


Figure 5: Data Management Architecture

scripts (prescriptive). Here, the model-based paradigm fits particularly well with the necessity for metadata descriptions, i.e. instead of encoding extraction and transformations in program code they are declaratively defined in documents that are accessible for the domain experts.

The resulting architecture is shown in Fig. 5. The bottom half of the figure shows the data layer. The data “flows” from left to right, starting from the LIS database with the raw data. In the first step, the contents of relevant tables are exported in the form of *comma separated values (CSV)* files, where the contents of the columns containing sensitive information are hashed. In the second step, this data is transformed into an event log structure. This transformation step has to address the challenges related to data quality, see Sect. 4.1. Eventually, the event log is replayed on the process model to obtain performance data about case and activity durations.

The top-half of Fig. 5, contains the metadata documents. The database is described via SQL Create Table statements, which were manually extracted from a PDF provided by the LIS supplier. An Excel sheet declares the columns, which are extracted and which column contents are hashed. In our case, Excel turned out to be a viable compromise for a tool that domain experts are familiar with and which, simultaneously, can easily be integrated in automated toolchains. Similarly, the declarations about how event codes from the LIS map to the individual process steps are defined in an Excel sheet. For the latter, we first created a *domain model* of histopathology. The domain model has the form of a *class diagram* and is encoded using *Ecore* [31], a standard serialization format in the context of model-based engineering. Moreover, there is the *extensible event stream (XES)* schema definition [32], which defines a standard for representing event logs, and the process model defined as a coloured petri net, see Sect. 4.2.

All these documents are inter-related because they refer to each other’s elements, e.g. the transition names in the CPN-model must correspond to activity names defined in the domain model. These relations are visualized as cyan-coloured links in Fig. 5.

For the foundation of this infrastructure, we built on CorrLang[3], an academic prototype tool addressing semantic interoperability via mediation, based on a textual domain-specific language, which was developed in the context of the first author's PhD thesis [33]. The tool establishes generic relations (the cyan links) between the various metadata documents, which are interpreted to perform the extraction and transformation on the data level[4].

6. Conclusion

To summarize the content and contributions of this state-of-the-project report: we began by introducing (digital) pathology with an emphasis on not only focusing on classic data science for image analysis, but also consider process data science for event data stored in health care information systems. Concretely, we want to gain insights about the specimen preparation process in the lab, as this constitutes a significant amount of time within the diagnostic process. There are several reports on successful applications of process mining in the health care domain [12, 34]. Also, the *ProHealth* workshop series (now *KR4HC*) offers a significant body of knowledge about applications of process-centric approaches within healthcare. However, to our knowledge, none of these domains have addressed pathology so far.

The main contributions of this paper are (a) an experience report (Sect. 4) about conducting a process mining project in pathology (currently in the *reporting* phase of Fig. 3 and stage *two* of the L^* model), and (b) a conceptual approach for exploiting project documents as an executable specification for a data transformation pipeline (Sec. 5). Our experience report comprises insights that, we believe, have received less attention in the process mining literature. Especially the conceptual mismatch between atomic token-based workflow modelling languages and the flow of specimens/blocks/slides in the pathology laboratory is to highlight here.

Acknowledgments

The present study is part of the project "PiV – Pathology services in the Western Norwegian Health Region: a centre for applied digitization". The project is funded by the Western Norway Regional Health Authority. We also would like to thank the anonymous reviewers for their helpful remarks.

References

[1] OECD, Health at a Glance 2013: OECD Indicators, Organisation for Economic Co-operation and Development, Paris, 2013. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2013_health_glance-2013-en.

[2] OECD, Improving Health Sector Efficiency: The Role of Information and Communication Technologies, Organisation for Economic Co-operation and Development, Paris, 2010. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/improving-health-sector-efficiency_9789264084612-en.

[3] https://www.corrlang.io/

[4] A demonstration of these definitions is found in: https://github.com/webminz/piv-data-mgmt

[3] P. Haried, C. Claybaugh, H. Dai, Evaluation of health information systems research in information systems research: A meta-analysis, Health Informatics Journal 25 (2019) 186–202. doi:10.1177/1460458217704259.

[4] R. S. Mans, W. M. P. van der Aalst, N. C. Russell, P. J. M. Bakker, A. J. Moleman, Process-Aware Information System Development for the Healthcare Domain - Consistency, Reliability, and Effectiveness, in: S. Rinderle-Ma, S. Sadiq, F. Leymann (Eds.), Business Process Management Workshops, Springer, Berlin, Heidelberg, 2010, pp. 635–646. doi:10.1007/978-3-642-12186-9_61.

[5] L. Nguyen, E. Bellucci, L. T. Nguyen, Electronic health records implementation: an evaluation of information system impact and contingency factors, International Journal of Medical Informatics 83 (2014) 779–796. doi:10.1016/j.ijmedinf.2014.06.011.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88. doi:10.1016/j.media.2017.07.005.

[7] J. Ker, L. Wang, J. Rao, T. Lim, Deep Learning Applications in Medical Image Analysis, IEEE Access 6 (2018) 9375–9389. doi:10.1109/ACCESS.2017.2788044.

[8] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, J. Saltz, Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives, Journal of Pathology Informatics 9 (2018) 40. doi:10.4103/jpi.jpi_69_18.

[9] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases, Journal of Pathology Informatics 7 (2016) 29. doi:10.4103/2153-3539.186902.

[10] D. Komura, S. Ishikawa, Machine Learning Methods for Histopathological Image Analysis, Computational and Structural Biotechnology Journal 16 (2018) 34–42. doi:10.1016/j.csbj.2018.01.001.

[11] W. M. P. v. d. Aalst, Process Mining: Data Science in Action, Springer, 2016.

[12] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, Journal of Biomedical Informatics 61 (2016) 224–236. doi:10.1016/j.jbi.2016.04.007.

[13] C. A. Petri, Kommunikation mit Automaten, Doctoral Thesis, Technische Hochschule Darmstadt, 1962.

[14] F. W. Taylor, The Principles of Scientific Management, Harper & Brothers Publishers, New York, 1911.

[15] T. H. Davenport, J. E. Short, The New Industrial Engineering: Information Technology and Business Process Redesign, MIT Sloan Management Review (1990). URL: https://sloanreview.mit.edu/article/the-new-industrial-engineering-information-technology-and-business-process-redesign/.

[16] M. Hammer, Reengineering Work: Don't Automate, Obliterate, Harvard Business Review (1990). URL: https://hbr.org/1990/07/reengineering-work-dont-automate-obliterate.

[17] L. Cao, Data Science: A Comprehensive Overview, ACM Computing Surveys 50 (2017) 43:1–43:42. doi:10.1145/3076253.

[18] T. H. Davenport, D. J. Patil, Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review (2012). URL: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-

21st-century.

[19] W. M. P. van der Aalst, A. J. M. M. Weijters, Process mining: a research agenda, Computers in Industry 53 (2004) 231–244. doi:10.1016/j.compind.2003.10.001.

[20] W. M. P. van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, A. J. M. M. Weijters, Workflow mining: A survey of issues and approaches, Data & Knowledge Engineering 47 (2003) 237–267. doi:10.1016/S0169-023X(03)00066-1.

[21] A. Rozinat, W. M. P. van der Aalst, Conformance checking of processes based on monitoring real behavior, Information Systems 33 (2008) 64–95. doi:10.1016/j.is.2007.07.001.

[22] W. M. P. van der Aalst, Data Scientist: The Engineer of the Future, in: K. Mertins, F. Bénaben, R. Poler, J.-P. Bourrières (Eds.), Enterprise Interoperability VI, Springer International Publishing, Cham, 2014, pp. 13–26. doi:10.1007/978-3-319-04948-9_2.

[23] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0 Step-by-step data mining guide, Technical Report, The CRISP-DM consortium, 2000.

[24] R. J. C. Bose, R. S. Mans, W. M. van der Aalst, Wanna improve process mining results?, in: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2013, pp. 127–134. doi:10.1109/CIDM.2013.6597227.

[25] v. d. Aalst, W.M.P., Extracting event data from databases to unleash process mining, BPM reports, BPMcenter. org, 2014.

[26] C. W. Günther, W. M. P. van der Aalst, Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics, in: G. Alonso, P. Dadam, M. Rosemann (Eds.), Business Process Management, Springer, Berlin, Heidelberg, 2007, pp. 328–343. doi:10.1007/978-3-540-75183-0_24.

[27] W. M. P. Van Der Aalst, P. Barthelmess, C. A. Ellis, J. Wainer, Proclets: a framework for lightweight interacting workflow processes, International Journal of Cooperative Information Systems 10 (2001) 443–481. doi:10.1142/S0218843001000412.

[28] K. Jensen, Coloured Petri Nets, Monographs in Theoretical Computer Science An EATCS Series, Springer, Berlin, Heidelberg, 1997. URL: http://link.springer.com/10.1007/978-3-642-60794-3. doi:10.1007/978-3-642-60794-3.

[29] A. Rozinat, R. S. Mans, M. Song, W. M. P. van der Aalst, Discovering simulation models, Information Systems 34 (2009) 305–327. doi:10.1016/j.is.2008.09.002.

[30] M. Brambilla, J. Cabot, M. Wimmer, Model-Driven Software Engineering in Practice, 2nd ed., Morgan & Claypool Publishers, 2017.

[31] D. Steinberg, F. Budinsky, E. Merks, M. Paternostro, EMF: Eclipse Modeling Framework, Pearson Education, 2008.

[32] IEEE, Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams, IEEE Std 1849-2016 (2016) 1–50. doi:10.1109/IEEESTD.2016.7740858.

[33] P. Stünkel, A framework for multi-model consistency management, Doctoral Thesis, Høgskulen på Vestlandet, Bergen, 2022. URL: https://hdl.handle.net/11250/2837740.

[34] R. S. Mans, W. M. P. van der Aalst, R. J. B. Vanwersch, A. J. Moleman, Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions, in: R. Lenz, S. Miksch, M. Peleg, M. Reichert, D. Riaño, A. ten Teije (Eds.), Process Support and Knowledge Representation in Health Care, Springer, Berlin, Heidelberg, 2013, pp. 140–153. doi:10.1007/978-3-642-36438-9_10.