Audit Trails in OpenSLEX: Paving the Road for Process Mining in Healthcare

E. González López de Murillas[1], E. Helm[2], H.A. Reijers[1,3], and J. Küng[4]

[1] Department of Mathematics and Computer Science
Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Research Department of e-Health, Integrated Care
University of Applied Sciences Upper Austria, Hagenberg, Austria
[3] Department of Computer Science
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[4] Institute for Applied Knowledge Processing
Johannes Kepler University, Linz, Austria
e.gonzalez@tue.nl, emmanuel.helm@fh-hagenberg.at,
h.a.reijers@tue.nl, jkueng@faw.jku.at

Abstract. The analysis of organizational and medical treatment processes is crucial for the future development of the healthcare domain. Recent approaches to enable process mining on healthcare data make use of the hospital information systems' Audit Trails. In this work, methods are proposed to integrate Audit Trail data into the generic OpenSLEX meta model to allow for an analysis of healthcare data from different perspectives (e.g. patients, doctors, resources). Instead of flattening the event data in a single log file the proposed methodology preserves as much information as possible in the first stages of data extraction and preparation. By building on established standardized data and message specifications for auditing in healthcare, we increase the range of analysis opportunities in the healthcare domain.

Keywords: process mining, healthcare, audit trails, meta model

1 Motivation

Process mining provides an a-posteriori empirical method to discover process models from observed system behaviour. By applying the techniques of process mining to the healthcare domain, valuable insights regarding e.g. clinical practice, performance bottlenecks, and guideline compliance can be gained [13]. However, the healthcare domain presents certain challenges to traditional process mining approaches. Healthcare processes are highly dynamic, highly complex, increasingly multi-disciplinary, and, generally, ad hoc [12].

In [6] the authors tried to overcome the problems of distributed, heterogeneous data sources in healthcare by analyzing their minimum common ground – the Audit Trail (AT). Based on internationally agreed upon standards, these ATs contain only basic information about what happened even though these can be enriched to be useful for various business intelligence (BI) analysis.

This work will go further by solving a number of remaining issues of the previous approach (1) by explaining how process mining can be incorporated in the analysis of ATs, and (2) by providing the basis for automated process mining in different contexts via the OpenSLEX meta model [11].

The remainder of this paper is structured as follows: Section 2 introduces some background information. Then, Section 3 discusses the problem that motivates this work. The proposed approach is described in Section 4. Finally, Section 5 concludes the paper and gives a short outlook on future work.

2 Background

As has been mentioned before, the goal of this work is (1) to solve the data issues of the previous approach, and (2) to enable automated process mining on the extracted and transformed data. Before we attempt to tackle these challenges, it is necessary to introduce some background information about the data that we want to analyze, the ATs, and the meta model that we propose to represent the extracted information, OpenSLEX.

2.1 Standardized Audit Trails

The international non-profit organization *Integrating the Healthcare Enterprise* (IHE) aims to improve the integration and interoperability of healthcare IT systems. Founded in 1998 by radiologists and IT experts, it defines how established standards, like DICOM and HL7, can be implemented to overcome common interoperability problems in healthcare. The initial focus was on radiology but nowadays IHE covers use cases in different healthcare domains. Their *integration profiles* are the basis for systems of major vendors and national and international healthcare programs.

One of the basic IHE Integration Profiles dealing with IT infrastructure in healthcare, the *Audit Trail and Node Authentication* (ATNA) profile, defines how to build up a secure domain that provides patient information confidentiality, data integrity, and user accountability [7]. A secure domain can scale from department to enterprise to cross-enterprise size. To ensure user accountability, ATNA specifies the use of a centralized ARR where all Audit Messages are stored. In a joint effort IHE, HL7, DICOM, ASTM E31, and the Joint NEMA/COCIR/JIRA Security and Privacy Committee defined the structure of the Audit Messages using XML schema. The normative specification of the messages is defined in the DICOM standard PS3.15: A.5 Audit Trail Message Format Profile [2]. The original intention of ATNA event audit logging was to provide surveillance logging functions to detect security events and deviations from normal operations. It was not designed for forensic or workflow performance analysis. However, the integration profile states that forensic or workflow analysis logs may also use the same XML schema and IHE transactions [7] and recent developments propose the use of ATNA for keeping track of the whole workflow [3].

Cruz-Correia et al. analyzed the quality of hospital information systems (HISs) Audit Trails in Portugal [4]. They pointed out the potential use of ATs for the improvement of the HIS by applying data mining, machine learning, and process mining techniques. The authors highly recommend the use of standards to record ATs – including ATNA – but criticize the lack of completeness and overall data quality in the data they analyzed.

2.2 OpenSLEX Meta Model

Data extraction and transformation are, very often, the most time-consuming stages of a process mining project. The difficulty to tackle these tasks comes from the variability on data representations in which the original data can be found. Most of the applications of process mining in real-life systems provide ad-hoc solutions to the specific environment of application. Some examples of these systems are SAP [8, 9, 15] and other ERPs [10]. Nevertheless, efforts have been made to develop standards for data representation in process mining. The IEEE XES standard [1] is the most important example, being extensively used both in academic and industrial solutions. However, despite its success at capturing event data in an exchangeable format, something that this standard misses is the data perspective on the original system.

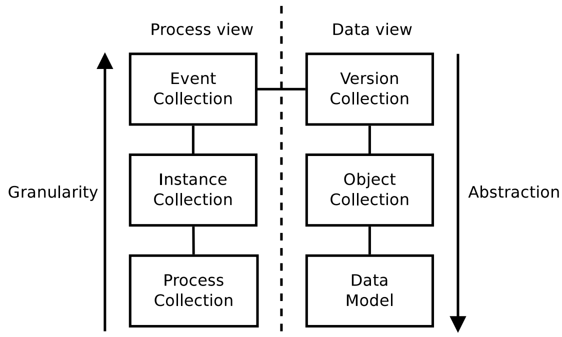

Fig. 1. Diagram of the OpenSLEX meta model at a high level.

With the purpose of mitigating the limitations of current event data representation standards, in previous work we proposed OpenSLEX [11], which provides a meta model that takes into account not only the process view (*events*, *instances*, and *processes*), but also the data view (*data models*, *objects*, and *object versions*). Figure 1 shows a high level description of the meta model, where we see how granularity of data increases inversely proportional to the level of abstraction. In other words, a *data model* is a more abstract representation of the data than the *objects* or the *object versions*, while the latter has a greater level of granularity than the *data model.* The same can be said about the process view, where *processes* are abstract descriptions of the *events*, which are much more granular data. A more detailed description of the meta model is available online[5].

Additionally, the fact that in this meta model the process side is combined with the data side allows to capture a richer snapshot of the system under study. Unlike other meta models, like the one proposed in XES, which requires the existence of a case notion to group events into process instances, OpenSLEX enables the adoption of different perspectives. Events are stored independently of any case notion. Afterwards, one or many case notions can be defined, generating the cases that will group events in different event logs. This is the key to enable multi-perspective process mining on the extracted data. The fact that not a single case notion is enforced during the data extraction phase avoids data loss. Additionally, it enables the application of automated techniques that correlate events in multiple ways to show different processes or perspectives coexisting in the same system.

To summarize, OpenSLEX provides a layer of standardization of the representation of data, while considering both process and data views, unlike other existing flatter event models. This makes it possible to decouple the application of analysis techniques from data extraction and transformation stages. Additionally, it enables smarter ways to analyze the information, considering the data side to enrich the process perspective.

3 Problem

The analysis of healthcare information systems is a challenging task. This is due to the unstructured nature of the processes, the large variety of data schemas and systems used in HISs, privacy issues, etc [12]. These aspects make the application of process mining a time-consuming process. Some efforts have been made in order to standardize the way HISs communicate and transmit event data. The ATNA integration profile by IHE specifies the structure and content of messages needed for auditing. The profile describes how user accountability, especially regarding the use of protected health information (PHI), can be ensured by using a centralized Audit Record Repository (ARR) [7]. Therefore, every access to and transfer of PHI is recorded. Although the approach presented in [6] showed that the information recorded in audit trails is sufficient to enable process mining, several major issues remained unsolved. (1) The approach was not able to automatically identify traces. (2) The manually chosen, fixed trace identifier (the patient ID) lead to snapshot-problems and limited the possible process mining perspectives. (3) The hard-wired mapping of fields from ATNA audit messages to event logs sometimes lead to incorrect mappings.

In addition to tackling the mentioned issues this paper also addresses the topic of BI applications on top of ARRs, specifically in an automated manner. According to the integration profile the ATNA ARR is expected to have analysis and reporting capabilities, but those capabilities are not defined as part of the

[5] https://github.com/edugonza/OpenSLEX/blob/master/doc/meta-model.png

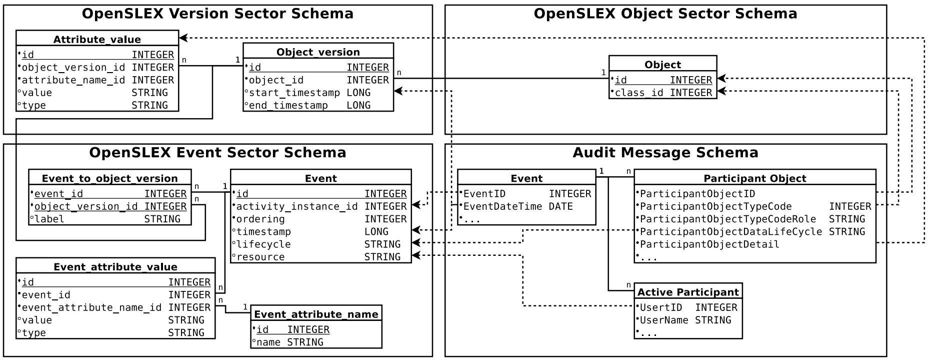

Fig. 2. The dashed lines show the mapping of the fields of Audit Messages to the OpenSLEX meta model.

profile [7]. The question is, *how* can BI applications be put on top of an ARR? And how can this be done in an automated way, assuming the data is stored and managed to be analyzed?

The following section describes how ATNA messages can be mapped into the OpenSLEX meta model, in order to tackle the issues of the previous approach, and the type of data transformation required to enable automated process mining.

4 Approach

The use of ATNA messages in order to extract event data for process mining has been previously demonstrated in [6]. Now we aim at stepping up in the level of generalization, using the OpenSLEX meta model as an intermediate format for data collection. This meta model can be seen as a data schema for a data warehouse, acting as an ARR, capturing event data, together with data objects, data models, and historical data, ready to be exploited by the existing process mining techniques.

4.1 Mapping of ATNA messages to OpenSLEX

The specific characteristics of ATNA messages makes them great candidates for event data extraction. Figure 2 shows how different fields of the ATNA message can be mapped to fields of the OpenSLEX meta model. The minimally required attributes in order to obtain events are activity names and timestamps. These two attributes can be directly mapped to the ATNA message's fields *EventID* and *EventDateTime*, respectively. In addition, *Active Participant* fields such as *UserID* and *UserName* show valuable resource data to enrich the events. However, what makes ATNA messages specially attractive from the process mining perspective is the presence of *Participant Object* data. The fields within this part of the message contain not only object data information such as role (*ParticipantObjectTypeCodeRole*) and life-cycle (*ParticipantObjectDataLifeCycle*), but also object type (*ParticipantObjectTypeCode*) and unique object identifiers (*ParticipantObjectID*), which enable the traceability of data changes and behavior at object level. Additionally, detailed value pair data (*ParticipantObjectDetail*) of the participating object can be present. Such key-value pairs represent a snapshot of the relevant attributes of a participant object at the time of occurrence of the event, which can be seen as an object version. Object versions reveal the evolution of objects through time, and are related to the events that caused the modifications.

Data extraction and transformation are difficult tasks that require a significant amount of domain knowledge to be carried out. It is common that, during this transformation of data, choices are made that affect the final picture we obtain of the system under study. Considering the ATNA message fields we just discussed, we seem to be able to capture event information, which may be mapped into the OpenSLEX corresponding elements. The next section explains the transformation of the captured event data in order to infer new information. This will allow to obtain a more complete picture of the whole system.

4.2 Transformation of ATNA messages

OpenSLEX provides the meta model to shape the extracted information, with the purpose of minimizing the impact of the data extraction and transformation stages on the result of the analysis. Transforming the ATNA messages into this new representation enables the application of process mining without any semantic loss of the original data. This is achieved by considering the data view in addition to the process view, avoiding flattening multidimensional data into simple event logs. Figure 3 shows the steps in the data transformation process in order to capture a picture as close as possible to the original system:

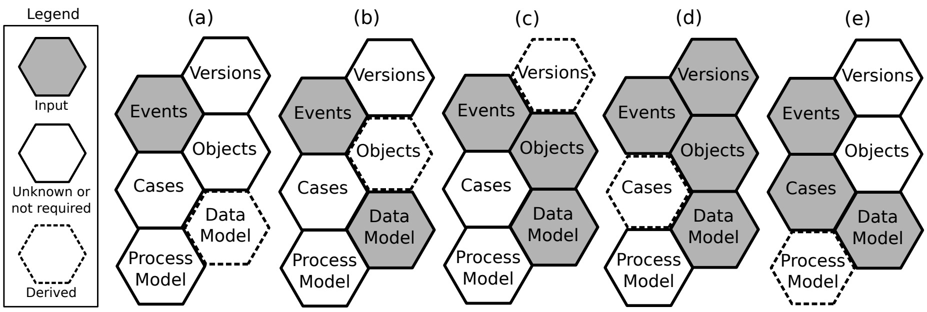

Fig. 3. Inference of the missing elements in the meta model, starting from the events (a) and finishing mining a model (e).

a) Figure 3.a shows the situation in which we only obtained event information from the system. This matches the situation we face when dealing solely with ATNA messages. These messages are, in the end, events emitted by different actors in the healthcare ecosystem under study. These events contain valuable information that will let us infer some of the other sectors of the meta model. Using the information in the *ParticipantObjectTypeCode* field, we can infer the classes involved in the data model of the system. The *ParticipantObjectDetail* key-value pairs provide the information about the attributes of such classes.
b) Figure 3.b represents the next step in which, after discovering the data model underneath the global system, object instances are mapped into it. To do so, the field *ParticipantObjectID* helps us to identify the unique entities that exist for each data class.
c) Figure 3.c depicts the subsequent step, in which we infer the object versions involved in the process. These object versions are object snapshots at a certain moment in time, and they can be reconstructed by matching the key-value pairs in the *ParticipantObjectDetail* field of the ATNA events, to the object id obtained in the previous step (Figure 3.c). Reconstructing these event versions will help us understand the object's life-cycle within the process. And not only that, but applying primary and foreign key discovery techniques [14, 16] will make possible to uncover links between objects belonging to different classes. In the next step we will see how these connections can be exploited to correlate events in different ways.
d) So far we have been able to capture events, infer the data model, and extract object and object version information from the data. However, a case notion is needed in order to group events in cases. These cases will build the event logs necessary for process the mining analysis. It is in this step (Figure 3.d) in which one of the main benefits of this technique arises: the independence of the case notion from the event capturing phase. Events can be correlated in many different ways. One of them is to select a common attribute as a case identifier, which is the most common way to build logs nowadays. However, our meta model gives us an advantage with respect to traditional methods: the existence of links between events and object versions. As has been described in the previous step, relations between object versions can be discovered. This means that objects can point to others, as they do in databases with foreign keys, e.g., a treatment object points to a doctor and a patient. This enables a new way to correlate events that might not share any common attribute (doctor and patient events), by means of a linking object (treatment). The data model structure discovered from the data will determine all the possible combinations (case notions) that can be made in order to build event logs, making possible to have a multi-perspective view on the data.
e) Only when case notions have been discovered, and the logs representing the different views have been built, we can proceed with the process discovery. Figure 3.e shows the step in which process models are discovered using existing process mining techniques. This is the step that enables further analysis

of the data combining process and data views in a centralized and comprehensive manner.

The steps above describe the transformation of ATNA messages into the OpenSLEX format. Sometimes this requires to infer the missing bits of information from the ones available in the ATNA message's attributes. As a result, we obtain a global view of the data and process sides, minimizing data loss during the extraction and transformation processes. The following section discusses in further detail the benefits of this transformation.

5 Conclusion and Future Work

The aim of the presented work is twofold. (1) To solve the issues found in the previous approach, trying to decouple the data extraction of audit trails from the application of analysis techniques. To do so, a mechanism is provided to integrate audit trails into the OpenSLEX meta model proposed in [11]. (2) To enable the application of new process mining techniques in an automated way, without the need for extensive domain knowledge, thanks to the standardization of the data representation proposed by OpenSLEX.

In order to solve the issues found in the previous approach, a new way to extract data from ATNA messages had to be developed. This new approach integrates ATNA messages into the OpenSLEX meta model to obtain a global, multi-perspective view of the data, reducing data loss as much as possible. This is achieved by a comprehensive data extraction process. Additionally, inference of incomplete data becomes possible by analyzing the participant objects indicated within the ATNA messages. All these data aspects of the process are rarely exploited in an integrated way by the existing event models such as XES. The proposed OpenSLEX meta model combines these data aspects with the process view to provide an integrated and complete picture of the system under study. On top of that, it provides the tools to discover the underlying data schema of the fragmented pieces of data captured in an heterogeneous environment such as current healthcare information systems.

Another one of the benefits of this transformation is that, collecting all the information in a standardized meta model enables the application of analysis techniques independently of the origin of data. Decoupling analysis from data extraction makes it possible to apply many analysis techniques with minimum effort. The standardization of queries, the application of automated machine learning, automated trace identification, and multi-perspective analysis are some of the techniques that become possible to apply when using a standardized and complete representation of data like the one that OpenSLEX provides.

The presented work is a first step in order to enhance the possibilities of process mining in healthcare. The fact that now we are able to process ATNA messages, allows us to exploit new possibilities, like the recent Standardized Operational Log of Events (SOLE). SOLE is a technical framework supplement of IHE that proposes the use of the ATNA message specification and IHE transaction (ITI-20) to communicate events regarding the whole radiological workflow

[3]. SOLE requires the use of the SWIM (SIIM Workflow Initiative in Medicine) workflow lexicon to define the semantics of the audit messages. This term list was created by the Society for Imaging Informatics (SIIM) to "improve the quality of information available to people managing imaging departments" [5] and was recently added to the RadLex[6] ontology. With audit messages describing events from start to end, the entries in an ARR can be used to analyze the complete workflow of an imaging facility. We plan to further develop our approach and do a case study based on real life event data from a radiological facility using the SOLE profile.

References

1. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. IEEE Std 1849-2016 pp. 1–50 (Nov 2016)
2. Association, N.E.M.: Nema ps3 / iso 12052, digital imaging and communications in medicine (dicom) standard. Website (2016), available free at http://medical.nema.org/
3. Committee, I.R.T.: Standardized operational log of events (sole). IHE Radiology Technical Framework Supplement Rev. 1.0 Draft for Public Comment, 51 (2017)
4. Cruz-Correia, R., Boldt, I., Lapão, L., Santos-Pereira, C., Rodrigues, P.P., Ferreira, A.M., Freitas, A.: Analysis of the quality of hospital information systems audit trails. BMC medical informatics and decision making 13(1), 84 (2013)
5. Erickson, B.J., Meenan, C., Langer, S.: Standards for business analytics and departmental workflow. Journal of Digital Imaging 26(1), 53–57 (2013)
6. Helm, E., Paster, F.: First steps towards process mining in distributed health information systems. International Journal of Electronics and Telecommunications 61(2), 137–142 (2015)
7. IHE: Audit trail and node authentication (atna). IHE IT Infrastructure (ITI) Technical Framework Volume 1, Integration Profiles ITI-TF-1, 71–83 (2016)
8. Ingvaldsen, J.E., Gulla, J.A.: Preprocessing support for large scale process mining of SAP transactions. In: Business Process Management Workshops. pp. 30–41. Springer (2008)
9. Mahendrawathi, E., Astuti, H.M., Wardhani, I.R.K.: Material movement analysis for warehouse business process improvement with process mining: A case study. In: Asia Pacific Business Process Management, pp. 115–127. Springer (2015)
10. Mueller-Wickop, N., Schultz, M.: ERP event log preprocessing: Timestamps vs. accounting logic. In: Design Science at the Intersection of Physical and Virtual Design, Lecture Notes in Computer Science, vol. 7939, pp. 105–119. Springer Berlin Heidelberg (2013)
11. González López de Murillas, E., Reijers, H.A., van der Aalst, W.M.P.: Connecting databases with process mining: A meta model and toolset. In: International Workshop on Business Process Modeling, Development and Support. pp. 231–249. Springer (2016)
12. Rebuge, Á., Ferreira, D.R.: Business process analysis in healthcare environments: A methodology based on process mining. Information Systems 37(2), 99–116 (2012)
13. Rojas, E., Munoz-Gama, J., Sepúlveda, M., Capurro, D.: Process mining in healthcare: A literature review. Journal of Biomedical Informatics 61, 224–236 (2016)

[6] RadLex ontology: http://www.radlex.org

14. Sismanis, Y., Brown, P., Haas, P.J., Reinwald, B.: Gordian: efficient and scalable discovery of composite keys. In: Proceedings of the 32nd international conference on Very large data bases. pp. 691–702. VLDB Endowment (2006)
15. Štolfa, J., Kopka, M., Štolfa, S., Kobĕrskỳ, O., Snášel, V.: An application of process mining to invoice verification process in sap. In: Innovations in Bio-inspired Computing and Applications, pp. 61–74. Springer (2014)
16. Zhang, M., Hadjieleftheriou, M., Ooi, B.C., Procopiuc, C.M., Srivastava, D.: On multi-column foreign key discovery. Proceedings of the VLDB Endowment 3(1-2), 805–814 (2010)