Extracting Event Data from Document-Driven Enterprise Systems

Diego Calvanese[1 [0000−0001−5174−9693]], Mieke Jans[2 [0000−0002−9171−2403]], Tahir Emre Kalayci[3 [0000−0001−6228−1221]], and Marco Montali[1 [0000−0002−8021−3430]]

[1] Free University of Bozen-Bolzano, 39100 Bolzano, Italy
{calvanese, montali}@inf.unibz.it
[2] Hasselt University, 3500 Hasselt, Belgium
mieke.jans@uhasselt.be
[3] Virtual Vehicle Research GmbH, 8010 Graz, Austria
emre.kalayci@v2c2.at

The preparation of input event data is one of the most critical phases in process mining projects. Different frameworks have been developed to offer methodologies and/or supporting toolkits for data preparation. One of these toolkits, called OnProm, frames the problem of data preparation in process mining as a data access and integration problem and is meant to work on arbitrary, legacy databases. However, in many settings, the input database is not a legacy one, but is structured with conceptually understandable object types and relationships that can be effectively employed to support business users in the extraction process. This is, for example, the case for document-driven enterprise systems. Therefore, we propose a guided approach, erprep, to support a group of business and technical users in setting up OnProm with minimal effort.

The OnProm approach [1,2] leverages the *ontology-based data access* (OBDA) paradigm [5] and is based on the use of an ontology that captures the semantics of the domain of interest. It requires to go through three phases: *(1)* creating the domain ontology, based on the domain expert's knowledge, *(2)* mapping this domain ontology to the underlying database structure, and *(3)* annotating the domain ontology to indicate where to find cases, events, and their attributes. There are two important open issues with this approach. The first issue is that, in general, phases (1) and (2) require genuine human effort, with data engineers interacting with domain experts, handling the full complexity of "ontology-based data access". There is no guidance available for this. At the same time, there is no guiding principle for the third phase (annotating the domain ontology to extract an event log), aside from the very low-level indication that "events require the presence of corresponding timestamps". The second issue is that there is a lack of validation of OnProm in real-life case studies.

The erprep guided approach provides a series of methodological steps for a group of experts collaborating in a process mining project. The approach is inspired by the procedure in [3] and systematically explores it as an intermediate layer between the group of users and OnProm, providing guidance in taking decisions on the log structure in a conscious manner [4]. The main purpose of erprep is to drive extracting event data from document-driven enterprise systems in agreement with the questions the group wants to find an answer for. At the

same time, the approach relieves them from manual, ad-hoc data preparation procedures. During these steps, the group of experts operates over different information sources that, behind the scenes, correspond to the information sources used by OnProm to effectively obtain the XES log in agreement with the perspective, granularity, and scope chosen by the group. In the description of the steps, we distinguish in particular the steps where the expert group directly interacts with OnProm from those where the group operates over intermediate information sources that are then (semi-)automatically translated into OnProm.

We assume that the group of experts consists of: *(i)* a *domain expert* who knows, at the business and operational level, the organisational domain, its relevant processes, its internal structure, and who wants to use process mining to answer specific, business-relevant questions; *(ii)* a *data engineer* who masters the information system(s) used by the organisation at the technical level, can access the schema of the underlying relational database(s), and knows how to formulate queries on top of it; *(iii)* a *process mining expert* who comes with in-depth knowledge of event data formats (in particular, XES). In a real setting, also depending on the size and complexity of the organisation, such competencies may often belong to different people who have to team up for a process mining project. The erprep approach, presented in this work, comprises five steps and are explained in a concrete fashion, using a running example.

Following these five steps, finally, OnProm is equipped with all the required information to be able to automatically generate an XES event log from the document-driven database, in agreement with the established mappings to the ontology and case/event annotations. Different analyses can then be easily conducted by simply varying the case and/or event annotations, invoking OnProm to obtain the corresponding, different XES logs.

We applied the erprep approach in a case study, where we obtained the following indicators, witnessing feasibility; *(i)* using the sales document perspective, 238 806 traces, 1 345 269 events, and 10 489 169 attributes have been extracted in 241 seconds, resulting in a log of 338 MB; *(ii)* using the invoice perspective, 247 051 traces, 1 742 127 events, and 10 497 414 attributes have been extracted in 362 seconds, resulting in a log of 457 MB.

References

1. Calvanese, D., Kalayci, T.E., Montali, M., Santoso, A.: OBDA for log extraction in process mining. In: RW Tutorial Lectures. LNCS, vol. 10370. Springer (2017)
2. Calvanese, D., Kalayci, T.E., Montali, M., Tinella, S.: Ontology-based data access for extracting event logs from legacy data: The onprom tool and methodology. In: Proc. BIS. LNBIP, vol. 288. Springer (2017)
3. Jans, M., Soffer, P.: From relational database to event log: Decisions with quality impact. In: Proc. of BPM Workshops. LNBIP, vol. 308. Springer (2017)
4. Jans, M., Soffer, P., Jouck, T.: Building a valuable event log for process mining: an experimental exploration of a guided process. Enterprise Information Systems **13**(5) (2019)
5. Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R., Zakharyaschev, M.: Ontology-based data access: A survey. In: Proc. IJCAI (2018)