Towards more structured data quality assessment in the process mining field: the DaQAPO package

Niels Martin[*a,b], *Greg Van Houdt*[b], *Gert Janssenswillen*[b]

[a] *Research Foundation Flanders (FWO), Egmontstraat 5, 1000 Brussels, Belgium*
[b] *Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium*

* *Corresponding author: niels.martin@uhasselt.be*

ABSTRACT

Process mining is a research field focusing on the extraction of insights on business processes from process execution data embedded in files called event logs. Event logs are a specific data structure originating from information systems supporting a business process such as an Enterprise Resource Planning System or a Hospital Information System. As a research field, process mining predominantly focused on the development of algorithms to retrieve process insights from an event log. However, consistent with the “garbage in - garbage out"-principle, the reliability of the algorithm’s outcomes strongly depends upon the data quality of the event log. It has been widely recognized that real-life event logs typically suffer from a multitude of data quality issues, stressing the need for thorough data quality assessment. Currently, event log quality is often judged on an ad-hoc basis, entailing the risk that important issues are overlooked. Hence, the need for a more structured data quality assessment approach within the process mining field. Therefore, the DaQAPO package has been developed, which is an acronym for Data Quality Assessment of Process-Oriented data. It offers an extensive set of functions to automatically identify common data quality problems in process execution data. In this way, it is the first R-package which supports systematic data quality assessment for event data.