

UHASSELT

KNOWLEDGE IN ACTION

ENHANCING DATA QUALITY
ASSESSMENT IN PROCESS MINING:
EXTENDING DaQAPO's
FUNCTIONALITIES

Hale YURTTUTAN
Supervisor: Prof. dr. Niels MARTIN



Introduction

Process mining focuses on identifying, tracking, and enhancing processes using information from event logs[1]. The principle of "garbage in, garbage out" highlights the importance of input data quality in process mining[3]. In this context, DaQAPO is developed to identify various data quality issues prior to process mining[2]. However, there exists a research gap as DaQAPO only covers a subset of data quality issues.



Research Questions and Methodology




How can the functionalities of DaQAPO be extended to provide improved support for data quality assessment of process data?

- 1)Which process data quality issues are currently not supported in DaQAPO?
- 2)Which assessment functions can be developed to address process data quality issues not yet covered by DaQAPO?
- 3)Can the novel assessment functions generate insights into the data quality of a real-life event log?

The design science research methodology is applied to extend the functionality of DaQAPO.



Selected Data Quality Issues for Functionality Extension

	Data quality issue	Detection
 Resource-related	Imprecise resource	Detects imprecise resource identifiers based on sample user input, identifying rows where the resource does not match the expected detailed format.
	Resource inconsistency	Detects violation of user-defined resource consistency rules within the same case, identifying cases where user-specified activities are performed by different resources.
	Resource-activity mismatch	Counts resource-activity combinations per user-defined resource, sorting and showing activities with significantly lower occurrence, indicating potential data logging issues.
 Timestamp-related	Imprecise timestamp	Detects timestamps that deviate from the sample user input format, displaying frequent activities and resources associated with imprecise timestamps.
	Same timestamp issues	Detects different activity sets that commonly have the same timestamp, indicating potential form-based logging.
 Activity-related	Synonymous labels	Detects a list of activity pairs that have never been logged together for a case, indicating potential synonymous labels.
	Incorrect case	Detects violation of user-defined rules specifying activities that should not co-occur, indicating potential data logging issues.



Results & Discussion

The application of the extended DaQAPO functionalities is demonstrated using a publicly available Dutch academic hospital event log[4]. It provides a realistic scenario for testing the generalizability and effectiveness of the developed functionalities. The results confirm that the outputs of these functionalities are relevant and are capable of identifying data quality issues.



Conclusion

The development of new functionalities in these areas broadens DaQAPO's ability to identify a wider range of data quality issues. From a broader perspective, this study is not merely an extension of DaQAPO's functionalities but also a contribution to the understanding of data quality issues and their assessment in process mining.

[1] Bose, R. J. C., Mans, R. S., & Van Der Aalst, W. M. (2013). Wanna improve process mining results? In *Proceedings of 2013 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 127–134).

[2] Martin, N., Van Houdt, G., & Janssenswillen, G. (2022). Daqapo: supporting flexible and fine-grained event log quality assessment. *Expert Systems with Applications*, 191, 116274.

[3] Suriadi, S., Andrews, R., ter Hofstede, A. H., & Wynn, M. T. (2017). Event log imperfection patterns for process mining: towards a systematic approach to cleaning event logs. *Information Systems*, 64, 132–150.

[4] van Dongen, B. (2011). *Real-life event logs - hospital log*. Eindhoven University of Technology. Retrieved from <https://data.4tu.nl/articles/12716513/1>