Plan Overview

A Data Management Plan created using DMPonline.be

Title: Robust Multi-Modal Prediction in Business Processes

Creator: Jari Peeperkorn

Principal Investigator: Jari Peeperkorn

Data Manager: Jari Peeperkorn

Project Administrator: Jari Peeperkorn

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Jari Peeperkorn

Data Manager: Jari Peeperkorn

Project abstract:

In recent years, machine learning models have been shown to be able to improve the predictive accuracy of predictive process monitoring models. However, industry adoption of these types of models has been rather limited. Recent studies have suggested these models, often relying on one process executions as input, can lack in generalization and robustness. The aim of the RoMuMoP project is to bridge the gap between theoretical development and practical application. This will be done firstly by introducing a holistic evaluation framework able to measure both technical metrics like accuracy and robustness, while also incorporating managerial aspects like explainability, reliability, and costs. Next, we pioneer the development of multi-modal prediction algorithms, able to take multiple data sources into account, such as event data, process models, and textual descriptions. Further, by incorporating zero-shot learning and transfer learning, we aim to tackle issues like data scarcity and lack of generalization. We demonstrate the practical utility of our testbed and multi-modal algorithms in three diverse domains: finance, logistics, and manufacturing, underscoring the potential transformative impact of the project on business processes.

ID: 214240

Start date: 01-10-2024

End date: 30-09-2027

Last modified: 03-04-2025

Robust Multi-Modal Prediction in Business Processes
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
Publicly available business process event logs, video, and textual descriptions	Publicly available data will be used to experimentally validate the developed methodologies	Reuse existing data	Digital	Observational	tabular data video data textual data	*TBD*	NA
Business process event logs, video, and textual descriptions shared by industry partners	Confidential (or sharable) data from industry partners will be used to experimentally validate the developed methodologies.	Collect new data	Digital	Observational	tabular data video data textual data	TBD	NA
Business process event logs and textual descriptions of generated artificial data	Simulated data with similar properties as industry partner's data will be used to experimentally validate the developed methodologies	Generate new data	Digital	Observational	tabular data textual data	TBD	NA
Code	Implementations of developed methodologies	Generate new data	Digital	Software	computational script	TBD	NA
Documentation code	Documentation of developed implementations	Generate new data	Digital	Textual	documentation	TBD	NA

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

An overview of frequently used publicly available process event data can be found here:
https://www.processmining.org/event-data.html
Video and event data example
https://zenodo.org/records/12671568

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

When personal data is used, anonymization or full replacement by simulated data will be used before sharing.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

The potential for commercial valorization relates to the methodlogies that will be developed, which are independent of datasets.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

The developed code will be accompanied by code documentation and the papers that describe the methodologies that are implemented. We plan to make our code publicly available on GitHub. README files will accompany the code in order to facilitate others to reuse it.
The publicly available datasets that will be used are well-documented.
For the private datasets agreements with companies will have to be made. The simulated data sets will be shared with documentation.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

TBD

3. Data storage & back-up during the research project

Where will the data be stored?

GitHub

How will the data be backed up?

Back-ups made by KUL FEB faculty ICT.
Code (intermediate and finished) is uploaded (and published) on GitHub.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

The code and data will probably require little storage space.
For the video files, a solution can be found if needed.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Standard safety measures are employed to safeguard work in progress. Code published on GitHub is available for others to use freely.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

TBD (will not be much normally)

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data will be preserved for 10 years according to KU Leuven RDM policy

Where will these data be archived (stored and curated for the long-term)?

GitHub (and other sites).
KUL FEB drives

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

TBD (probably none)

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Other, please specify:

Code will be available as open data.
SImulated/artificial data as well.
Some partner company data will most likely not be available to share publicly.

If access is restricted, please specify who will be able to access the data and under what conditions.

NA

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights

Where will the data be made available? If already known, please provide a repository per dataset or data type.

GitHub

When will the data be made available?

Upon publication of research results

Which data usage licenses are you going to provide? If none, please explain why.

CC-BY 4.0 (data)
MIT licence (code)

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

What are the expected costs for data sharing? How will these costs be covered?

None

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Jari Peeperkorn

Who will manage data storage and backup during the research project?

Jari Peeperkorn

Who will manage data preservation and sharing?

Jari Peeperkorn

Who will update and implement this DMP?

Jari Peeperkorn