Deep Process Model Forecasting

A Data Management Plan created using DMPonline.be

Creators: Johannes De Smedt, n.n. n.n.

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Johannes De Smedt, n.n. n.n.

Data Manager: Johannes De Smedt, n.n. n.n.

Grant number / URL: C14/23/031

ID: 202110

Start date: 01-10-2023

End date: 30-09-2027

Project abstract:

This project envisions to introduce a new take on process analytics by introducing a deep learning-based approach to forecast process models which simultaneously supports the discovery of process models from data, i.e., to understand how the process behaved in the past, and the prediction of process models, i.e., to understand how the process will behave in the future. Currently, both activities are treated separately, although they are learned from the same data.

To achieve this, the investigators will build on their recent efforts on a first incarnation of process model forecasting based on univariate statistical time series analysis and use them towards deep learning solutions. In a second instance, the short term-focused predictive monitoring and long-term model-wide forecasting will be combined into a single deep learning model based on recurrent and graph neural networks optimized with a process-specific loss function which will also support the discovery of process models over time directly. The project envisions significant predictive performance improvement and uncovering the interrelations between short- and long term, element-specific and model-wide predictions packaged in a user-friendly software tool.

Through a set of benchmarks, the model will be enhanced with guidelines in a Predictive Process Mining Framework (PPMF) which will equip practitioners with high-quality process-oriented decision support. The results will help process experts to simplify their predictive

Last modified: 06-10-2023

Deep Process Model Forecasting

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.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N**(ew data) or **E**(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
BPI Challenge 2011-2023	Event log data collected from various information systems used through the process mining research community for benchmarking and experimental evaluation.	☐ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☐ Experimental ☐ Compiled/ aggregated data ☐ Simulation data ☒ Software ☐ Other ☐ NA	☐ .por ☐ .xml ☐ .tab ☐ .csv ☐ .pdf ☐ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: .xes ☐ NA	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ 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:

All data is hosted at: https://data.4tu.nl/articles/
E.g., BPI 12 has a DOI of https://doi.org/10.4121/

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- 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.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or 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

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keepdata 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 data is already available and well-documenten on dedicated sites such as (https://www.win.tue.nl/bpi/2012/challenge.html).
For outcomes produced by the research in the project based on these available data, GitHub pages will be made available which contain the algorithms for pre-processing, data analysis, visualisation of results, and so on.

Will a metadata standard be used to make it easier tofind and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

Data Storage & Back-up during the Research Project

Where will the data be stored?

- OneDrive (KU Leuven)
- Personal network drive (I-drive)

For online processing, the data will be stored on the local desktop computers of the academics involved in the research

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- No (explain solution below)

No needed, all data is freely available.

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

NA

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

NA

Data Preservation after the end of the Research Project

Which data will be retained for 10 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)?

- Shared network drive (J-drive)

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

NA

Data Sharing and Reuse

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

- Other (specify below)

Already available open access

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 per dataset or data type where appropriate.

- No

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other data repository (specify below)

All data is hosted at: https://data.4tu.nl/articles/

When will the data be made available?

- Other (specify below)

Already available

Which data usage licenses are you going to provide?

If none, please explain why.

- Other (specify below)

https://data.4tu.nl/articles/_/12721292/1

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, my dataset already has a PID

Available for all the event logs listed above, e.g. for BPI 12: https://data.4tu.nl/articles/dataset/BPI_Challenge_2012/12689204

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

None

Responsibilities

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

The Main PI

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

The Main PI/postdoctoral research of the project

Who will manage data preservation and sharing?

The Main PI/postdoctoral research of the project

Who will update and implement this DMP?

The Main PI