DMP title

Project Name Automated Process Discovery of IoT Business Processes (APDIoT) - DMP title

Principal Investigator / Researcher Yannis Bertrand

Description The goal of the project is to investigate new process mining techniques that use IoT data. As such, we are using (and combining) two types of data: 1) logs of the execution of business processes and 2) data collected by sensors that track environmental parameters.

Institution KU Leuven

1. General Information

Name applicant

Estefanía Serral Asensio

Jochen De Weerdt

Francesco Leotta

FWO Project Number & Title

Automated Process Discovery of IoT Business Processes (APDIoT)

Affiliation

- KU Leuven
- Other

Francesco Leotta: Sapienza Università di Roma

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Type of data	Format	Volume	How created
Generated smart home data	XES	1-10 MB	Generated using a smart home data simulator
Event log data from partner company	CSV	100-150 MB	Record of all the activities performed by the operators of a manufacturing company
Planned process data from partner company	Copy of relational database (.bacpac)	25 MB	Database of the planning of the manufacturing company
Sensor data from partner company	CSV	10-50 GB	Sensor measurements of various interesting parameters during the execution of the process
Bees activity sensor data	CSV	1-10 MB	Sensor measurements of the movement and foraging behaviour of bumble bees

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

Privacy Registry Reference:

Short description of the kind of personal data that will be used:

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, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- No

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- Yes

The data provided by a partner company (data 2-3-4 in the data description) are subject to an NDA that restricts dissemination of the data.

Concretely, the data may only be processed by members of the project, and cannot be shared outside the team working on the project.

It is not excluded that a new agreement with more flexible terms might be negociated with the partner company in the future. If this were the case, the DMP would be updated accordingly.

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

1. Generated smart home data: consists of three XES datasets: one containing high-level activities, one with lower-level activities and one sensor log. A PDF document describing the datasets and the setup in more detail is available.

2. Event log data from partner company: contains reports of each activity performed, specifying the plant, tank, high-level activity, batch number, low-level activity, timestamp and various process parameters.

3. Planning data from partner company: whole planning database of the company, containing information on the batches, recipes, adjustments, workflow, etc. A database schema detailing the tables, the variables they contain and the relationships between them exists.

4. Sensor data from partner company: measurements of pressure, humidity, temperature, flow, etc. at various points in the process. Measurements are taken every second. An .xls sheet describing the sensors is available.

5. Bees activity sensor data: log of the visits of bees on a robotic flower. The dataset contains the time of the visit, the ID of the flower and the duration of the visit, together with environmental data collected by additional sensors.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

The data we will use in the scope of the project are not properly described by usual metadata standards. In addition to this, we cannot share too much detail regarding the data shared by the partner company.

However, metadata already exists for the generated smart home sensor data, and metadata will be created for the other data sources (within the limits set by the third-party agreement for data sources 2-3-4).

5. Data storage and backup during the FWO project

Where will the data be stored?

The data will be kept on the KU Leuven OneDrive. The company data is the only one that could be sensitive, but there is currently no specific obligation to store them in an environment providing more safety.

How is backup of the data provided?

OneDrive automatically makes backups of the data at regular intervals.

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

There is largely enough capacity for the data we plan to use now on the KU Leuven OneDrive (2TB available for < 100GB foreseen).

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

No additional cost is foreseen as long as OneDrive can be used.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

We rely on the KU Leuven OneDrive authentication layer. If a new agreement with our partner company were to set specific security requirements, we would contact ICT services to discuss a solution that also takes into account the potentially high volume of sensor data that we would receive.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

1.: Simulated smart home data will be kept available.

2., 3., 4.: We will likely not be authorised by the company to keep their data after the end of the project.

5.: Bees behaviour data will be kept available.

Where will the data be archived (= stored for the longer term)?

The data that can be kept will be stored in an X drive folder managed by Prof. Serral. This drive is a KU Leuven server/storage configuration optimised for "strictly confidential" research data.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

We expect no additional cost as the size of the data should remain below 2TB.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- Yes. Specify:

2., 3., 4.: The data have been obtained from a partner company and cannot be shared outside the team working on the project.

Which data will be made available after the end of the project?

1. and 5. will be made available.

2.,3.,4. will likely not be shared at the end of the project.

Where/how will the data be made available for reuse?

Where/how will the data be made available for reuse?

- In an Open Access repository
- Other (specify):

1. and 5. could be made accessible in an open access repository.
We do not foresee to make 2., 3., 4. publicly available as there is a third-party agreement with the company constraining the transfer and sharing of the data. If that were to change, the most likely would be to make a part of the data available under conditions agreed upon with our partner company. We would also update the DMP accordingly.

When will the data be made available?

- Immediately after the end of the project

1.is already available (at http://pros.webs.upv.es/sites/bp-meet-iot2022/#six).
5. can be made available shortly after the end of the project.
If we could make 2., 3., 4. available, we cannot say yet when that would be possible.

Who will be able to access the data and under what conditions?
The raw datasets 1. and 5. will be accessible freely online to anyone interested.
2., 3, 4. would have to be anonymised and preprocessed to prevent a third party from gaining knowledge of the inner working of our partner company. Furthermore, we would only grant access to the data for research purposes.

What are the expected costs for data sharing? How will the costs be covered?
We do not expect specific costs for data sharing at this point in time. If this were to change, we would update the DMP accordingly.

8. Responsibilities
Who will be responsible for data documentation & metadata?
Yannis Bertrand (PhD student)

Who will be responsible for data storage & back up during the project?
Yannis Bertrand (PhD student)

Who will be responsible for ensuring data preservation and reuse ?
Estefanía Serral Asensio (promotor of the project)

Who bears the end responsibility for updating & implementing this DMP?
The PI bears the end responsibility of updating & implementing this DMP.