DMP title

Project Name C3 - Het automatiseren van datatransformaties - DMP title

Project Identifier C3/21/061

Grant Title C3/21/061

Principal Investigator / Researcher Luc De Raedt

Project Data Contact Wannes Meert, wannes.meert@kuleuven.be

Description Data is everywhere, and being able to efficiently analyse and interpret this data allows companies to, amongst other thigns, save money and increase productivity. Existing tools allow users with varying skills levels to perform complex tasks, such as clustering, predictive modeling and forecasting, without writing code, by automatically selecting and configuring suitable algorithms for each task. These tools, however, assume that the data is in an appropriate shape; a single table in which all values are appropriately formatted. This assumption is often violated in practice, as a lot of data is stored in semi-structured formats, such as spreadsheets, and are formatted for human readability rather than for consumption by algorithms. Getting this data into a suitable shape is therefore time consuming, even for experienced data scientists, with up to 80% of time spent on data preparation tasks like restructuring tables and formatting values. Existing tools aimed at reducing the effort to transform various data, such as spreadsheets, into appropriate formats, however, make the strong assumption that users already know exactly what the final data should look like, both in terms of layout and formatting. The barrier for non-experts to get started with analysis thus remains high, and even experts are still spending most of their time on the preparation steps when presented with ill-formatted data. As part of the ERC Advanced Grant SYNTH project, the team of professor Luc De Raedt has already shown that predictive approaches are promising for suggesting effective layouts and formats for tables and values, respectively. The goal of this project is to leverage research prototypes with functionality required for practical use, an effort guided by a number of use cases, and to have all components in place for building a minimal viable product and pursue a spin-off based on this technology. With our solution, we thus aim to close the gap between having access to data, and quickly being able to generate value from this data, by allowing users to focus on its content, rather than on its structure and format.

Institution KU Leuven

1. General Information

Name of the project lead (PI)

Luc De Raedt

Internal Funds Project number & title

C3/21/061 - “Het automatiseren van datatransformaties.”

2. Data description

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

- Reuse existing data

2.2. What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

- A table layout transformation datasets, by using spreadsheets and tables from open data portals and templates.These will be Excel spreadsheets (tutorials, exercises, statistics), which are proprietary but publicly available.
- A (unsupervised) feature selection dataset, by using the collaborative data science platform Kaggle (https://kaggle.com) to determine how data scientists use data.
- Part of the project is to define use cases with industrial partners. The DMP will be augmented once it is clear which cases and datasets are considered (we do not intend to use any sensitive or personal data. If, in the unlikely case, we deem a such a dataset valuable we will first obtain aprovel through a PRET application).

3. Ethical and legal issues

3.1. Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

No.

We do not intend to use any sensitive or personal data. If, in the unlikely case, we encouter a such a dataset when searching for additional use cases and deem it valuable enough for the research, we will first obtain aprovel through a PRET application.

3.2. 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.

3.3. Does your research 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?

Yes. While none of the data will be protected by us (it is from external sources), the computer programs and models developed in this C3 project will be protected to achieve the valoriation goals of this C3. We are in communication with LRD to follow up on this.

3.4. 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 regarding reuse and sharing are in place?

No. Not on the IP generated in this project.

The data used to validate the results of the project will not be shared further as they are already publicly available.

4. Documentation and metadata

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

The code generated in this project will be documented following standard practices (inline code documentation and tutorial workbooks).

4.2. Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.

Standard methods such as inline code documentation and tutorial workbooks are provided

5. Data storage and backup during the project

5.1. Where will the data be stored?

The data and code will be stored on the servers of the Dept CS and KU Leuven. Code will be tracked in the Gitlab repository server. All data and other documents are kept on DTAI's internal secure NetApp storage server.

5.2. How will the data be backed up?

Gitlab and DTAI's NetAPP servers are automatically backed using KU Leuven's backup service. In addition NetApp has built-in disk failure recovory by means of redundancy.

5.3. 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.

DTAI's internal NetApp is a multi-terrabyte storage server, so well beyond the needs of this project.

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

DTAI's NetApp storage is a strategic long-term investment for which funds are foreseen in various projects and reserves.

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

DTAI's NetApp storage allows directory and file-based access to individual accounts. By default all directories are closed and only accessible to project members.

6. Data preservation after the end of the project

6.1. Which data will be retained for the expected 10 year period after the end of the project? If only a selection of the data can/will be preserved, clearly state why this is the case (legal or contractual restrictions, physical preservation issues, ...).

All code will be retained for 10 year. Copies of the data used to validate the results and results are also kept for 10 years in snapshots on DTAI's NetApp server.

6.2. Where will these data be archived (= stored for the long term)?

On DTAI's NetApp server with a backup to KU Leuven's backup servers.

6.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?

The NetApp is a strategic long-term investment by the DTAI section. The costs are covered by ongoing projects, and reserves are available if necessary. The system is maintained by the Dept CS sysadmin team.

7. Data sharing and re-use

7.1. 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 or because of IP potential)?

All data is reused data from other sources and will not be distributed by us. Copies will be retained internally for reprodicibility purposes.

Software developed in this project is proprietary with the aim for valorisation and will not be shared freely.

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

Software will be offered under a (commercial) software license. Ideally this will be through a spin-off company that obtains the rights to this software from KU Leuven.

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

- Other (specify):

Only commercially.

7.4. When will the data be made available?

When a minimal viable product is realized.

7.5. Who will be able to access the data and under what conditions?

Only project members.

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

None.

8. Responsibilities

8.1. Who will be responsible for the data documentation & metadata?

Luc De Raedt and Wannes Meert

8.2. Who will be responsible for data storage & back up during the project?

Wannes Meert

8.3. Who will be responsible for ensuring data preservation and sharing?

Wannes Meert

8.4. Who bears the end responsibility for updating & implementing this DMP?

The end responsibility for updating and implementing the DMP is with the supervisor (promotor).