# Effective utilization of costly data for informed decision making
# Application DMP

Questionnaire

**Describe the datatypes (surveys, sequences, manuscripts, objects … ) the research will collect and/or generate and /or (re)use. (use up to 700 characters)**

During this research, I will use input data, gather computational results, write code, and write papers. The input data can be either (1) artificially generated data, including existing problem sets taken from literature, or (2) real-life data. All artificial data will be made available online and its generation described in detail. Real-life data will most likely not be used, but if so then it will be provided in an anonymized way, not contain personal information, and will be safely stored and never be published. My code and experimental results will be accessible online and my papers will be shared on an open-access repository, such as arXiv and/or SSRN, to ensure transparency.

**Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)**

I will take personal responsibility for making both my code and generated datasets accessible on my GitHub repository to facilitate easy replication of the results. Additionally, to enhance transparency further, completed manuscripts will be shared on an open-access repository (arXiv / SSRN). I take responsibility for this process both during and after the research project. Furthermore, research data will be stored on the KU Leuven OneDrive for Business, during as well as after the project, where the data is safe and automatically backed up. While I will be responsible for managing the data during the project, Roel Leus, the PhD supervisor, will take responsibility afterward.

**What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)**

NA

**Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)**

NA

**Which other issues related to the data management are relevant to mention? (use up to 700 characters)**

NA

# Effective utilization of costly data for informed decision making
# 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 |
| | | *Please choose from the following options:*<br>• Generate new data<br>• Reuse existing data | *Please choose from the following options:*<br>• Digital<br>• Physical | *Please choose from the following options:*<br>• Observational<br>• Experimental<br>• Compiled/aggregated data<br>• Simulation data<br>• Software<br>• Other<br>• NA | *Please choose from the following options:*<br>• .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, …<br>• NA | *Please choose from the following options:*<br>• <100MB<br>• <1GB<br>• <100GB<br>• <1TB<br>• <5TB<br>• <10TB<br>• <50TB<br>• >50TB<br>• NA | |
| Artificial data from literature | Many papers utilize data that is artificially generated and for which the generation is documented. This data is often freely available online and will be used in our experiments. | Reuse existing data | Digital | Other | NA (depending on dataset, currently using dataset with extension .dat) | < 1GB | |
| Newly generated artificial data | When certain types of datasets are not yet available, we will generate our own datasets. We will document how this data is generated and make the data freely available online. | Generate new data | Digital | Other | .csv | < 1 GB | |
| Code | Code written to run experiments. | Generate new data | Digital | Software | .cpp/.h/... | <100 MB | |
| | | | | | | | |

**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:**

For our current project we are using the data used in the paper Time-critical testing and search problems - ScienceDirect . This

data has been personally provided by Ben Hermans and is stored on OneDrive.

**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.**

- No

**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.**

- No

**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).**

Generated data and written code will be made available online and will be accompanied by a README.txt file in which the data is explained.
In the paper, the generation of the datasets and the steps of the algorithms will be explained.
Both the code to generate the data and the algorithms that are coded will be made available through Git Hub.

**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.**

- No

Generated data will be made available online together with a README.txt file explaining the data files.

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

**Where will the data be stored?**

Data currently used for our research project is stored on OneDrive. OneDrive for Business is a Microsoft cloud solution to securely store, manage and share documents and files. KU Leuven has signed a license agreement with Microsoft for OneDrive for Business for staff and students as part of the Microsoft 365 Education plan.

**How will the data be backed up?**

Full backup in non-Microsoft data center.

**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

250 GB is available, while less than 1 GB is needed.

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

On OneDrive for business files are private unless explicitly shared. Currently data is only shared with Ben Hermans, who provided the data.

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

OneDrive for business is free for staff of KU Leuven, meaning that currently no costs are involved.

#### 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...).**

Newly generated data will be retained as long as possible after the end of the project, such that other researchers can run experiments using the same data and take a look at the code that was developed. We will preserve data for at least 10 years, but we aim for 25 years or more.

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

We plan on archiving generated datasets on KU Leuven RDR, which is KU Leuven's institutional research data repository for the publication of research data. It is a Dataverse.org based platform to upload, describe, and share research data. A published dataset in KU Leuven RDR gets its own DOI and is registered in Lirias.

Code that was written will be archived on GitHub, allowing not only share the code but also manage versions and collaborate with others.

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

For the generated datasets there won't be any costs as every researcher can store 50 GB per year for free.
For sharing code the GitHub free plan currently suffices.

**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

Newly generated artificial data will be made openly available such that others can run experiments on the same data sets.
Software will be made publicly available for others to use.

**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.**

- No

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

We plan on making the datasets available on KU Leuven RDR, which is KU Leuven's institutional research data repository for the publication of research data.
We plan on making code available through GitHub.

**When will the data be made available?**

Around the time of publication of results.

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

For the data that we generate, we will use CC-BY-4.0.
For the code that we have written we will use the MIT license.

**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?**

We don't expect to have any costs.

**6. Responsibilities**

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

Emmeline Perneel

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

Emmeline Perneel

**Who will manage data preservation and sharing?**

Emmeline Perneel

**Who will update and implement this DMP?**

Emmeline Perneel