FWO DMP Template - Flemish Standard Data Management Plan

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. General Project Information	
Name Grant Holder & ORCID	**Pieter Vanmechelen** **0000-0002-3733-2880**
Contributor name(s) (+ ORCID) & roles	**Giovanni Samaey** **Promotor** **0000-0001-8433-4523** **Geert Lombaert** **Co-promotor** **0000-0002-9273-3038**
Project number[1] & title	3E200531 Multilevel Markov Chain Monte Carlo methods for Bayesian full-field data assimilation
Funder(s) GrantID[2]	**1SD1823N**
Affiliation(s)	■ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: Provide ROR[3] identifier when possible:
Please provide a short project description	Markov chain Monte Carlo (MCMC) methods are an indispensable tool in sampling probability distributions for which a direct sampling is impossible. In Bayesian data assimilation, these methods are used to sample the posterior probability distribution for a parameter that needs to be estimated based on an accept-reject procedure, given a prior distribution for the parameter and a model that allows computing the likelihood of the data for a given value of the parameter. This project deals with the development and analysis of MCMC methods when the available data has a full-field nature. Markov chain Monte Carlo methods face two problems when confronted with full field data. First, computing the likelihood can be a very expensive step in MCMC sampling, especially when the corresponding model contains multiple time scales and is high-dimensional. Second, the acceptance rates of standard MCMC procedures drop as the data dimensionality increases. In this project, we will develop and analyse multilevel Markov chain Monte Carlo methods that increase the computational efficiency of Bayesian posterior sampling by making use of a hierarchy of levels at different resolution, and by explicitly exploiting spatial correlation that is present in full-field data. While being application-independent in nature, the developed methods will be applied to problems in damage assessment in civil engineering.

[1] "Project number" refers to the institutional project number. This question is optional since not every institution has an internal project number different from the GrantID. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

[3] Research Organization Registry Community. https://ror.org/

2. 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[4].

				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
		☐ 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: ☐ NA	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
Simulation code	MATLAB scripts	New data	Digital	Software	.m	< 1 GB	/
Numerical experiments	Results produced by running MATLAB scripts	New data	Digital	Experimental data	.mat	< 100 GB	/

[4] Add rows for each dataset you want to describe.

Manuscripts	Paper and thesis text files	New data	Digital	Text	.tex and .pdf	< 1 GB	/
Papers	Literature study	Existing data	Digital	Text	.pdf	< 100 GB	/

GUIDANCE:

DATA CAN BE DIGITAL OR PHYSICAL (FOR EXAMPLE BIOBANK, BIOLOGICAL SAMPLES, …). DATA TYPE: DATA ARE OFTEN GROUPED BY TYPE (OBSERVATIONAL, EXPERIMENTAL ETC.), FORMAT AND/OR COLLECTION/GENERATION METHOD.

EXAMPLES OF DATA TYPES: OBSERVATIONAL (E.G. SURVEY RESULTS, SENSOR READINGS, SENSORY OBSERVATIONS); EXPERIMENTAL (E.G. MICROSCOPY, SPECTROSCOPY, CHROMATOGRAMS, GENE SEQUENCES); COMPILED/AGGREGATED DATA[5] (E.G. TEXT & DATA MINING, DERIVED VARIABLES, 3D MODELLING); SIMULATION DATA (E.G. CLIMATE MODELS); SOFTWARE, ETC.

EXAMPLES OF DATA FORMATS: TABULAR DATA (.POR,. SPSS, STRUCTURED TEXT OR MARK-UP FILE XML, .TAB, .CSV), TEXTUAL DATA (.RTF, .XML, .TXT), GEOSPATIAL DATA (.DWG,. GML, ..), IMAGE DATA, AUDIO DATA, VIDEO DATA, DOCUMENTATION & COMPUTATIONAL SCRIPT.

DIGITAL DATA VOLUME: PLEASE ESTIMATE THE UPPER LIMIT OF THE VOLUME OF THE DATA PER DATASET OR DATA TYPE.

PHYSICAL VOLUME: PLEASE ESTIMATE THE PHYSICAL VOLUME OF THE RESEARCH MATERIALS (FOR EXAMPLE THE NUMBER OF RELEVANT BIOLOGICAL SAMPLES THAT NEED TO BE STORED AND PRESERVED DURING THE PROJECT AND/OR AFTER).

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.	The only existing data reused is the collection of papers used in the literature study. Sources are various journals in which these papers were published.
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, please describe these issues further and refer to specific datasets or data types when appropriate.	☐ Yes, human subject data ☐ Yes, animal data ☐ Yes, dual use ■ No If yes, please describe:

[5] These data are generated by combining multiple existing datasets.

Will you process personal data[6]? If so, briefly describe the kind of personal data you will use. Please refer to specific datasets or data types when appropriate. If available, add the reference to your file in your host institution's privacy register.	☐ Yes ■ No If yes: - Short description of the kind of personal data that will be used: - Privacy Registry Reference:
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 ☐ No If yes, please comment: The source code used to generate results will be made publicly available, hence can be used by others as example for commercial implementation of our methods. However, the code is not directly usable commercially as it is designed with academic purposes in mind.
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 to what data they relate and what restrictions are in place.	☐ Yes ■No If yes, please explain:
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 to what data they relate and which restrictions will be asserted.	☐ Yes ■No If yes, please explain:

[6] See Glossary Flemish Standard Data Management Plan

3. 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).	**Code is documented and has an accompanying README. The code is in version control, where the git history allows to trace back changes. Computational experiments are explained carefully in papers and are tested for reproducibility. Additionally, published code is accompanied with a dockerfile such that a virtual machine can be set up with the right environment to run the published codes. Using these codes and dockerfile based virtual machines, each of the published computational experiments can be reproduced.**
Will a metadata standard be used to make it easier to **find 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. *Repositories could ask to deliver metadata in a certain format, with specified ontologies and vocabularies, i.e. standard lists with unique identifiers.*	■Yes ■No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: For computational experiments we encode metadata about the running environment for the code in dockerfiles. If no, please specify (where appropriate per dataset or data type) which metadata will be created: Metadata about how to reproduce results will be documented in papers and README files. **Note:** in the future, we plan to automate metadata generation in our research group using iRODS. The pilot for this change in data management flow still has to be set up.

4. Data Storage & Back-up during the Research Project

Where will the data be stored?	**Data is stored locally on a departmentally issued laptop and depending on data type, back-ups are stored:** • **For version-controlled files, the KU Leuven Gitlab service is used** • **Literature is maintained in a reference manager (Zotero)** • **Other files are backed up on the KU Leuven OneDrive or the NextCloud server of the Department of Computer Science.** **After the research:** • **Where possible, all relevant data/code will be made publicly available with the corresponding papers** • **Manuscripts and all corresponding data are stored with read-only access at the Department of Computer Science.**
How will the data be backed up? *What storage and backup procedures will be in place to prevent data loss? Describe the locations, storage media and procedures that will be used for storing and backing up digital and non-digital data during research.*[7] *Refer to institution-specific policies regarding backup procedures when appropriate.*	**All files are synced with departmental file servers to ensure daily backup. The sole exception is the papers of the literature study which are stored in Zotero. This all is in accordance with the wider NUMA DMP policy, used by almost all researchers in our section.**

[7] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

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 ☐ No If yes, please specify concisely: All data is digital and currently backed up with plenty of storage space left. If in future this space is insufficient, more can be requested at the departmental level. If no, please specify:
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *Clearly describe the measures (in terms of physical security, network security, and security of computer systems and files) that will be taken to ensure that stored and transferred data are safe.* [7]	This issue is covered by the NUMA data management plan.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	Costs are fully covered by the NUMA research section.

5. 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...).	**Where possible, all relevant data/code will be made publicly available with corresponding papers.** **Manuscripts and all corresponding data are stored with read-only access at the Department of Computer Science.**
Where will these data be archived (stored and curated for the long-term)?	**Department of Computer Science**
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	**Costs will be fully covered by the NUMA research section.**

6. 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. *Note that 'available' does not necessarily mean that the data set becomes openly available, conditions for access and use may apply. Availability in this question thus entails both open & restricted access. For more information: https://wiki.surfnet.nl/display/standards/info-eu-repo/#infoeurepo-AccessRights*	■ Yes, in an Open Access repository ■Yes, in a restricted access repository (after approval, institutional access only, …) ☐ No (closed access) ☐ Other, please specify:
If access is restricted, please specify who will be able to access the data and under what conditions.	The promotor of this project, Giovanni Samaey, has full access to all repositories and can regulate access to the data.
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.	☐ Yes, privacy aspects ☐ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ■ No If yes, please specify:
Where will the data be made available? If already known, please provide a repository per dataset or data type.	**KU Leuven gitlab servers**

When will the data be made available? *This could be a specific date (dd/mm/yyyy) or an indication such as 'upon publication of research results'.*	**Open access repositories will be created upon publication of the corresponding paper.** **Restricted access repositories can be made available upon request.**
Which data usage licenses are you going to provide? If none, please explain why. *A data usage license indicates whether the data can be reused or not and under what conditions. If no licence is granted, the data are in a grey zone and cannot be legally reused. Do note that you may only release data under a licence chosen by yourself if it does not already fall under another licence that might prohibit that.* *Example Answer: E.g. "Data from the project that can be shared will be made available under a Creative Commons Attribution license (CC-BY 4.0), so that users have to give credit to the original data creators."* [8]	**This is still under consideration.**
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *Indicate whether you intend to add a persistent and unique identifier in order to identify and retrieve the data.*	■Yes ☐ No If yes: We will use a KU Leuven system that is currently in development.

[8] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

What are the expected costs for data sharing? How will these costs be covered?	**Costs are not precisely known, but expected to be modest.**

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	**Pieter Vanmechelen**
Who will manage data storage and backup during the research project?	**Pieter Vanmechelen**
Who will manage data preservation and sharing?	**Giovanni Samaey**
Who will update and implement this DMP?	**Pieter Vanmechelen**