FWO DMP Template - Flemish Standard Data Management Plan

Version KU Leuven

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	**Felipe Kenji Nakano 0000-0002-4884-9420**
Contributor name(s) (+ ORCID) & roles	**Celine Vens 0000-0003-0983-256X: PI** **Fabian Güiza Grandas** 0000-0001-7026-0957**: Collaborator**
Project number [1] & title	235924N - Novel tree-ensemble based methods for weakly-supervised structured output prediction with applications in biomedicine
Funder(s) GrantID [2]	
Affiliation(s)	X KU Leuven □ Universiteit Antwerpen □ Universiteit Gent □ Universiteit Hasselt □ Vrije Universiteit Brussel □ Other: ROR identifier KU Leuven: 05f950310

[1] "Project number" refers to the institutional project number. This question is optional. 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.

Please provide a short project description	Recent studies in structured output prediction, an umbrella term for machine learning tasks where multiple outputs must be predicted, have identified the veracity of the data as a major challenge. More specifically, structured output prediction datasets present noise in the output space due to faulty equipment, high cost of annotation or high volume of data, meaning that they are weakly-supervised. State-of-the-art methods, however, often disregard such weak-supervision, hindering their performance. In this project, I will investigate weakly-supervised structured output prediction. Concretely, I will develop novel methods based on tree-ensembles and deep-forest architectures which can handle such noise in two predictive tasks of interest: hierarchical multi-label classification (outputs are correlated according to a hierarchy) and multi-task learning (multiple outputs of different types). My methods will be validated on benchmark datasets and also in two biomedical applications: protein function prediction (hierarchical multi-label classification) and intensive care unit acquired-weakness (ICU-AW) prediction in multi-task learning. Here, I propose an innovative project, with local and international collaborators, which addresses emerging problems in machine learning, and consequently advances the state-of-the-art. The developed methods will be made publicly available to facilitate reproducible research and interdisciplinary collaboration.

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 [3].

				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
Hierarchical multi-label classification benchmark datasets (WP1)	Benchmark datasets related to protein function prediction.	☐ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	.csv	☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
Multi-label classification benchmark datasets (WP2)	Benchmark datasets related to multiple domains of knowledge	Reuse existing data	Digital	Numerical	.csv	< 1GB	
Multi-target regression datasets (WP2)	Benchmark datasets related to multiple domains of knowledge	Reuse existing data	Digital	Numerical	.csv	< 1GB	

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

Intensive care unit acquired-weakness (WP2, Task 2.4)		Reuse existing data	Digital	Numerical	.csv	< 1GB	

GUIDANCE:
The data description forms the basis of your entire DMP, so make sure it is detailed and complete. It includes digital and physical data and encompasses the whole spectrum ranging from raw data to processed and analysed data including analysis scripts and code. Physical data are all materials that need proper management because they are valuable, difficult to replace and/or ethical issues are associated. Materials that are not considered data in an RDM context include your own manuscripts, theses and presentations; documentation is an integral part of your datasets and should described under documentation/metadata.
RDM Guidance on data

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.	Hierarchical multi-label datasets: Originally proposed in 2007 by my PI, Celine Vens, in the publication: https://doi.org/10.1007/s10994-008-5077-3 and updated by us in 2018, in the publication: https://doi.org/10.1186/s12859-019-3060-6 Multi-label classification and multi-target regression benchmark datasets: Datasets related to multiple domains of knowledge. We have collected and pre-processed them in a previous work of ours: https://doi.org/10.1016/j.patcog.2021.108211 Intensive care unit acquired-weakness: This dataset was collected during the EPaNIC-trial (NCT00512122) and it contains information related to patients who underwent ICU-AW assessment on day 8 from ICU-admission onward: doi: 10.1056/NEJMoa1102662
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.	☒ Yes, human subject data; provide SMEC or EC approval number: ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☐ No Additional information: Intensive care unit acquired weakness: Leuven University Hospital Ethics Committee (ML4190)

Will you process personal data[4]? 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).	☒ Yes (provide PRET G-number or EC S-number below) ☐ No Additional information: The trial associated to the intensive care unit acquired weakness dataset, EPaNIC, was the last one before the implementation of the s-numbers. Thus, it presents only a ML number (ML4190). The only S-number that may be associated to it (s50404) belongs to a sub-study.
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 developed software may result in valorisation potential, e.g. for companies developing intensive care unit acquired-weakness treatments. In that case, the exact software license that will be used should be discussed with LRD and will be updated later
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: Exploitation restrictions depend on the dataset: 1. Benchmark datasets: No restrictions 2. Intensive care unit acquired-weakness dataset: We can only work on the work packages and tasks as described in the data use agreement. This agreement further stipulates the authorship rules for publications resulting from working on this data. Dissemination of the patient data from the Intensive care unit acquired-weakness is not possible. The patient data remains the property of the medical centers in each case.
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:

[4] 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). *RDM guidance on documentation and metadata*.	Documentation will include information on how to pre-process the datasets (which variables to select, how to deal with missing data, etc.), as well as the exact validation procedure (train/test split, cross validation, evaluation measures) and parameter optimization procedure. Although a detailed methodology section will be required in any publication about the project, we will have a more elaborate description of the exact protocol in an extra document, this may be a readme.txt file or a lab notebook. For the generated computer programs, detailed source code documentation and a manual will be added.
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: If no, please specify (where appropriate per dataset or data type) which metadata will be created: The metadata of the generated software includes programming language, author, version, date of creation,... Such metadata is automatically created by a git repository like KU Leuven GitLab.

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

Where will the data be stored? *Consult the interactive KU Leuven storage guide to find the most suitable storage solution for your data.*	☒ Shared network drive (J-drive) Benchmark datasets ☒ Personal network drive (I-drive) Benchmark datasets ☒ OneDrive (KU Leuven) Benchmark datasets ☐ Sharepoint online ☐ Sharepoint on-premis ☐ Large Volume Storage ☐ Digital Vault ☐ Other: Intensive care unit acquired-weakness: my own hard drive
How will the data be backed up? *What storage and backup procedures will be in place to prevent data loss?*	☒ Standard back-up provided by KU Leuven ICTS for my storage solution ☒ Personal back-ups I make (specify) ☐ Other (specify)
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 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.* *Guidance on security for research data*	• Hard drive of my laptop: has a bitlocker, laptop itself is password-protected. • OneDrive for Business: secured by two-factor authentication. • KU Leuven GitLab: secured by two-factor authentication. • The latter 2 storage media are supported by the KU Leuven infrastructure, and are ensured to stay within a datacenter in Europe

What are the expected costs for data storage and backup during the research project? How will these costs be covered?	No substantial costs for storage are expected.

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...). *Guidance on data preservation*	☒ All data will be preserved for 10 years according to KU Leuven RDM policy ☐ All data will be preserved for 25 years according to CTC recommendations for clinical trials with medicinal products for human use and for clinical experiments on humans ☐ Certain data cannot be kept for 10 years (explain) The reused patient data can not legally be preserved by us after the research project. The responsible for the preservation of the intensive care unit acquired weakness data is Dr. Ir. Fabian Güiza Gandas.
Where will these data be archived (stored and curated for the long-term)? *Dedicated data repositories are often the best place to preserve your data. Data not suitable for preservation in a repository can be stored using a KU Leuven storage solution, consult the interactive KU Leuven storage guide.*	☒ KU Leuven RDR ☐ Large Volume Storage (longterm for large volumes) ☒ Shared network drive (J-drive) ☒ Other (specifiy): Github repositories
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	The expected costs for preservation are negligible and will be covered by the budget of the PI.

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, as open data ☐ Yes, as embargoed data (temporary restriction) ☐ Yes, as restricted data (upon approval, or institutional access only) ☐ No (closed access) ☐ Other, please specify: • Raw data reused in this project and processed data will not be made available for legal and ethical reasons. The processed data can easily be recovered from the raw data using the software code generated during this project. • Analysed data will be included in a publication, possibly as supplemental materials. • Code may be made available on a project-per-project basis, e.g. in a hosted Git repository. Access to these repositories will not be restricted, as they do not consist of sensitive personal data (although care must be taken to curate the analysed data, e.g. graphs and tables, to fulfil this condition).
If access is restricted, please specify who will be able to access the data and under what conditions.	Not applicable
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: **Data related to intensive care unit acquired weakness may not be shared to 3rd parties without the permission of my collaborator Fabian Güiza Grandas**

Where will the data be made available? If already known, please provide a repository per dataset or data type.	☐ KU Leuven RDR ☒ Other data repository (specify) ☐ Other (specify) In an Open Access repository, such as GitHub.
When will the data be made available?	☒ Upon publication of research results ☐ Specific date (specify) ☐ Other (specify)
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.* *Check the RDR guidance on licences for data and software sources code or consult the License selector tool to help you choose.*	☐ CC-BY 4.0 (data) ☐ Data Transfer Agreement (restricted data) ☐ MIT licence (code) ☐ GNU GPL-3.0 (code) ☒ Other (specify) The exact license of the software has to be decided and will be discussed with LRD. For internal use in the research group, the software will be stored on the network drives.
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, a PID will be added upon deposit in a data repository ☐ My dataset already has a PID ☐ No A DOI can be generated for any GitHub repositories
What are the expected costs for data sharing? How will these costs be covered?	No costs for data sharing are expected, creation of a public GitHub repository is free.

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Myself
Who will manage data storage and backup during the research project?	Myself
Who will manage data preservation and sharing?	PI
Who will update and implement this DMP?	PI and myself