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	**Steven Janssens, 0000-0003-1246-9474**
Contributor name(s) (+ ORCID) & roles	**Yves Bawin, lab technician, 0000-0002-1663-6535**
Project number [1] & title	G0GDY23N, Harmonizing plant metabarcoding systems in Europe to support monitoring activities in plants and associated functional networks of organisms
Funder(s) GrantID [2]	FWO (G0GDY23N)
Affiliation(s)	KU Leuven 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	Plant metabarcoding, which involves analyzing environmental DNA (eDNA) to identify taxa, can be standardized and automated, making it suitable for high-throughput, large-scale, and long-term monitoring. This technique provides a scale and accuracy in biodiversity surveys that was previously unattainable. In Europe, initiatives for nature conservation rely on species presence, and indicator species form the foundation of decision-making in conservation. However, public and private sector stakeholders need rapid, accurate, and inexpensive methods to monitor plant biodiversity. The transnational and transdisciplinary METAPLANTCODE project aims to test and optimize pan-European case studies on metabarcoding, provide best practice recommendations, optimize analysis pipelines for species identification, and create easy-to-use reference databases to be implemented in European and national infrastructures, in collaboration with BIOSCAN Europe, ELIXIR communities and others. The project will identify and specify gaps, publish best practice documents on FAIR data publishing of plant metabarcode data to GBIF and the INSDC databases, and implement ELIXIR-compatible multimodal DL models in novel tools for stand-alone metabarcoding analyses using different data sources. The project will also enhance species identification accuracy through GBIF records and metadata (GBIF, EUNIS, Biolflor, by sequence data, text data, taxonomic classification datasets, and ecological analyses) and map regional, national, and international botanical taxonomic checklists, red lists, and floras to the Catalogue of Life (COL) through COL ChecklistBank. Furthermore, taxonomic and florsitic literature will be semantically enriched with new entity recognition and relationship extraction modules to support the enhanced identification of species via domain-specific descriptive/phenotypic features (e.g., habitats, features, soil characteristics, biotic interactions). An interface will be provided to link taxonomic names to treatments, identify homonyms and synonyms, and facilitate the conversion and annotation of flora, red lists, and ecological treatments. All METAPLANTCODE products will be available at project end FAIR+. The project will support knowledge transfer with associated partners and stakeholders from the start. Relevant stakeholders will be identified, priorities set, communication channels established, monitored, and revised as needed. Greater stakeholder engagement, training, and outreach efforts will be undertaken to ensure that plant metabarcoding becomes a routine standard for biodiversity monitoring in Europe and beyond in the future.

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
Sequencing data	Raw sequencing data from environmental samples.	☒ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Numerical ☐ Textual ☐ Software ☒ Other: sequences	.txt, .fastq, .fasta, .bam, .pdf	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ > 5 TB ☐ NA	
Samples	Samples and their DNA extracts taken from the environment (soil, insects, and air)	☒ Generate new data	☒ Physical				Frozen samples: tubes stored at -20 °C
Scripts	Code written for the analysis of metabarcoding data	☒ Generate new data	☒ Digital	☒ Software	.txt, .py, .R	☒ < 1 TB	
Biodiversity data	Species diversity lists obtained from environmental samples	☒ Generate new data	☒ Digital	☒ Other: species lists	.txt, .pdf	☒ < 1 GB	

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

Literature	Collection of scientific publications about metabarcoding studies	☒ Reuse existing data	☒ Digital	☒ Textual	.pdf	☒ < 100 GB	

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.	Scientific publications were retrieved from multiple scientific databases: - Google Scholar: https://scholar.google.com/ - Scopus: https://www.elsevier.com/products/scopus/search - Limo: https://kuleuven.limo.libis.be/discovery/search?vid=32KUL_KUL:KULeuven - Web of Science: https://www.webofscience.com/wos/woscc/basic-search
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: Publicly shared anonymized datasets
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:

[4] See Glossary Flemish Standard Data Management Plan

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:
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:

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*.	We will keep track of the following in distinct documents and electronic notebooks: - Sampling strategy (.one) - List of abbreviations (.docx) - Lab protocols and optimization steps (.docx, .one, .pdf) - Relevant analysis scripts (.py, .R) - Raw data (specific file format according to data type) - Processed data (specific file format according to data type) - List of scientific publications will be made available on Zenodo. - Physical data: (DNA) samples will be stored in a stabilizing agent at -20°C for at least ten years after the end of the project.
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: We will standardize the data generated by this project using specific standards for different data types: - Biodiversity data will be published following the Darwin core standards (https://www.dcc.ac.uk/resources/metadata-standards/darwin-core) - Genomic data will be made available using the genome metadata standards (https://www.dcc.ac.uk/resources/metadata-standards/genome-metadata) - All other data will be reported using the Dublin core standards (https://www.dcc.ac.uk/resources/metadata-standards/dublin-core) If no, please specify (where appropriate per dataset or data type) which metadata will be created: No additional metadata will be created.

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) ☒ Personal network drive (I-drive) ☒ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☒ Large Volume Storage ☐ Digital Vault ☐ Other:
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 on my personal OneDrive. ☒ Other (specify) An agreement has been made that all data will be stored and backed-up at the servers of Meise Botanic Garden by the ICT service of Meise Botanic Garden.
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, all the data is already available and safe in the servers of Meise Botanic Garden. ☐ 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*	Researchers that are not involved in the project will not have access to data from ongoing studies. Data from finalized studies will be made publicly available under appropriate copyright licenses. Access to the data is controlled by the ICT service of Meise Botanic Garden and is password protected.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	No additional costs for data storage are required.

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)

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 (specify): the data will be archived at the servers of Meise Botanic Garden, which is maintained by the ICT service of Meise Botanic Garden. Data associated to scientific publications will be stored in the European Nucleotide Archive (ENA) and the Global Biodiversity Information Facility (GBIF). Scripts will be made publicly available on GitLab and GitHub.
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	No additional costs for data storage are required.

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: The data will be made publicly available after publication via the required link in the publication. Moreover, biodiversity data will be freely available on GBIF. Upon publication, the data will be free to use for scientific purposes, referring to the original publication, but not for commercial use. Data will be released under a CC-BY 4.0 reuse licence.

If access is restricted, please specify who will be able to access the data and under what conditions.	
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 RDR ☒ Other data repository (specify): - All scripts will be available on a personal gitlab account (https://gitlab.com/ybawin). - All sequencing data will be deposited at the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/browser/home). - Biodiversity data will be made available via GBIF (https://gbif.org/). ☐ Other (specify)
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)
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
What are the expected costs for data sharing? How will these costs be covered?	No additional costs expected for data sharing.

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	The PI (Steven Janssens) and the day-to-day manager of the project; currently: Yves Bawin.
Who will manage data storage and backup during the research project?	The PI (Steven Janssens) and the day-to-day manager of the project; currently: Yves Bawin; both in collaboration with the ICT service of Meise Botanic Garden.
Who will manage data preservation and sharing?	The PI (Steven Janssens) and the day-to-day manager of the project; currently: Yves Bawin.
Who will update and implement this DMP?	The end responsibility for updating and implementing the DMP is with the PI, Steven Janssens.