1166125N - ECOGraph - A graph-based approach to species distribution modelling
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**
GeoPlant: presence-absence data	Plant species occurrences (only presence-absence)	Reused	Digital	Observational (compiled)	.csv	< 1 GB	
GeoPlant: environmental predictor data	Predictor rasters, values, satellite images & cubes associated with the species occurrences	Reused	Digital	Compiled/aggregated	.tif, .csv, .jpeg, .pt	< 100 GB	
GeoPlant/Malpolon: data loaders & baseline models	Code for loading data and modelling baselines	Reused	Digital	Code	.py/.ipynb/.ckpt/.yaml	< 10 MB	
Plant species presence-absence data	Full-track (non-anonymized, presence-absence) plant occurrence data from databases (e.g. GBIF, EVA, Grassplot, sPlot)	Reused	Digital	Observational	.csv	< 1 GB	
Scripts	Scripts for data loading, processing, model architectures	Generated	Digital	Code	.py/.ipynb (maybe .R)	< 10 MB	
Landscape graphs	Output of graph generation from raster data	Generated	Digital	Compiled/aggregated	.shp, .txt, .gpkg (or others, depending on processing)	< 100 GB	
Plant community graphs	Output of graph generative models & SDMs	Generated	Digital	Compiled/aggregated	.shp, .txt, .gpkg (or others, depending on processing)	< 100 GB	
Environmental predictor data	Predictor data other than from the GeoPlant dataset, or extended upon. From similar sources.	Reused	Digital	Compiled/aggregated	.tif, .csv, .jpeg, .pt	< 100 GB	

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:

- GeoPlant data: https://www.kaggle.com/datasets/picekl/geoplant or http://arxiv.org/abs/2408.13928 (related deep learning Python framework MALPOLON: https://github.com/plantnet/malpolon)
- Presence-absence occurrence data: GBIF (https://www.gbif.org/), European Vegetation Archive (EVA; https://euroveg.org/eva-database/), sPlot (https://www.idiv.de/research/projects/splot/), Grassplot (https://edgg.org/databases/GrassPlot)

- Environmental predictor data (other than from GeoPlant, non-exhaustive): land cover & elevation from NASA Earthdata portal (https://search.earthdata.nasa.gov/search), soil from Soilgrids (https://soilgrids.org/), satellite data (Landsat, Sentinel-2) from Ecodatacube (https://stac.ecodatacube.eu/), climate from CHELSA (https://chelsa-climate.org/)...

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

The only aspect relevant here may be that locations of sensitive/vulnerable species will be processed (for Plant species presence-absence data (self-compiled) from databases). This can easily be solved by anonymizing species, as is already the case for GeoPlant data and will probably be a condition for the use of those data.

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.

- Yes

Yes -> probably: please see the ethics question as well. Plant species occurrence data derived from databases may be subject to dissemination restrictions before use, if the species are vulnerable/sensitive. This will not affect the research.

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).

Documentation and adding metadata will happen at several levels to keep data FAIR:

- Overall documentation (.txt file) on the project (a content file with a general overview of the data and where the used data and backups are found).

- Documentation & metadata at the level of different data types/datasets by using frequent semi-automated creation of README.txt files, containing a.o. information below additional to the KUL template, which are some of the most relevant contents for all data (no scripts):
  - Data source
  - Used scripts for processing
  - Date
  - Units
  - Coordinate reference systems/projections/extent
  - Variable explanations
- Main procedures are scripts:
  - In the scripts, concise but sufficient comments will be included to explain the processes, as well as an overall explication.
  - Overview of packages & dependencies by a conda environment yml file
  - When new functions/classes are defined, parameters and variables will be explained in the docstring.
- A GitHub repository will be used for version control and will be accompanied by the same (possible some additional) README files.

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

In the fast-track phase of the project, the GeoPlant data will be used, which already exists. Main project outputs will be models and thus scripts, to which no direct standards will be associated. Similar to above, language and packages (e.g. conda environment) will be associated, as well as other computational requirements. When creating the full-track data, metadata will describe data sources, processing steps and formats.

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

Where will the data be stored?

The main storage place is the KU Leuven personal OneDrive, to the extent that storage quota allow at the moment. If necessary, other data can be stored on the KU Leuven network drives or Sharepoint Lab sites after discussion with the supervisor.
In general, it should be noted that currently, no extra storage is required and if this would change, volumes will stay relatively low.

How will the data be backed up?

Two additional storage sites will be used for data and script backups, to be made at regular intervals:

- The KU Leuven Personal I: Network Drive (up to 50 GB), to be extended by the KU Leuven Large Volume L: Network Drive if necessary, with consent of the supervisor
- An external hard drive purchased for backups (LaCie 2 TB)

I do not expect to need extra backups for very large raster files, as I will not generate them myself, but use them from existing sources, which will act as a backup themselves.
The main output scripts will be held at a GitHub repository as well.

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

Although the total storage volume from the data table sums up to +/- 500 GB, this should be a large overestimation as a result of

the intervals used. I expect that the total data volumes will not exceed the 250 GB of OneDrive, or even will be much lower. The I: Network data (50 GB) should be sufficiently large initially, but can be extended (by other network drives) as mentioned earlier. The 2 TB hard drive should definitely fulfill storage requirements and will be kept at the office. My main outputs will however just stay scripts, whereas a lot of the data does not need a backup as it originates from other sources.

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

All storage spaces will be either personal or within the lab at KU Leuven, with managed access only for supervisors and collaborators when they would have a contribution. The hard drive is kept at the office, which is locked when nobody is present.

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

Expected costs should not exceed those for 1 TB/year, assuming at max ~ €100/year, or probably less depending on the need for and type of storage solutions used. These costs can be covered by part of the FWO project budget.

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...).

It should be possible to preserve all data for the 10 year KU Leuven RDM policy, as it consists of only digital material. The main data types will be the models and codes, whereas datasets will only be preserved if they are self-compiled.

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

Most materials will be stored on the KU Leuven RDR repository where possible, code from GitHub will be kept there as well.
If in need of other storage solutions over time, this will be discussed together with the supervisor (but not expected to).

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

Most data preservation should remain free, but as mentioned above, extensions remain to be seen later on in the project. In the case storage fees would be required, a budget can be made available.

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 a restricted access repository (after approval, institutional access only, …)
- Yes, in an Open Access repository

- Final models and scripts for data preparation wil be made openly available.
- Self-compiled data may either be made openly available and/or with restricted access

If access is restricted, please specify who will be able to access the data and under what conditions.

- Self-compiled occurrence data may comprise occurrences of rare species, which may be vulnerable. If the data would be made open access, species names would be made anonymized, whereas non-anonymized species data could be requested under specific conditions. This will depend on the final format and possible own agreements with 3rd party data providers.

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.

- Yes, Ethical aspects
- Please see the explanation above about the vulnerable species occurrences.

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

- Code candidate repositories: RDR, GitHub
- Occurrence candidate repositories (to be decided on if necessary): Seafile, Pangaea

Finished state if:

- Models & code are optimized and working properly. Adaptations can always be made on GitHub using version control, whereas other data will only be uploaded on other repositories at the end of the project or if a part of it (work package) is considered finished.

When will the data be made available?

(Pleas also see above:)

- On GitHub: as soon as code is properly working, as it can be regularly updated
- Occurrence/graph data: at the end of the project or if they are finished to a standard shape (e.g. similar to the GeoPlant dataset.

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

For code: licences compatible to the used packages/software..., probably one of GNU General Public License, MIT or 3-Clause BSD License.
For occurrence/graph data: again dependent on the sources, but open data: probably CC-BY.

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?

This remains to be seen. For models and code, free repositories will be used.

6. Responsibilities

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

Reinout Vandenabeele

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

Reinout Vandenabeele

Who will manage data preservation and sharing?

Reinout Vandenabeele & Stef Lhermitte

Who will update and implement this DMP?

Reinout Vandenabeele