Project[1] Number: [741413

Project Acronym: [HOPE]

Project title: [Humans on Planet Earth]

DATA MANAGEMENT PLAN

[1] The term 'project' used in this template equates to an 'action' in certain other Horizon 2020 documentation

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of the data collection is to synthesise, as far as is possible, global fossil pollen data, identify evidence of human activity ("events") in each individual pollen dataset and assess ecosystem properties before and after the occurrence of human activity. These analyses will provide the necessary information to address HOPE's main objectives and test its main hypothesis.

What types and formats of data will the project generate/collect?

The project will generate several types of data:

A) Fossil pollen database – compiled fossil pollen datasets from public and non-public sources
B) R code and packages – all methods (R code) used to process and analyse the data and produce the outputs
C) Processed data – state-of-the-art curated fossil pollen sequences on a global scale (as far as is possible)
D) Taxonomic harmonisation tables – necessary for standardisation of taxonomy of the pollen taxa before analyses
E) Database of human events – the identification of human activity in the assessed fossil pollen sequences
F) Archaeological database – a global compilation of archaeological radiocarbon dates gathered from publicly and non-publicly available sources
G) Result outputs – since HOPE data are characterised by a complex database structure (multiple tables for a single sequence) and most work is conducted in the R environment (R Core Team, 2021), all data outputs use R objects (".rds" files).

Will you re-use any existing data and how?

Yes, we will re-use (i) palaeoecological datasets currently compiled in the global centralised database called Neotoma, (ii) data from the global Pangaea data portal, and (iii) archaeological data of radiocarbon dates for estimating human densities from the PEOPLE3K working group of the Past Global Changes (PAGES) project.

What is the origin of the data?

A) Fossil pollen database:
    a) Neotoma Paleoecological Database – an open-access multiproxy community database for the Quaternary-Pliocene - https://www.neotomadb.org/
    b) Pangaea – open-access data publisher for Earth and Environmental Science - https://www.pangaea.de/
    c) Indo-Pacific database – personal contact with collaborators
    d) Data from Latin America gathered by personal contact with collaborators based on a literature review by Flantua et al. 2015[#]
    e) Data from China obtained from collaborators in China based on an overview of potential pollen datasets from China published in Herzschuh et al. (2019)*
B) R code and packages: produced by HOPE
C) Processed data: produced by HOPE
D) Taxonomic harmonisation tables: produced by HOPE
E) Database of human events: produced by HOPE
F) Archaeological database: publicly available databases such as CARD (upon request) (https://www.canadianarchaeology.ca/) and from published studies

[#Flantua, S.G.A., et al. 2015. Updated site compilation of the Latin American Pollen Dataset. Review of Palaeobotany and Palynology 223: 104-115. https://doi.org/10.1016/j.revpalbo.2015.09.008

*Herzschuh, U., et al. 2019. Position and orientation of the westerly jet determined Holocene rainfall patterns in China. Nature Communications 10: 2376. https://doi.org/10.1038/s41467-019-09866-8]

What is the expected size of the data?

A) Fossil pollen database: up to 50 MB
B) R code and packages: up to 50 MB
C) Processed data: up to 1.5 GB
D) Taxonomic harmonisation tables: up to 5 MB
E) Database of human events: up to 5 MB
F) Archaeological dataset: up to 5 MB
G) Database of the age-depth models for each processed pollen dataset: up to 50 GB

To whom might it be useful ('data utility')?

To the palaeocommunity, macroecologists, biogeographers, archaeologists, palaeontologists, computer scientists, climate modellers, museums, government agencies, independent research consortia, and the general public.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Yes, they are.

What naming conventions do you follow?

We created a naming convention for HOPE purposes which is made available with our documentation of the databases.

Will search keywords be provided that optimize possibilities for re-use?

Yes. For the R code produced by this project, keywords will be provided in the Zenodo archives of Github repositories.

Do you provide clear version numbers?

Yes, for both data compilations and R packages and script.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

A) Fossil pollen database – metadata will specify the source of datasets, accessed data, source information, e.g. contact PI and reference publication.
B) R code and packages – R packages will be well documented according to recommended standards and have a clearly defined author and maintainer. Each R code will have a purpose and date as well as names of contributors.
C) Processed data – datasets will include a metadata file with date of processing and all configuration criteria used to produce such data.
D) Taxonomic harmonisation tables – metadata will describe the process of creation of each harmonisation table.
E) Database of human events – metadata will describe the process of detection of human events in each continent.
F) Archaeological database – metadata will include the date of compilation of the dataset and sources.
G) Age-depth models – metadata will describe the building process of the age-depth model for each processed dataset.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

A) Fossil pollen database – data from Neotoma, Pangaea, and some publications are open and comprise the bulk of the data HOPE works with. Data from private sources (China and Latin America) cannot be made public by us as this was a requirement set by the data providers when allowing us to use their data.
B) R code and packages – all data will be publicly available.
C) Processed data – data created from public sources will be publicly available at the end of the project (see A).
D) Taxonomic harmonisation tables – data will be publicly available at the end of the project.
E) Database of human events – data will be publicly available at the end of the project.
F) Archaeological database – data will be publicly available at the end of the project.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

Not applicable

How will the data be made accessible (e.g. by deposition in a repository)?

A) Fossil pollen database – most data are already available via online databases. Other files will be stored in ZENODO.
B) R code and packages – all code and packages will be stored in Github.
C) Processed data – all processed data created via public data be stored in ZENODO.
D) Taxonomic harmonisation tables – data will be stored in ZENODO.
E) Database of human events – data will be stored in ZENODO.
F) Archaeological dataset – data not already publicly available will be stored in ZENODO.

What methods or software tools are needed to access the data?

A) Fossil pollen database – R
B) R code and packages – R
C) Processed data – R
D) Taxonomic harmonisation tables – software that can read spreadsheets
E) Database of human events – R
F) Archaeological database – R

Is documentation about the software needed to access the data included?

No

Is it possible to include the relevant software (e.g. in open source code)?

Yes

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

A) Fossil pollen database – most data are already available via online databases. Other files will be stored in ZENODO.
B) R code and packages – all code and packages will be stored in Github.
C) Processed data – all processed data created via public data be stored in ZENODO.
D) Harmonisation tables – data will be stored in ZENODO.
E) Database of human events – data will be stored in ZENODO.
F) Archaeological database – data not already publicly available will be stored in ZENODO.

Have you explored appropriate arrangements with the identified repository?

Yes

If there are restrictions on use, how will access be provided?

Some data from China are restricted and requests for their use should be made to Professor Fahu Chen at the Chinese Academy of Sciences in Beijing. Alternatively, most of the private data from China can also be accessed from the published sources (Herzschuh, U., et al. 2019, https://www.nature.com/articles/s41467-019-09866-8#Sec10/). Some data for Latin America are restricted and requests for their use should be addressed directly to the authors of the corresponding data.

Is there a need for a data access committee?

No

Are there well described conditions for access (i.e. a machine readable license)?

Not applicable, all data (with the exception of some data from Asia and Latin America) are open access.

How will the identity of the person accessing the data be ascertained?

NA

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Yes

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

We shall use existing standards and vocabularies in the R environment.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Yes

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Yes

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

A) Pollen database – most of the datasets in the database are already available via online databases (CC BY 4.0 license). Other files will be similarly licenced
B) R code and packages – all code and packages will be stored in Github under a MIT License
C) Processed data – all processed data created from public data will be under a MIT License
D) Harmonisation tables – data will be under a MIT License
E) Database of human events – data will be under a MIT License
F) Archaeological database – data will be under a MIT License

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data, which will be made public, will be available at or soon after the end of the project.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?

The data, which will be made public, will be available at or soon after the end of the project.

If the re-use of some data is restricted, explain why.

For some datasets, we do not have the right to distribute them. This applies specifically to some data sets from Asia and Latin America, where other initiatives have compiled the databases and made these available for the project, but are not yet public access due to pending publications.

How long is it intended that the data remains re-usable?

Forever

Are data quality assurance processes described?

We will publish a series of methodological and data papers describing the data and data generation process, with a focus on data quality.

Further to the FAIR principles, DMPs should also address:

3. Allocation of resources

What are the costs for making data FAIR in your project?

Not sure!

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Covered by overhead costs at the University of Bergen

Who will be responsible for data management in your project?

Dr Alistair Seddon <alistair.seddon@uib.no>

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

No

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

ZENODO and Github have provisions in place. The data are not sensitive. The data are also stored locally at the University of Bergen, with backup routines.

Are the data safely stored in certified repositories for long term preservation and curation?

Yes

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

None that we are aware of.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Not relevant.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

We keep within the guidelines of the Norwegian Research Council and the University of Bergen.

7. Further support in developing your DMP

The Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.

The EUDAT B2SHARE tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

Useful listings of repositories include:

Registry of Research Data Repositories

Some repositories like Zenodo, an OpenAIRE and CERN collaboration, allow researchers to deposit both publications and data, while providing tools to link them.

HISTORY OF CHANGES		
Version	Publication date	Change
1.0	02.11.2017	▪ Initial version
1.2	03.02.2021	▪ Updated version to accommodate some private data where there were gaps in the publicly available data
1.3	01.07.2021	▪ Updated version to accommodate details about differences in data types (R code, processed data, archaeological data).

Other useful tools include DMP online and platforms for making individual scientific observations available such as ScienceMatters.