GELIFES DATA REPOSITORY - PROCEDURE

Version 3 - January 2015

1 PURPOSE

All data published within the institute[1] will be archived for purposes of both verification (safeguarding scientific integrity) and safekeeping of valuable datasets (remote backup). This is done to accommodate increasing awareness that institutes should always keep copies of original datasets, which can be used in case of doubt of scientific integrity of members of the institute. In addition, safe storage of original data is important to prevent unwanted loss of datasets, for example in case of fire damage, loss or theft of data storage media and archives, and to safeguard the possibility of re-analyses or re-use of datasets.

2 WHICH STUDIES REQUIRE A DATA DEPOSIT AND WHICH NOT

The following projects always require a data deposit in the institute's data repository:

1. All publications in a scientific journal or book where the institute or a research group within the institute is the work address of the first or equal author (also in case of multiple addresses of the first or equal author).
2. All MSc reports done within the institute, supervised by an institute staff member
3. External MSc studies done at another institute, but with an institute staff member being responsible for the final grading and/or when the student is registered with the RUG
4. All PhD theses done within the institute, with an institute professor as first promotor
5. All external PhD theses (e.g., done at a KNAW or NWO institute) with an institute professor as first promotor

NOTE: chapters in a PhD thesis that have been previously published in a journal/book and have been archived as a publication, must be archived (again) as part of the PhD thesis archive. Similarly, chapters of a PhD thesis that have been archived as part of the PhD thesis archive and are published in a journal/book at a later moment, must be archived again as a publication.

For the following studies a data deposit in the repository is voluntarily:

1. All other unpublished studies done within the institute (e.g. pilot studies)
2. Publications with a first author from a different institute
3. PhD theses with a non-institute professor as first promotor
4. External MSc studies where the student temporarily registers at a different university for the duration of the study

3 WHO IS RESPONSIBLE FOR DEPOSITING THE DATA ARCHIVES

1. MSc students deposit the data with their daily supervisor.
2. PhD students and postdocs deposit the data with the data manager responsible for their group when publishing a paper.
3. Staff members (assistant, associate, full professors) deposit the data with the data manager for their own publications and of the PhD thesis archives that are handed in with them, and for the MSc projects that they supervise.

[1] Note: 'institute' refers to both GELIFES and its predecessors CEES and CBN

4. The data manager checks the archives for completeness and file integrity, and uploads the archive to the repository.
5. The data manager administrates the **Data Management Plans** of all required projects and monitors the punctual delivery of data archives according to the DMP.
6. The data manager reports to the director of the institute.
7. The data managers are Joke Bakker (institute coordinator and repository admin; former CEES groups Microbial Ecology, BESO, Theoretical Biology, Evolutionary Genetics); Jan van den Burg (former CEES groups Animal Ecology, Cocon, MarBEE, MarEcon, Plant Ecophysiology; Leon Steijvers (former CBN groups Behavioural Biology, Behavioural Physiology, Chronobiology, Molecular Neurobiolgy, Neuroendocrinology).

4 WHERE & HOW LONG IS THE DATA STORED

The repository is safely stored in the central data wiki on the RUG webhosting server:

http://gelifesdata.webhosting.rug.nl/ , which is backed-up daily.

Regular data archives will be stored for a period of at least 10 years; long-term databases or other important data sets can be stored for longer periods.

5 WHO HAS ACCESS TO THE DATA

Data archives are stored **per senior staff member responsible for a group of junior researchers** and **per year within such PI-groups.** The data archives are only accessible by the senior staff members of a PI-group; each PI-group has only access to their own data.

Staff members can view (= download) data files from the archive of their projects, but not edit or remove any uploaded file. When additions or corrections to a file are needed, a new version or supplement must be uploaded where the initial version is preserved.

Requests for data by external scientists are to be judged by the original contributor of the data.

6 WHEN TO DEPOSIT THE DATA

FOR REGULAR PUBLICATIONS IN PEER-REVIEWED SCIENTIFIC JOURNALS AND BOOK CHAPTERS:

Each MSc, PhD student, postdoc and other researchers (including professors) should compile a documented archive of all data underlying a publication and send this in for archiving in the repository **within 3 months after the paper appears online**, including 'early online' publication.

FOR DATA COLLECTED IN THE CONTEXT OF AN MSC STUDY:

All data should be deposited no later than the date of handing in the final version of the MSc report. A grade for the project will only be awarded when all data (including documentation) are deposited with the daily supervisor of the project (PhD student, postdoc, staff member etc.). The daily supervisor should deposit the data in the repository no longer than **1 month after the grade has been awarded** to the student.

FOR DATA COLLECTED IN THE CONTEXT OF A PHD STUDY:

The documented data archive of the study should be deposited with the promotor upon handing in the final manuscript for the manuscript committee. The promotor will only sign the approval form of the PhD thesis when the data archive of the study has been handed in. Also parts/chapters that have already undergone a previous storage procedure upon publication of the paper are to be included in the PhD study archive as one of the folders

for each chapter. The promotor deposits the data archive of the thesis in the repository **within 1 month after the thesis has been handed in**.

7 HOW TO DEPOSIT THE DATA

The data are deposited by sending an email with the full reference of the publication, and the zip file of the data archive to gelifes-data@rug.nl. For sending larger files (>5 MB) the service http://bars.rug.nl can be used, or other file transfer services such as Unishare (preferred), DropBox, Google Drive, WeTransfer.

For MSc ad PhD thesis archives, start the file name always with a standard prefix:

PhD_thesis_<lastname>_<year>.zip
MSc_thesis_<lastname>_<year>.zip

For papers the prefix can be omitted, so directly:

<author>_<journal>_<year>.zip

<author-1>_<author-2>_<journal>_<year>.zip

<author>_etal_<journal>_<year>.zip

Provide with each data archive that is submitted a complete, formatted reference of the publication in the metadata file:

- For MSc theses: include the s-number of the student, the project title and the names of the supervisors.
- For PhD theses: add the thesis title and names of the (co-)promotors.
- For journals and book chapters, use the Annual Review of Ecology, Evolution and Systematics journal format for the bibliographic reference. Do not abbreviate the author list to et al., and include the DOI of the paper (can be found through Web of Science).

To ensure general accessibility of the archives, all file names, metadata and other description files and comment lines in code must be in **English**. File names should **not** contain special characters other than dashes (-) or underscores (_); use underscores instead of spaces.

8 HOW TO ORGANIZE THE DATA ARCHIVE

Data storage will be organised per publication, i.e. MSc students' theses, PhD theses and papers in scientific journals, in a single .**zip archive per publication** and should include:

1. the **final version** of the document,
2. all **primary (raw) and secondary (processed) data** underlying the document. Primary data include a scanned pdf of original data sheets, field books etc. by preference; if this is not possible raw data should be at least included in some electronic way, i.e., in a table, spreadsheet or text file,
3. all **program code and scripts** used to produce the final results such as figures, tables, statistical analyses etc.,
4. **relevant metadata**: a text file describing the data sources in relation to (corresponding sections of) the document,
5. for PhD theses: these four elements **per chapter**.

PRIMARY DATA INCLUDE ALL SOURCES OF RAW DATA:

1. scanned field logs, lab journals, score forms,
2. pictures of gels, microscopic observations,
3. output from data loggers,

4. video and audio recordings,
5. webcam/photo identification files: only when the resulting IDs are NOT included in other primary data sets, e.g. in field journals or data files,
6. sequencing and genotyping data,
7. micro array and hi throughput data: only if NOT stored in public database on publication.

SECONDARY DATA INCLUDE ALL PROCESSED DATA FILES, PROGRAM CODE AND SCRIPTS USED IN THE PREPARATION OF THE DOCUMENT:

1. spreadsheets, databases, graphics,
2. output from statistical packages,
3. output of geographic information systems,
4. simulated datasets: only if NOT possible to reproduce from program code.
5. program code in C/C++, NetLogo, Matlab, Maple, Mathematica etc. of all programs developed to produce the published results & all associated parameter files,
6. R scripts (statistical analysis, graphs, etc.), Python or other batch scripts used for data processing,
7. specific program code/scripts, e.g. Z-Tree or other software packages used to produce or process primary data.

DATA COLLECTED AND/OR STORED EXTERNALLY:

1. Data that is collected and stored at an external institute, falls under the responsibility of the external institute and need not be deposited in the repository; however, this is ONLY the case for raw, primary data! The database and institute should be referred to in the corresponding metadata file (read_me_first.txt) of the archive. All processed, secondary data such as spreadsheets, databases, scripts, code etc. that is used for the thesis/publication must be included in the archive file.
2. Large primary data sets such as sequencing data that are stored elsewhere in a public database need not be deposited in the repository; again, this is ONLY the case for primary data. The link to the storage location should be included in the metadata file. All secondary data must be included in the archive file.
3. Large long-term databases that are used for many projects need not be stored with each project. The raw/primary data, i.e., scans/photos of lab and field journals, should be stored once, and can be updated yearly with new contributions if applicable. The relevant version of this database archive should be referred to in subsequent project/publication archives. Secondary data must always be included in each project/publication archive file.

DOCUMENT:

1. final version of MSc thesis, both as a text file (rtf, doc, docx) and a as a pdf version,
2. final version of PhD thesis, both as a text file (rtf, doc, docx) and a as a pdf version,
3. final version of journal paper as a pdf file,
4. optional for archiving and follow-up purposes: RIS-format file of all literature references used in the project.

METADATA:

read_me_first.txt file to be included in each folder in the zip archive; this file should contain info on the folder's contents and how these relate to the document:

1. the description/reference of the publication (for students: include your s-number; for publications: include the doi);
2. an overview of the contents of the zip archive and how this relates to the publication: indicate how the data was processed, which data was used in which chapter & to produce what results/plots, or which parameters were used to generate the simulated datasets used for the publication;

3. when including other file formats than listed in section 2.9: indicate which computer programs were used plus link to source site when relevant;
4. if relevant: a short description plus links for data that is archived externally;
5. contact details of the author

9 PREFERRED DATA FORMATS

The strongly preferred way of storing all data is as tab- or comma-delimited text files with variable names in the first line, with an associated R script that reads the data file, as this makes data robust towards future changes in software and data file formats. For other data types, consider using the suggested file formats below (based on the KNAW-DANS Preferred Formats overview, May 2013) for similar reasons of compatibility and future accessibility:

Do not use spaces in file names, but use underscores_ instead. If you use special file types that cannot be stored in a general format, consider uploading a copy of the associated software package.

Data type	**Preferred format**	**Acceptable format**
Documents	PDF/A (.pdf) Unicode TXT (.txt)	OpenDocument Text (.odt) MS Word (.doc, .docx) Rich Text File (.rtf) PDF (.pdf) Non-unicode TXT (.txt)
Spreadsheets	PDF/A (.pdf) Comma separated values (.csv)	OpenDocument Spreadsheet (.ods) MS Excel (.xls,.xlsx)
Databases	ANSI SQL (.sql) Comma separated values (.csv)	MS Access (.mdb, .accdb) dBase III or IV (.dbf)
Statistical data	R	SPSS Portable (.por) SAS transport (.sas)
Audio	WAVE (.wav)	MP3 AAC (.mp3)
Video	MPEG-2 (.mpg, .mpeg, …) MPEG-4 H264 (.mp4) Lossless AVI (.avi)	QuickTime (.mov)
Pictures (raster)	JPEG (.jpg, .jpeg) TIFF (.tif, .tiff)	
Pictures (vector)	PDF/A (.pdf) Scalable Vector Graphics (.svg)	Adobe Illustrator (.ai) PostScript (.eps) PDF (.pdf)
Geographical data	Google Earth Keyhole Markup Language (.kml, .kmz) Geographical interchange standard geoTIFF (.geotiff)	ESRI Geodatabase ESRI Shapefiles (.shp and accompanying files) ERDAS Imagine (.img)

10 SUMMARY: HOW TO SET UP A REPOSITORY DATA FILE

For all MSc projects, PhD projects and publications with first author (also) in GELIFES, all data must be deposited in the repository. This includes all data, to be collected in a .zip file:

1. Primary (raw) data, i.e. all measurements, pictures, scans of lab and field books, score forms, databases, etc.
2. Secondary (processed) data, i.e. spreadsheets, statistical data, plots, figures, etc.
3. Program code & scripts, e.g. R-scripts used for statistical analyses or plots with the relevant parameter files, Mathematica, MatLab, C++ code
4. Manuscript: as MS Word and pdf file, plus all supplementary material
5. ****VERY IMPORTANT** metadata**: read_me_first.txt file with:
    a. the description/reference of the publication (for MSc students: include your s-number; for publications: include the doi);
    b. an overview of the contents of the zip archive and how this relates to the publication (indicate how the data was processed, which data was used in which chapter & to produce what results/plots);
    c. which computer programs were used plus link to source site when relevant;
    d. if relevant: a short description plus links for data that is archived externally;
    e. contact details of the author/supervisors

6. for a PhD thesis: the above for each chapter, plus a Word/pdf document of the complete thesis:
    - If a chapter has been published previously, you can of course re-use the data file of that particular publication.
    - If a chapter is to be published at a later moment, it must be archived again, since usually the final accepted publication will be different from the first draft manuscript.

This manual is available in more detail and with examples on our website: http://www.rug.nl/research/institute-evolutionary-life-sciences/data-management/ (also for download).

The complete zip file should be sent to gelifes-data@rug.nl , if necessary via Unishare, WeTransfer or DropBox in case of very large files.