RESEARCH DATA MANAGEMENT ESRIG

INDEX

1. SUMMARY AND ACTION POINTS

- As of September 2015, a Research Data Management Plan (RDMP) is obligatory for all research projects (MSc, PhD, postdoc and other) carried out within ESRIG. A RDMP safeguards documentation of data of all publications derived from the project and of unpublished results that are to be published at a later stage.
- The Research Data Management Plan (RDMP) is the starting point for all research projects in the institute where a senior ESRIG staff member (PI) holds final responsibility. The RDMP should describe the collection, processing, storage and archiving of all project data.
- The senior staff member (PI) is ultimately responsible for a research project. Therefore, he/she has the final responsibility for correct and timely planning, management and archiving of all project data.
- All primary and secondary data of a project are to be archived in the institute repository located on the Y-drive according to the procedure described in ANNEX I; archives will be stored for at least 10 years. Alternatively, data may be stored in the specially designed database structure maintained by the research group CIO.
- The purpose of the institute data repository is archive-only; access to archive files is restricted to senior staff members ultimately responsible for the corresponding research projects.
- Central administration of RDMPs, monitoring of data archive deposits and maintenance of the repository is coordinated by the institute's scientific coordinator.
- Time-line: RDMPs will be installed as of the academic year 2015/2016 for new researchers (PhD, postdoc) and new Master projects. For current projects (PhD, postdoc, Master projects) they are not mandatory, since the requirement for a RDMP should be communicated at the beginning of a project, and therefore included in TSPs and MSC guidelines.
- The scientific coordinator will be hired ASAP at the beginning of the academic year 2015/2016. RDMP implementation will be evaluated by a RDMP-MT on a yearly basis. The first evaluation will be done January 2016.

2. BACKGROUND

ESRIG is a multidisciplinary research institute, consisting of 6 research groups. Within ESRIG research data needs to be properly archived. The proposed Research Data Management Plan (RDMP) is meant for purposes of both verification (safeguarding scientific integrity) and safekeeping of valuable datasets. Scientific integrity is protected by assuring that all (published) scientific results can be traced back to the original raw data in an unambiguous way. Furthermore, RDMP's guarantee that all research data produced within the institute or group are available for follow-up research. In almost all cases, ESRIG acts as the legal representative of the University of Groningen and holds ownership of all research data generated and stored within the institute. If however this cannot be the case due to contractual obligations or collaborative work with colleagues from other institutions, the RDMP explicitly addresses this ownership situation.

This research data management plan describes how every scientist (and some of the research technicians) within ESRIG must deal with research data during his/her research and once the research project has been completed.

According to university/faculty rules (and in compliance with many funding agencies) each research project should have its own RDMP. Mostly, a research project is defined as a PhD project, a Postdoc project or a Master project (when not directly associated with a PhD or postdoc project). In the case of a PhD project, the TSP should refer to an existing RMDP, in the other cases (Postdoc or Master projects), the PI or supervisor should make sure that a RDMP is prepared. In general, all new researchers should be clearly instructed about RDMP obligations (by their PIs or supervisors).

To make this task as easy as possible, ESRIG has developed a template RDMP, which can easily be adapted to specific research projects or other individual needs. It furthermore identifies persons who have specific responsibilities (supervisors, data manager).

ESRIG's template RDMP contains a General project information form and a Questionnaire to safeguard the above two goals. Further tips and instructions are described under point 3 (General data policy and RDPM Guidelines). In fact the RDPM does no more than establishing ways of conduct that are (should be) quite common in research. This means a RDMP can be relatively short. However, if we use the obligation to have an RDMP available to also organise our data archiving system, the RDMP gets more detailed, but also more useful. For data archiving, the former research institute CEES has already made a well-designed (and practical) instruction for all their researchers and students, and within (now) GELIFES these RDPM's have been used as a pilot since 2013. Furthermore, our Research Data Management Template is based (with adjustments) on the established RDMP that was developed for the three technical universities in the Netherlands. Yet, next steps will need to include work in the realm of data archiving arrangements, and, of course, active surveillance by the research supervisors as well as the new scientific coordinator of ESRIG. This latter person needs to ascertain that tasks associated with the RDMP are, in fact, performed.

3. GENERAL DATA POLICY AND RDPM GUIDELINES

Each project and on-going activity within ESRIG, including but not only MSc and PhD projects, is covered by this RDMP. All researchers, technicians and students active within ESRIG are supposed to know the contents of this RDMP and to act accordingly. New researchers and students should be briefed by their supervisor and informed as part of the curriculum. From the start of their projects, researchers and students need to comply with the points given below.

A. Data Collection

The type of scientific output that requires data deposit
For the following types of scientific output a data deposit is mandatory:

1. All publications in a scientific journal or book where ESRIG or an ESRIG group is the work address of the first author (also in case of multiple addresses of the first author)
2. All MSc reports generated within ESRIG, supervised by an ESRIG staff member
3. External MSc reports done at another institute (mostly, but not exclusively, EES master thesis students), but with an ESRIG member as first supervisor, i.e. being responsible for the final grading, unless other written arrangements have been made with the other institute at the start of the project.
4. All PhD theses executed within ESRIG, with an ESRIG professor as first promotor
5. All external PhD theses (e.g. done at a KNAW or NWO institute, or in industry) with an ESRIG professor as first promotor, unless other arrangements have been made with the other institute at the start of the project.

For all other publications with ESRIG involvement (such as second promotor, co-authorship, etc) storage is recommended, but voluntary.

ESRIG groups produce a high variety of data types, so data need to be clearly defined. For experimental work, **primary**, **"raw data"** are those from some measurement instrument prior to any data analysis process. Subsequently, depending on the necessary data analysis treatment, there are various forms of **intermediate** data, until the **final** data (including the published manuscript itself) are those that typically occur in scientific presentations of the work (oral or written). Modelling usually depends on input data from other sources. Crucial for the modelling work is the code which transfers these input data into the final model outcome data. It is obvious that the code itself, along with its input parameters and other "choices" is as crucial for the work as the input and output data. This model thus has to be stored along with the data in order to preserve the work.

In the RDMP, a short description of the data needs to be given, including the (prospected) amount and content. If possible a rough estimate of the number of files is given. What type of data are generated impacts how the data will be managed and how long it needs to be preserved. There are four main types of research data (where often also **primary, intermediate, final** data can be distinguished):

- **Observational data (incl. monitoring)**: captured in real time, typically cannot be reproduced exactly

- **Experimental data**: from labs and equipment, can often be reproduced but may be expensive to do so

- **Simulation data**: from models, can typically be reproduced if the input data, the model code (*) and design of simulation experiment (setup of parameter space and method of exploration (selecting and executing simulation runs)) are known

- **Derived or compiled data**: after data mining or statistical analysis has been done, can be reproduced if analysis is documented

Data types could include text, numbers, images, 3D models, software, audio files, video files, reports, surveys, etc. (*): the Model code includes either the complete code in MatLab, R, Excel, or the code used as “middleware”, e.g. in System Dynamics model (in e.g. Stella), Agent-Based Model (e.g. in NetLogo or other), scenario model in PowerPlan, Times etcetera.

Handling of primary data

Each researcher, including OBP working within a defined project, is obliged to write down his/her daily achievements, decisions, (experimental or model) try-outs and anything else of importance in a laboratory logbook. Classically, these are notebooks with pre-numbered pages, none of which may be removed. More and more, however, notes are being taken electronically, usually in a word processing document or a spread sheet. While this lacks the traceability and thus integrity of a paper laboratory notebook, the advantages are obvious. The best of both worlds would be an electronic laboratory notebook. Software Packages for that exist, but while such a system is not yet in place, ESRIG tolerates the use of electronic files for taking daily notes. However, they must be regularly stored, and backed-up. These files (and papers) are an important part of the research work package that needs to be stored.

Primary data include a scanned pdf of paper laboratory logbooks, or the electronic form of this (the above-mentioned daily notes in word processing files). If the primary data are generated by own (experimental) work, they should either be stored in some electronic way, (in a table, spreadsheet or text file), or should be traceable in a well-maintained and documented laboratory database. If primary data are from elsewhere (for example large data sets from other repositories) and these cannot be stored in the repository, there should be an unambiguous reference to these data (origin, way to get to them, version number etc.) Primary data also include the development of computer programs, input parameters and scripts used for modelling activities. If third party computer programs are in use, these should be properly and uniquely identified (origin, name, type, version number etc).

Finally, a special form of experimental data are on-going observations, usually called "monitoring". What makes them different from "normal" experiments, is that the work is never completed, and past data can be re-calibrated/adapted according to increasing knowledge of the observational process. For such data a database is the natural storage mode, which should contain the raw data as well as the data treatment process along with its modifications over time. Especially, data published or presented to international fora/data collections should be "frozen" in the database as a specific version. Another form of ”ongoing” data is is “Linked Open Data”, which forever changes, but of which we use a snapshot when we populate simulation models with data.

File formats

In planning a research project, it is important to consider which file formats will be used to store the data. In some cases, this will be dictated by the software that is used or the conventions within the scientific discipline, but in other cases a choice between several options needs to be made. These are likely to be some of the key factors in decision-making:

- what software and formats have been used in past projects.
- any discipline-specific norms (and any peer support that comes with them).
- what software is compatible with hardware already available.
- whether funding is available for new software for the job.
- how the researcher plans to analyse, sort, or store the data.

However, the researcher should also consider:

- what formats will be easiest to share with colleagues for future projects.
- what formats are at risk of obsolescence, because of new versions or their dependence on particular software and/or hardware
- what formats will be possible to open and read in the future.
- what formats will be easiest to annotate with metadata so that others can interpret them days, months, or years in the future.

In some cases, the researcher may be best off using one format for data collection and analysis and converting the data to a standard format for archiving once the project is complete. After conversions, data should be checked for errors or changes that may be caused by the export process. Given the high variety in research types within ESRIG, considerations and decisions based on the points above can be structured and streamlined on the Research Group level.

Version control

Because digital research data can so easily be copied, over-written or changed, researchers need to take steps to protect its authenticity. Research time is wasted and valuable data put at risk if researchers work with outdated versions of files.

Version control can prevent this. Control is particularly important if data is being used by multiple members of a research team, or if research files are shared across different locations.

A regime to synchronize different copies or versions of files will improve research efficiency and help guarantee the authenticity of the data. Good practice generally involves the keeping of a single master file, to which all changes are recorded. Version control mechanisms should be established and documented before any data is collected or generated. Options to realize version control will be investigated together with CIT. This might include a joint git repository for students, which allows supervisors to trace revisions.

B. Data Storage and Back-up

Data documentation needs to be done in such a way that secondary users will be able to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed. Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data.

Note that recent developments might prove useful as well, which will be explored. For example interesting work is going on with the concept of "ReproducibleResearch" (https://www.coursera.org/course/repdata). In particular, tools such as R Markdown (http://rmarkdown.rstudio.com/), the iPython Notebook (http://ipython.org/notebook.html) and more generally Jupyter (http://jupyter.org/) allows the user to render html files that show which blocks of code were used to create which graphs.

Set up of the data archive

Most of the research done in ESRIG will result at some point in scientific publications, PhD theses and MSc Theses. Once completed, data storage will be organized per publication or thesis in a single .zip archive and should include:

- the final version of the document. For PhD theses the archive should contain the separate publications/chapters as well as the final version of the completed dissertation.
- all primary (raw) and intermediate and final (processed) data underlying the document. The main criterion is that the final work should be reproducible with these data, but also primary data and results that did not make it to the final publication (because the analysis failed for example), but were important for the work, should be included.
- All program code, input data and input parameters and scripts used to produce the final results such as figures, tables, statistical analyses etc.
- relevant metadata: a text file describing the data sources in relation to (corresponding sections of) the document. N.B. Of course information presented in the publication itself (e.g. on data analysis) need not be duplicated.

The researcher needs to make sure that all file types can be read by packages generally available, or at least in possession by the research group for general and intended long(er) term use

The name of the .zip file should contain prefix_<last name>_<year>.zip, e.g.:

- *For MSc & PhD theses*: PhD_thesis_Name_2013.zip or MSc_thesis_Name_2013.zip, If this code type proves not sufficiently unique additional codes will be considered (report code, name of supervisor).
- *For journal papers*: <author(s)>_<journal>_<year>.zip, e.g. Author_PNAS_2008.zip or Author_etal_RCM_2009.zip,

Note: Underscores, not spaces in the file names. Also, each folder inside the ZIP archive should contain a read_me.txt file with info on the folder's contents.

Identifiers

An identifier is a reference number or name for a data object and forms a key part of your documentation and metadata. To be useful over the long-term, identifiers need to be unique (globally unique if possible) and persistent (the identifier should not change over time). The emerging identifier standard for publicly available datasets is the Digital Object Identifiers (DOIs). Although DOIs have been traditionally used for journal articles, they can now be assigned to datasets.

Back up policy.

It is the responsibility of the researcher to ensure that their research data is regularly backed-up and stored securely for the life of the project. It is good practice to store only what is needed to keep and keep at least three copies of crucial data. It is recommended that data is stored on the University's networked fileservers and copies kept on remote storage and /or portable storage. Generally there are four options for data storage:

Networked drives: University fileserver – As these are secure and backed-up regularly, they are ideal for master copies of research data.

Local drives: PCs and Laptops – Data can be lost because local drives can fail, or the computer may be lost or stolen. These are convenient for short-term storage and data processing but should not be relied upon for storing master copies, unless backed-up regularly.

Remote or Cloud storage – commonly used services, such as Dropbox and Google Drive, will not be appropriate for sensitive data, and their service level agreements should be studied before using them to store research data.

External portable storage devices – External hard drives, USB drives, DVDs and CDs. These are very convenient, being cheap and portable, but not recommended for long-term storage as their longevity is uncertain and they can be easily damaged.

The researcher may choose to only back up certain data, or to back up files you use every day more regularly than others. The basic rule of thumb is: the more important the data and the more often they change, the more regularly they need to be backed up.

If files take up a large amount of space and backing up all of them (or backing them up sufficiently frequently) would be difficult or expensive, the researcher may want to focus on backing up specific key information, programs, algorithms, or documentations that would be needed in order to re-create the data in case of data loss.

How to deposit the data

In the first phase, the ZIP file described above will be copied from for example the Version Control system (git repository) to a designated part of the Y-drive: for each base unit of ESRIG there will be a separate folder, with read/write rights only for scientific members of the specific base unit. The people mentioned above (main author, supervisors) are responsible for depositing the file.

When to deposit the data to the Y-drive.

For regular publications in peer-reviewed scientific journals and book chapters: the main author of the publication (usually the first author) should compile a documented archive as described above and make this available for storage within 3 months after the paper appears online.

For data collected in the context of an MSc study: all data should be deposited with the daily supervisor no later than the date of handing in the final version of the MSc report. A grade will only be awarded for the project when all data (including documentation) are deposited with the daily supervisor of the project . The daily supervisor should make the data available for storage before the final grade is 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 reading committee. 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. (This holds also the other way around: publications that have already undergone storage as part of a PhD thesis should also again be stored). The promotor should make the data available for storage no longer than two months after the approval of the thesis by the reading committee.

For data from continuous observations: for these data there is no typical point in time for archiving. Of course, when these data are subject of a publication or thesis, the above rules apply. In addition, the primary data must be stored in an appropriate database system. Data analysis procedures must be documented and stored as well, by the principal investigator responsible for the data, or by co-workers mandated by her/him. The database must also contain the final data (corrected, calibrated etc). Once a part of the data have been published or transferred to international data collections, those should be marked in the database and kept in "frozen" condition. In case for instance a recalibration requires these data to change, the prior "final" data should be kept as well.

C. Data Access and ownership

During the research project research data need to be kept safe and secure. The responsible researcher wants to determine who has access to the data and what they are authorised to do with it. Data security is needed to prevent unauthorised access or disclosure and changes to or destruction of data. The principle investigators are responsible for ensuring data security. The level of security required depends upon the nature of the data – personal or sensitive data need higher levels of security.

It is possible that remote access to the data is needed, if work is done from more than one location, or not at the university. A number of individuals may require access to the data, possibly with different privileges to read, write, update or delete. This may be accomplished by keeping a copy of the data on the university shared network file store, where it is password protected. The use of cloud storage to share data depends upon the level of security needed.

It is possible that a project may need to arrange for access to third party data that may have specific limitations in how they can be distributed (based on IP or the agreement by which the project obtained the data). When a research project has received data under confidentiality or other restrictions, these restrictions will have to be identified and explained in the data management plan.

Ownership of research data must be clarified prior to, or at the beginning of a project. Future storage and reuse are directly affected by the intellectual property rights of research data. Ownership of the data and copyrighted datasets will depend on whether the project was created as part of sponsored research; the employment status of the creator; whether third-party data has been utilised during the conduct of research, and, in case of an ‘encoded work’ whether substantial university resources were used in the creation of the encoded work.

The conditions under which the data may be made available by the data repository to other researchers are determined by the Principal Investigator depositing the data.

D. Data Sharing and Reuse

By publishing data, research will be made available to the scholarly community, who can study and build upon the work. At the end of a research project, the funding agency may require the investigator to share his/her research data, by publishing it with no access restrictions (open access). Some journal publishers also require the data supporting the research article to be published. In some research fields, international agencies collect and scrutinize data, and make them available in open databases. When disseminating the data, it needs to be considered who would be interested in the research findings, and the way how to reach this audience (by newsletter, community website, press release, attending seminars or conferences, etc.)

What also needs to be considered is how others will reuse the data. If the data are to be used as widely as possible, the Creative Commons Attribution Only licence (CC-BY), would be most useful. This license lets others distribute, remix, tweak, and build upon the work, even commercially, as long as they credit the investigator for the original creation.

Depositors can elect to apply an embargo to the research data so that public access is deferred for a specific period (typically no more than two years). Embargo may be appropriate in cases where the researcher needs to maintain the data in a managed repository environment while deferring any access to the data pending further data collection, analysis, publication of results, or if the data are subject to a patent application process. If data is generated using specifically-developed software, it may be necessary to provide a copy of the software, noting operating requirements, with the data.

E. Data Preservation and Archiving

It needs to be decided what data to preserve and what data to dispose of after the end of the project. The researcher may have his own view on how long the data need to be retained. This will be influenced by the discipline, the type of data created and whether further work or publications will be based on it. When uncertainties arise about what data might need to be held, advice from supervisors is needed. Data selected for long-term preservation will normally be submitted to a funder established data centre, disciplinary data repository or an institutional data repository.

Most funders regard costs for archiving the data or preparing it for archive as allowable as long as they are justified and incurred within the life of the project.

4. TIMELINE OF RDMP IMPLEMENTATION.

So far data storage within ESRIG highly varies between units and is often not structured and monitored. The RDMP protocol describes a mostly new procedure and time is needed for implementation. It is decided not to start with this protocol retroactively. The following implementation timeline is proposed:

- July/August 2015: all research units organize sufficient storage space on the Y-drive to which at least the chair of the group and the director of the institute will have access

- July/August 2015: ESRIG data management committee is installed. Committee members include Harro Meijer (ESRIG director), Anita Buma (Ocean Ecosystems), Gerard Dijkema

(IVEM) and –as soon as she/he is appointed- the scientific coordinator.

- September 2015: new PhD’s will add a project-RDMP to the TSP, other new researchers and Master students will describe a personal RDMP and store it in a dedicated folder on the Y-drive of the research unit.

- September 2015: possibility for Open repository (Version Control System) explored (pending appointment of scientific coordinator).

- Early 2016: the ESRIG data management committee will do a first evaluation, which will be repeated annually. During the first years, the Questionnaire DM will be evaluated as well, and updated if required.