# Good enough practices in data management and how to translate these in data management plans (DMPs)

Tanja Milotić, Peter Desmet and Stijn Van Hoey (LifeWatch, INBO)

## Table of contents

# PART I: Data management in biodiversity research

# Introduction

## About this document

This document provides some generic guidelines for the preparation of a data management plan (DMP) and gives guidance for research data management (RDM). These guidelines have been drafted as an output of the SAFRED project. The goal of the BRAIN project SAFRED (Saving Freshwater Biodiversity Research Data) was to achieve systematic recovery and publication of data generated in freshwater research. As many of the problems of retrieving old research data could have been avoided by good data management, these guidelines for research data management and DMP writing have been written in order to avoid similar problems in future projects and to stimulate project partners to improve their day-to-day data management practices. As the purpose of this project was to recover freshwater biodiversity data, examples are given for biodiversity data, but most of the guidelines are easily applicable for other fields of research as well.

The structure of the document is in line with the data life cycle. In case your funder or institution does not require the use of a specific DMP template, this 'generic' structure can be used in the DMP. For each section, we provide general guidelines and tips (*guidance*) which are complemented by specific *examples* for biological research where relevant.

## The research life cycle vs. the data life cycle

The workflow for conducting scientific research is comparable for most fields of study and can be summarized in the ***research life cycle*** (figure 1). Generally, in a first step, research starts from (new) **ideas and hypotheses**. In case those ideas are pursued, methods and techniques are selected or developed to **collect data** for testing the hypotheses. In a next step, the data is analyzed (e.g. statistical tests, modeling,...) and the results and data are visualized and interpreted. In a final step, the findings of the study are **published** as a scientific paper, report, thesis, etc. and can feed new ideas and hypotheses for new studies.

As the number of publications and their impact factors are commonly used to evaluate research careers or to make funding decisions, scientific publications are often regarded as the main research output. Although good research requires good data and a great amount of effort and money is invested in data collection, datasets are often considered as a by-product of scientific research that lose their value after publishing results. Recently, Vines et al. (2013) studied the

availability of research data from 516 ecological studies between 2 and 22 years old and found out that the odds of a dataset being reported as retrievable fell by 17% per year. The reduced availability of raw research data was often caused by the inability to contact the authors by email and broken or lost storage devices. Some of the unavailable datasets could be retrieved after considerable efforts of the authors, while other datasets are completely lost for science. Many datasets are unique due to the specific timing and location of research, and the combination of multiple local datasets could lead to a better understanding of larger-scale processes and changes (Wolkovich et al. 2012). Examples in biological sciences are observational data (such as species distributions at a certain location and moment, phenological data, meteorological data, satellite images, etc) and experimental data that cannot be replicated easily or at a moderate cost.

Recently, many publishers and funders have begun to realize the intrinsic value of research data for the scientific community and require proper data management and/or the publication of data. For example, *Nature* requires authors to make data "promptly available to readers without undue qualifications" and to disclose restrictions upon submission, and encourages researchers to deposit data in community-endorsed public repositories (Gibney and Van Noorden 2013). Funders, such as the EU-funded Horizon 2020 programme, increasingly recognize the value of research data and engages project proposers to plan data management even before the project is granted.

Similar to the workflow for conducting good scientific research, the different steps in research data management can be summarized in the ***data life cycle*** (figure 1). The first step in this cycle is the **collection of data** which may involve the collection of *new data* (through experiments, measurements, observations or simulations) and/or the reuse of *existing dataset*s. Next, **data quality** is tested and if needed corrected. Although in each step, it is wise to document the data collection process and corrections in the data quality check, it is also needed to **document** the dataset with metadata. As in the research life cycle, data and research results are **used** in publications and presentations. After finishing the research, or after a certain time period in long-term research projects, steps need to be made towards the long-term **preservation** of data. In the next and final step, data are published on general or discipline-specific platforms. As data become available and usable for the scientific community, the published data can lead to new ideas and hypotheses or can be applied in new research.

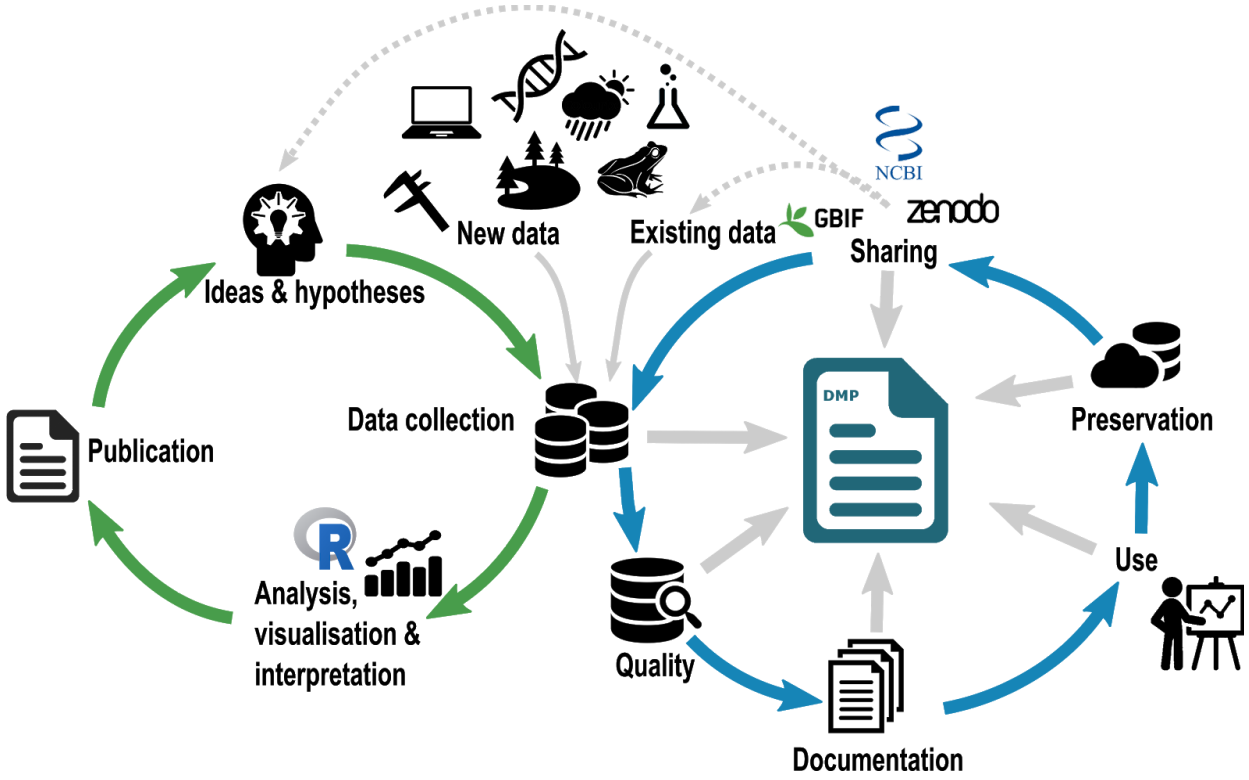



Figure 1 - The research life cycle (green cycle, left) vs. the data life cycle (blue cycle, right). Based on Michener 2015.

## Data management in practice: a survey of RDM practices among the SAFRED partners

During the SAFRED project, the current data management practices among the SAFRED partners were surveyed. The purpose of this survey was to document the used and known techniques and to identify weaknesses and missing steps. Two surveys were made aimed at either individual researchers to document day-to-day data management practices and research lab PI's to document the existence and use of an institutional or lab research data management policy. The questions asked in both surveys are available at the end of this section.

## Data management policy

Researchers pointed out the absence of a data management policy or data management guidelines in their labs. Research lab PI's often are aware of the existence of an institutional data management policy, although this is poorly implemented in the individual labs (figure 2).

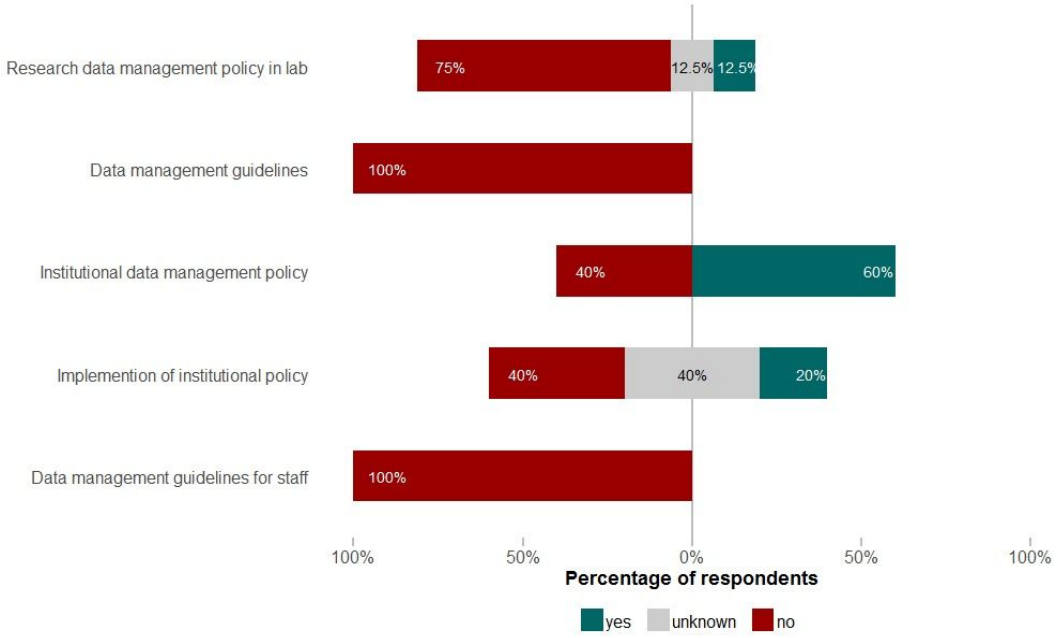



Figure 2: Existence and implementation of data management policies and guidelines in the surveyed research labs (survey questions R1, R2, L1, L2 and L3 respectively).

## Data types and file types

The diversity in research topics in the labs taking part in the SAFRED project leads to various data and file types (figure 3). Most of the file types are fit for storing tabular data, and some specific file types for genetic research (fasta, fastq) are commonly used.

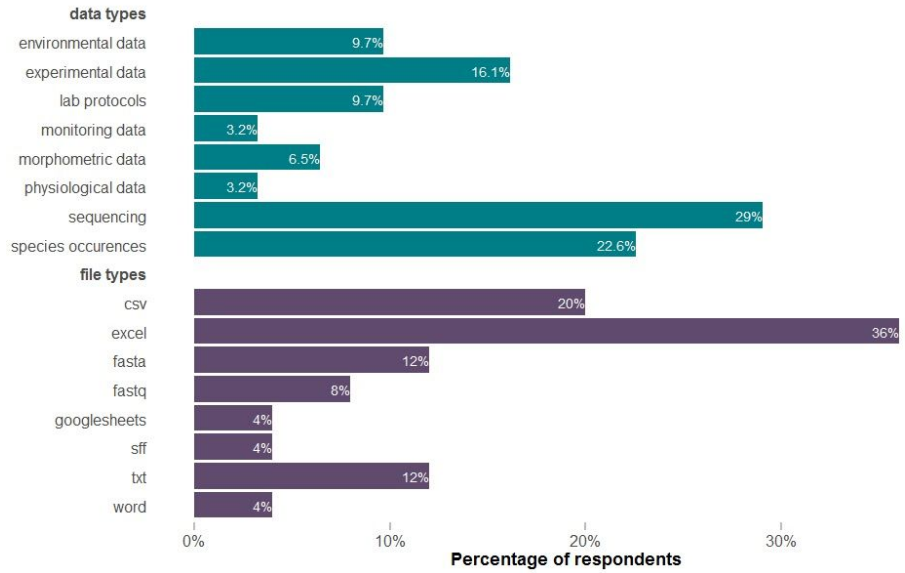



Figure 3: Commonly used data types and file types (derived from the answers on questions R3 and L5, and R4 and L6 for respectively data types and file types).

## Data organisation

Individual researchers use specific systems to structure files in folders (figure 4). However, clear naming rules, the use of version control for files (either by naming files or by using specific software) and the creation of metadata files describing the content of data folders and files is less commonly applied. Furthermore, the retrieval of old, important data files is impossible according to 25% of the interviewed researchers.

Lab managers are far less optimistic about the data management in their research lab (figure 5). No clear rules exist for naming files, structuring files in folders, writing metadata and the use of a version control system. This seems to reflect in the recovery rate of datasets used in publications as only 20% of the interviewed lab managers considers it possible to retrieve these data files if needed.

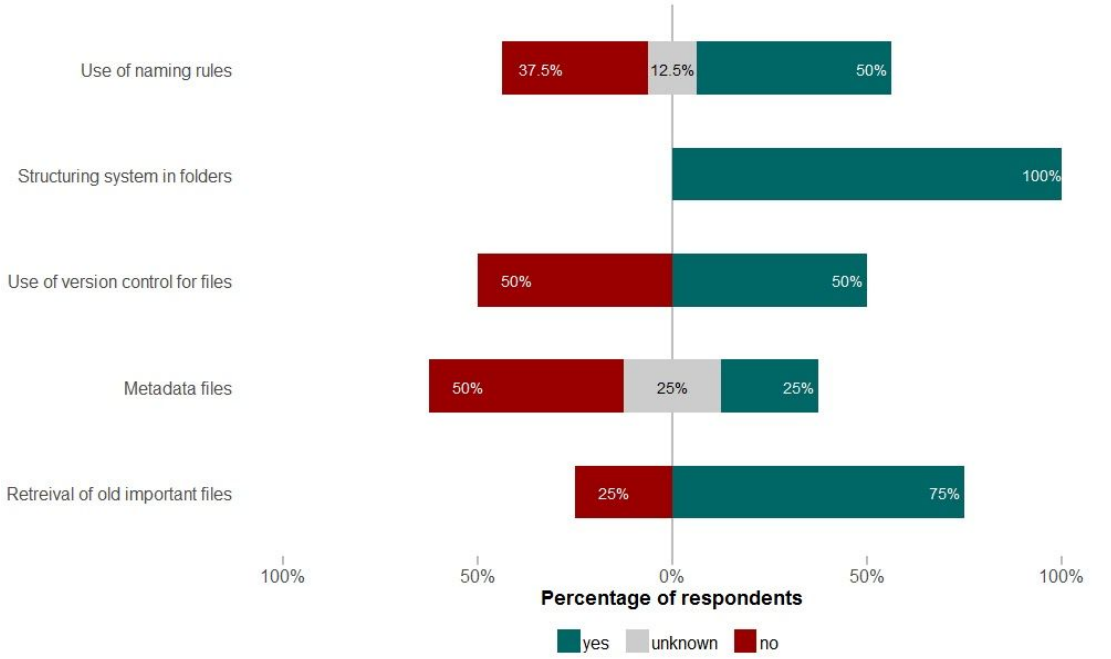



Figure 4: The use of data organisation practices by researchers (questions R9, R10, R11, R12 and R13).

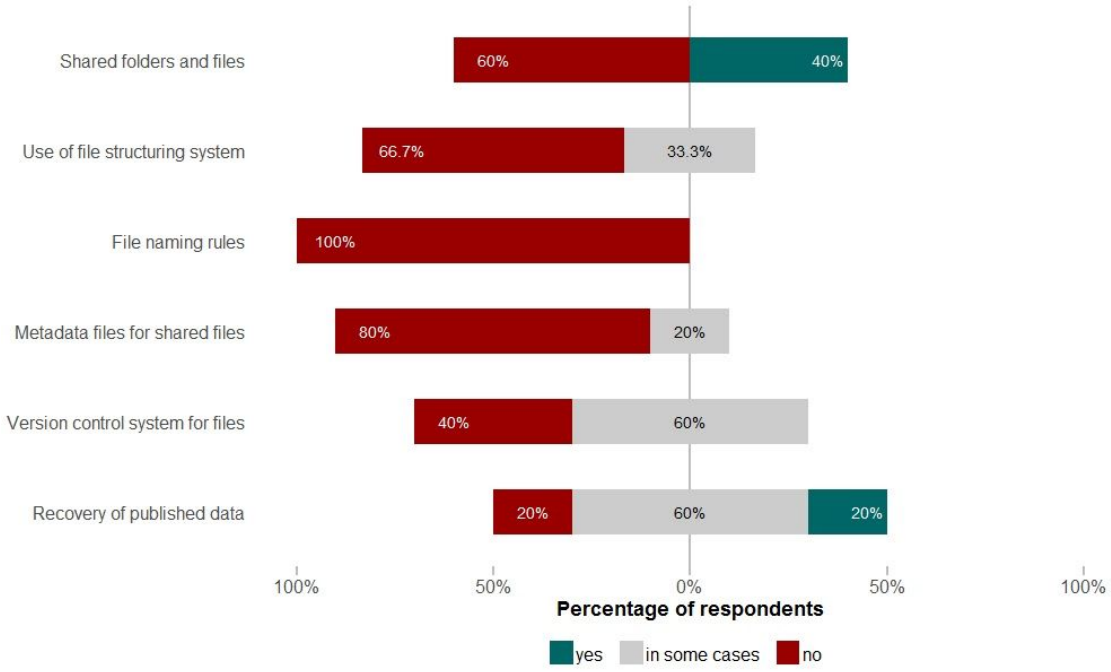



Figure 5: Data organisation in research labs (questions L9 - L14).

## Storage and back-up

Both researchers and lab responsibles are aware of the need to back-up data (figure 6). However, back-up guidelines are lacking and the back-up frequency is highly variable (figure 7).

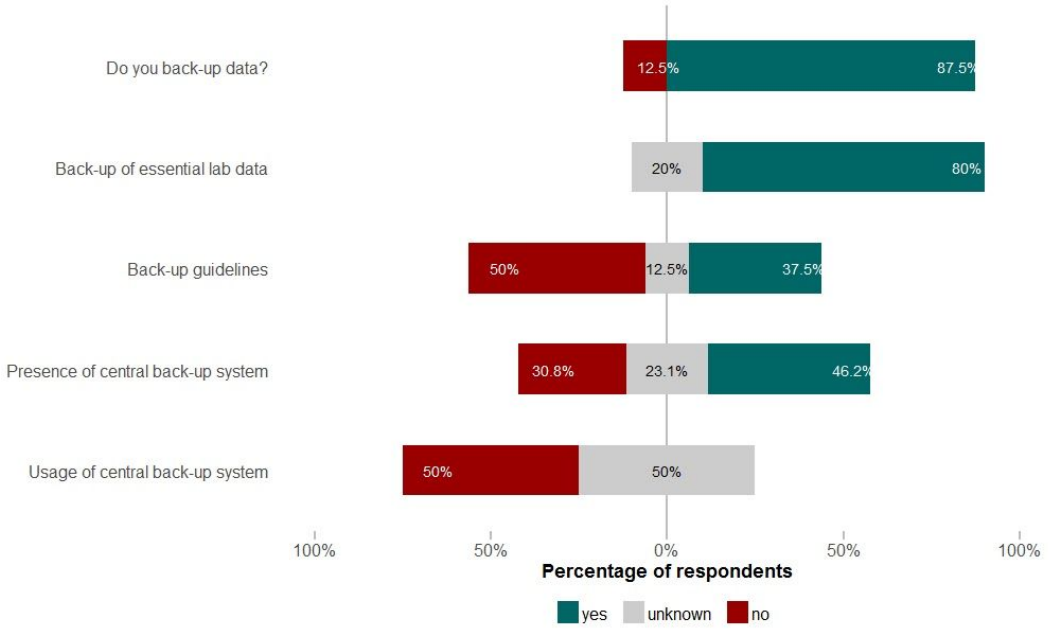



Figure 6: Storage and back-up strategy at the researcher and lab level (questions R15, L16, R18, R19, L19).

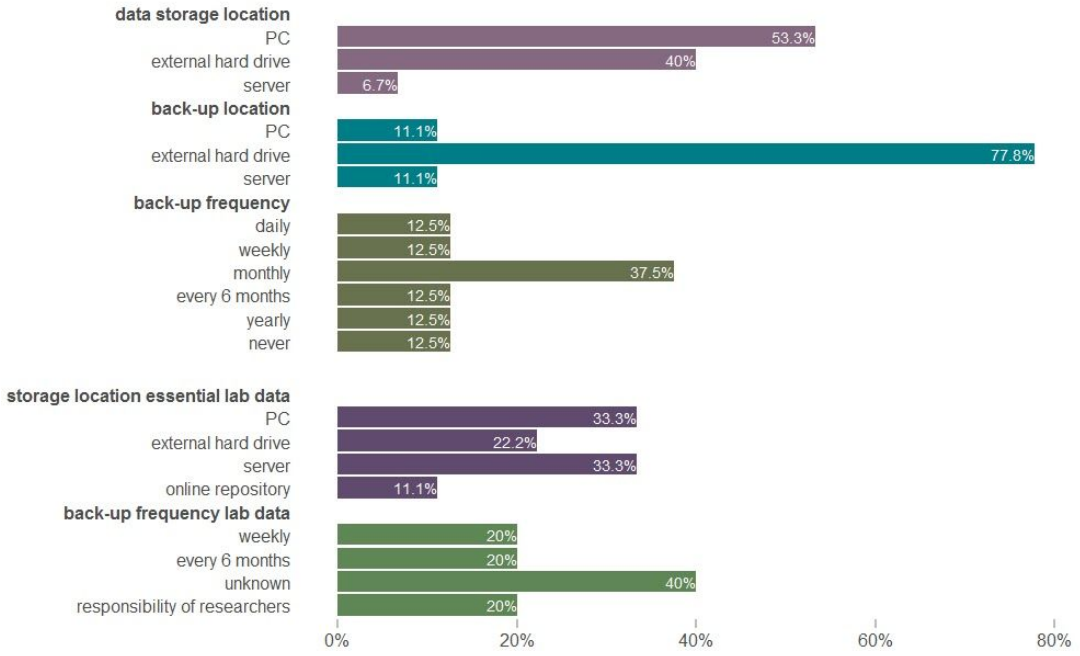



Figure 7: Storage media, back-up media and back-up frequency used by researchers (questions R14, R16, R17) and in research labs (questions L15, L18).

## Sharing and publishing

Not too surprisingly, the researchers associated with the SAFRED project are highly engaged in collaboration with peers with analyses of datasets collected by other researchers, and sharing and publishing their own datasets (figure 8).

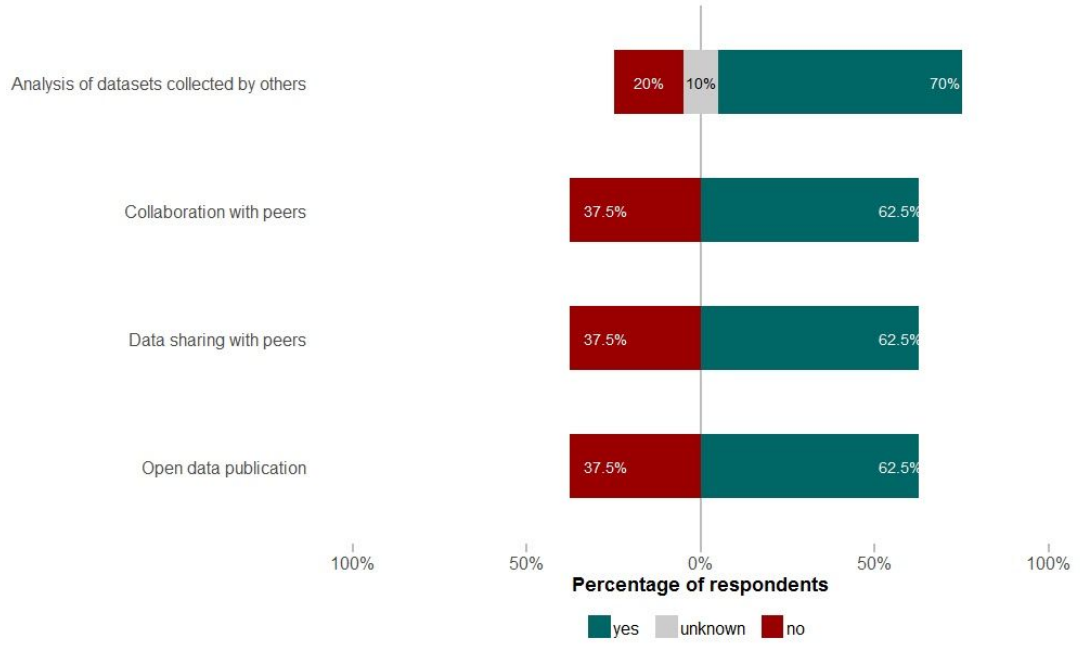



Figure 8: Sharing and publishing of datasets by researchers in the SAFRED project (questions R20-24).

## Data management plans

Few of the respondents have been involved in writing data management plans (DMP) yet, but most researchers are convinced that this will become part of the application process for projects in the future. The need for DMP support is high according to all respondents (figure 9).

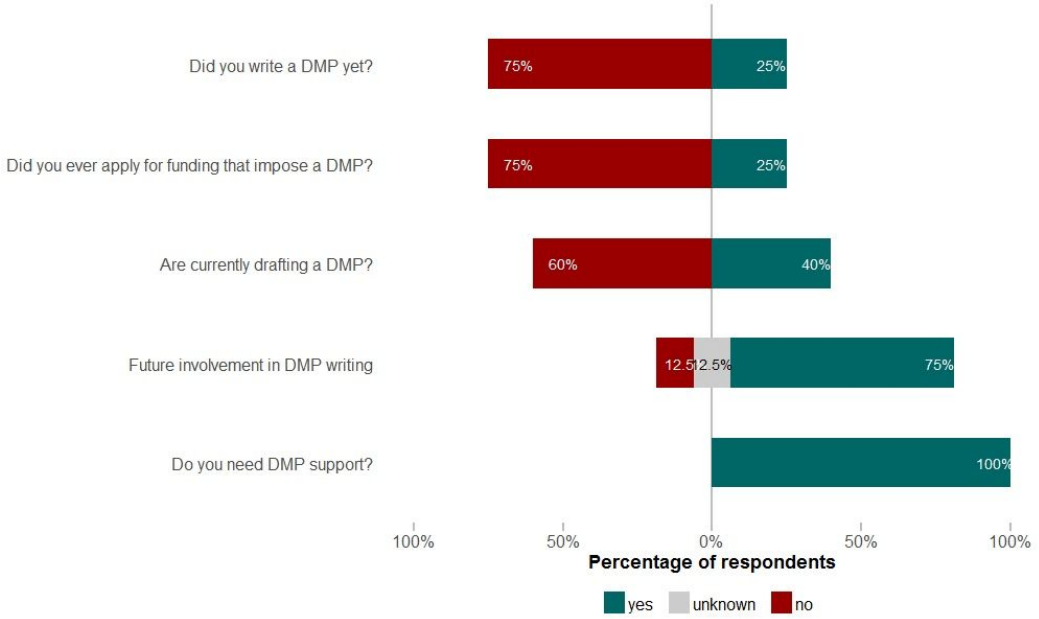



Figure 9: Experience in data management plan (DMP) writing and future expectations (questions R27+L24, R28+L25, L28, R29+L26, L28).

# Survey questions

## Questions asked in the survey for researchers

**General questions**
R1. Are you aware of a research data management policy at your university or department? (Y/N/don't know)
R2. Did you receive any guidelines or procedure regarding data management at the moment you joined the lab? (Y/N/don't know)

**Data collection**
R3. Which data do you generate for your work? Please, briefly describe the nature of these data, qualitative or quantitative, and the data types. Possible data types are: raw measurement data, processed data, lab protocols, calibration data, reference data (t0 measurements), measurement series (e.g., a yearly monitoring of value x on place y), experimental data, sequencing results,... (open ended question)
R4. In which format are these datasets? (e.g., MS Access, Excel, csv, txt, etc) (open ended question)
R5. What is the volume of your most important datasets? How many files? How many records are there (approximately) in a dataset? (open ended question)
R6. Did you analyze data produced by others? (Y/N/don't know)
R7. Do you use any large datasets which are shared with multiple collaborators (for input or use of data)? In which format and where are these stored? (open ended question)
R8. What products (e.g., research papers, reports, books,...) were or will be produced with these data? (open ended question)

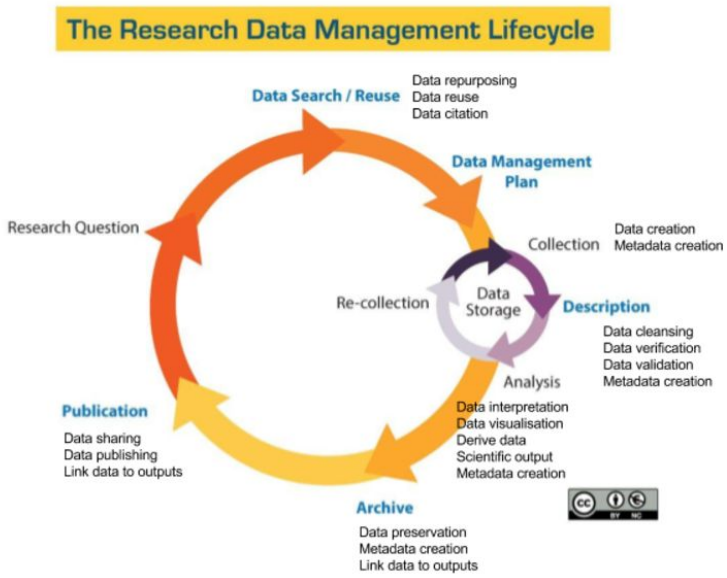
**Organising data**
R9. Do you use any naming rules for research files on your computer? Please briefly explain how. (open ended question)

R10. How do you structure files in folders? Per (sub)project? Is the method for classifying files in folders consistent among projects (e.g., separate folders for data, scripts, analysis results, etc)? (open ended question)
R11. Do you make metadata files describing datasets? (Y/N/don't know)
R12. Do you use a system for file versioning? How do you cope with different versions of a document/data file? (open ended question)
R13. If someone would ask you if you could send data used in a random publication you authored in 2012, would you be able to send the requested document (regardless of the fact you'd be willing to do this)? How much time would this cost you? (open ended question)

**Storage and back-up**
R14. Where do you store your research data? (open ended question)
R15. Do you make backups of your data? (Y/N/don't know)
R16. Where are backups stored? (open ended question)
R17. How often are backups made? (open ended question)
R18. Are you aware of any guidelines on backing-up data at your organization? (Y/N/don't know)
R19. Does your organization have a centrally managed backup system? (Y/N/don't know)

**Sharing and publishing data**
R20. Does your research require data sharing (e.g., between collaborators in a project)? (Y/N/don't know)
R21. If you answered yes to the previous question, who is your prime audience? Who do you share your data with? (open ended question)
R22. How do you share data with others? (open ended question)
R23. Are there any conditions that you consider prerequisite to sharing your data? (open ended question)
R24. Have you ever ever uploaded data as open data in a data repository (e.g., GBIF, zenodo, dryad?) (Y/N/don't know)
R24.1. If you answered yes, why? Required by a funding agency, consortium partner, publisher, on your own behalf? (open ended question)
R24.2. If you answered no, why not? (open ended question)

**Data management plans**
R25. Below, a picture is shown with a data management life cycle example. Please describe the parts of the data management life cycle that are a focus for you. (open ended question)




R26. Please describe how the data management life cycle for your project differs from this one, if at all? (open ended question)

R27. Did you write a data management plan (DMP) yet? (Y/N)
R28. Did you ever applied for a type of funding (such as Horizon 2020) that requires the preparation of a data management plan? (Y/N/don't know)
R29. Do you expect that you will have to write a data management plan in the future? (Y/N/don't know)

**Data management support**
R30. Do you have access to any technical assistance on research data management in your organization? (Y/N/don't know)
R31. Do you have specific questions or problems regarding data management which you would want to see covered in a general document with guidelines on research data management? (open ended question)
R32. Before we end, do you have any last comments that you would like to share? (open ended question)

## Questions asked in the survey for research lab PI's

**General questions**
L1. At your university or department, is there a research data management policy? (Y/N)
L2. If a research data management policy does exist, is it implemented in your department/research group? (Y/N)
L3. Are there any guidelines or procedure regarding data management when a new collaborator joins the lab? (Y/N)
L4. What happens with research data when a collaborator leaves the department? Is there a difference between finished and unfinished projects? (open ended question)

**Data collection**
L5. What would you consider as essential data within your department? Please, briefly describe the nature of these data, qualitative or quantitative, and the data types. Possible data types are: raw measurement data, processed data, lab protocols, calibration data, reference data (t0 measurements), measurement series (e.g., a yearly monitoring of value x on place y), experimental data, sequencing results,... (open ended question)
L6. What are the most common formats in which essential data are stored at the research group? (e.g., MS Access, Excel, csv, txt, etc) (open ended question)
L7. What is the volume of the most important datasets? How many records are there (approximately) in the essential datasets? (open ended question)
L8. Are there any large datasets with multiple users (for input or use of data)? In which format and where are these stored? (open ended question)

**Organising data**
L9. Are there any shared folders with collaborators in your research group? (Y/N)
L10. If you answered yes in the previous question, how do you structure files in folders? Per (sub)project? Is the method for classifying files in folders consistent among projects (e.g., separate folders for data, scripts, analysis results, etc)? (open ended question)
L11. Does your research group provide recommendations for consistent naming of data files? (open ended question)
L12. Does your research group imposes collaborators to write metadata files describing datasets? (open ended question)
L13. Does your research group use a system for file versioning? How do you cope with different versions of a document/data file? (open ended question)
L14. If someone would ask you if you could send data used in a random publication authored by a member of your research group in 2012, would you be able to send the requested document (regardless of the fact you'd be willing to do this)? How much time would this cost you? (open ended question)

**Storage and back-up**

L15. Where are the essential datasets of the research group stored? (open ended question)
L16. Are these data backed up? (Y/N/don't know)
L17. Where are backups stored? (open ended question)
L18. How often are backups made? (open ended question)
L19. Are there any guidelines on backing-up data? Does your organization have a centrally managed backup system? (open ended question)

**Sharing and publishing data**

L20. Are there any projects in your research group that require data sharing (e.g., between collaborators in a project)? (Y/N)
L21. If you answered yes to the previous question, how do you share data with collaborators in a project? How is this done in your own research group or between other institutes/universities or departments? (open ended question)
L22. Would your research group share its data as open data? Under which conditions? (open ended question)
L23. Has anyone in your research group ever ever uploaded data as open data in a data repository (e.g., GBIF, zenodo, dryad?) (Y/N/don't know)
L23.1. If you answered yes, why? Required by a funding agency, consortium partner, publisher, on your own behalf? (open ended question)
L23.2. If you answered no, why not? (open ended question)

**Data management plans**

L24. In your research group, are there any data management plans made yet? (Y/N/don't know)
L25. Has anyone in your research group ever applied for a type of funding that requires the preparation of a data management plan? (Y/N/don't know)
L26. Do you expect that you or a member of your research group will have to write a data management plan in the future? (Y/N/don't know)
L27. Do you have any technical assistance on research data management in your organization? (open ended question)
L28. Are you or a member of your research group currently working on a data management plan? Would you like a review of this plan? (open ended question)

**Data management support**

L29. Would you like a further survey on data management practices by a visit at your lab? (open ended question)
L30. Do you have specific questions or problems regarding data management which you would want to see covered in a general document with guidelines on research data management? (open ended question)
L31. Before we end, do you have any last comments that you would like to share? (open ended question)

# PART II: Writing a data management plan (DMP)

## Data management plans

Funders are increasingly recognizing the value of research data and often require to submit a data management plan with the research proposal. A ***data management plan*** (DMP) is a document that describes how data will be treated during a project and what happens with the data after the project ends. Such plans typically cover all or portions of the data life cycle: from data discovery, collection, and organization (e.g., spreadsheets, databases), through quality assurance and quality control, documentation (e.g., data types, laboratory methods) and use of the data, to data preservation and sharing with others (e.g., data policies and dissemination approaches) (Michener 2015). European funded programs (e.g., Horizon 2020, ERC) and some Belgian funders (e.g., FWO) have a specific template for drafting DMPs, but ideally a (basic) DMP is made when starting any new projects or research activities. Making a DMP in an early phase is beneficial to researchers and supervisors on the longer run as it provides an opportunity to compile an overview of the tasks in data management, assigns responsibilities for each task, and make a realistic estimate of the costs associated with data management.

### Why?

Although it requires extra efforts to make a data management plan, there are many good reasons to do so, including:

- Good practice: even though you have to spend valuable time at the start of a project, it is probable to pay off in increased efficiency in later steps in the project. In fact, you could compare making a DMP with making a lab protocol when doing experiments or measurements. In both cases, a plan should be made in a preparatory phase, while afterwards, when in later phases in the project it might be necessary to change the initial plan with new insights or solutions for unforeseen problems.
- It is mandatory for many funders (e.g., H2020).
- A DMP could serve as a guide book in case a collaborator suddenly leaves.
- To prevent the creation of a jungle of data: if in case people start thinking about the data they are about to collect, it is more probable that data will be collected and stored in a more structured way which improves the exploration and use of data in a later phase.
- To minimise the risks of data loss.
- To create a higher value for data by enabling their reuse (e.g., in meta-analyses) and to increase the reproducibility of research.
- To increase the verifiability of published research and to prevent accusations of bad science or fraud.
- Data often have a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding has ceased, follow-up projects

may analyse or add to the data, and data may be reused by other researchers. In order to facilitate reuse of data, datasets need to be well organized and well documented.

## When should you make a DMP?

- A first version of a DMP is ideally made before starting the collection of data in a project (see figure 1).
- Whenever major changes come up in the project scopes, used methods, type of collected data, etc. the DMP should be updated.
- A DMP It is also a useful tool at the start of a PhD as this encourages the student to overthink the whole process from data collection to publication and the next steps of data preservation and reuse.

## Writing a data management plan

In the review of Michener (2015), 10 simple rules for creating useful data management plans were listed (see box 1). In this section, we give guidance how to comply with these rules.

Box 1 - Overview of some recommended sections in a data management plan (Michener 2015).

**10 simple rules for creating a good data management plan**

1. Determine the research **sponsor requirements**
2. **Identify the data** to be collected
3. Define how the data will be **organized**
4. Explain how the data will be **documented**
5. Describe how **data quality** will be assured
6. Present a sound **data storage and preservation strategy**
7. Define the project's **data policies**
8. Describe how the data will be **disseminated**
9. Assign **roles and responsibilities**
10. Prepare a realistic **budget**

# 1. Project description

## 1.1. Plan details

This section contains some general information of the plan and the project, such as:

- ***Title***: this is typically the (working) title of the project or its acronym.
- ***Name and contact details*** of the PI and the data manager.
- ***Abstract***: provide a brief summary of the objectives and nature of the project to help others to understand the purpose of the data collection.
- Version: initial version at the start of the project / reworked version during the project
- What is the planned ***DMP update cycle***? Number of obligated updates / number of updates considered necessary irrespective of funder requirements

## 1.2. Purpose

**DMP question**

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

**Guidance**

Provide a brief overview of the objectives of the project and which data you will use to fulfill them.

## 1.3. Roles and responsibilities

**DMP question**

*Who will be responsible for data management in your project?*

**Guidance**

Depending on the scope and the size of the project, one person (e.g., the lab's data manager, or the PhD-student applying for a grant) might be responsible for data management of the entire project. In larger projects, especially international projects with multiple universities or research institutes involved, it is wise to assign data management responsibilities to multiple persons. It is

crucial to assign these roles before starting the project and to provide a backup responsible in case the assigned responsible should leave the organisation.

**Examples**

- In a PhD project: the PhD student is responsible for data management during the PhD project and manages data according to the lab's research data management policy. The supervisor of the PhD student also supervises the process of data management during the project. Once the project is finished, the key datasets are archived (according to the university's data management policy) and the responsible person for long-term preservation is the data manager of the departement.
- In a larger international project involving multiple partners: depending on the project it could be interesting to assign different responsibilities based on the timeline of the project. For example:
    - A data manager coordinates the process of data management during the project, but different sub-tasks are fulfilled by different partners or persons.
    - During data collection: partner 1 is committed to collect dataset X, while partner 2 is collecting dataset Y. During the data collection phase, each partner is responsible for managing the dataset they are collecting at that time using the data management rules that are agreed upon before starting the project.
    - After finishing the collection of data, partner 3 is responsible for checking data quality.
    - Data description phase: all partners are responsible for providing metadata for the datasets they collected, manipulated or checked for inconsistencies.
    - After finishing the project: partner 4 is responsible for the long-term storage of the research output, and for the publication and dissemination of data. That partner is also the main point of contact for questions about the dataset(s).

## 1.4. Budget

**DMP question**

*What are the costs for data management in your project? How will these costs be covered? What is the expected size of the dataset(s)?*

**Guidance**

In **European** (H2020, ERC) projects, data management costs are eligible as part of the grant. Other funders may have other rules, but still it is important estimate the costs for data management at the start of the project.

For **smaller projects** (e.g., PhD projects), you should consider whether you have enough server space to store the data, or whether you need an additional budget for storage (server space, additional hard-disks,...).

Also, you should estimate the costs and potential value of long-term preservation of data. Therefore, you should specify a target data archive for long-term storage, and the associated costs for using it (if any). If you are looking for long-term storage at your institutions' servers, also find out what the costs are and for what time period the storage of data is guaranteed.

**Examples**

The estimated costs should cover:

- The costs for server space (make an estimate of the size of the dataset(s) first!).
- Employment costs of a data manager.
- Costs and potential value of the long-term preservation of data.
- Data publication costs.

## 1.5. Sponsor requirements

Data management and DMP writing is becoming one of the standard tasks when applying for funding. European funding programs (ERC, Horizon 2020,...) and regional funders (FWO,...) have their own DMP templates that need to be filled out when applying (or shortly after the allocation of funding). These templates can be found at the funders websites, and a collection of funder templates are also included in the online DMP-tool dmponline.be, established for Belgian researchers. When you are working with a team on a DMP, this online DMP tool is recommended as it allows you to work in one document at the same time and provides the most recent versions of the DMP templates. Many universities and institutes[1] are already a member of the Belgian consortium, but if you're not a member you can sign in using an ORCID account. Non-members can contribute to a DMP created by a member, but are not able to create a new DMP entry.

[1] On September 21rst, 2018 DMPonline.be members were Instituut voor Natuur- en Bosonderzoek (INBO), Université Libre de Bruxelles, Universiteit Antwerpen, Universiteit Gent, Universiteit Hasselt, Vrije Universiteit Brussel, Wetenschappelijk Instituut Volksgezondheid – Institut Scientifique de Santé Publique (Sciensano), Université Catholique de Louvain, Université de Liège, Université de Mons, Université de Namur, and Vlaamse Instelling voor Technologisch Onderzoek (VITO).

# 2. Data acquisition

## 2.1. Origin of data

**DMP question**

*What is the origin of the data?*

**Guidance**

Provide a list with the datasets you will use or create in the project and make reference to their source (in case of existing data), or how you will obtain them (e.g., through measurements, modelling,...).

### 2.1.1. Reuse of existing data

**DMP question**

*Will you reuse any existing data and, if so, how?*

**Guidance**

- Search for existing datasets useful in the project.
- Check the conditions of reusing data:
    - What is the price?
    - Any restrictions on using the dataset?
    - Is it allowed to share data resulting from the project based on this third party dataset?
    - Does the data holder request co-authorship when using the data?
    - When making a data compilation: keep track of the source of the data and request permission for (re)publishing if needed.
    - If certain open datasets are vital for your work, consider to contact the provider of the data and acknowledge them accordingly.
- Cite the data sources in your work following the scientific conventions of citing sources: which data did you use, which version, download date, data publisher, and digital object identifier (DOI) of the dataset when available.

**Examples**

- In biology, GBIF.org is the standard for species occurrence data, while a broad range of datasets can be found in general-purpose open access online repositories such as zenodo.org, datadryad.org and figshare.com.

- re3data.org offers information of more than 2000 research data repositories and can be used to search for specific data repositories.
- At dataone.org, you can search for public datasets published on a wide number of repositories using keywords.
- Always check the terms of use of a dataset (see also licensing of open data).

### 2.1.2. Collection of new datasets

**DMP question**

*Which new dataset(s) will you collect? Which methods will you use to obtain them?*

**Guidance**

- Give a brief overview of the measuring techniques you will use.
- Give an overview of the geographic and taxonomic extent of the data.
- Describe the data flows for different types of (sub)datasets, e.g. field notes and measurements, separate samples for measuring chemical substances, extracting DNA, identifying organisms, etc. and how the data will be centralised at a later stage.

## 2.2. Data types

**DMP question**

*What types of data will be collected or generated during the project?*

**Guidance**

Research data is any information collected or created for the purpose of analysis to verify scientific claims. Research data can be classified in numerous ways. In this chapter, we try to give an overview of different data types based on the collection method, the degree of manipulation, and their physical nature.

### 2.2.1. Collection methods

Based on the methods and purpose of data collection, different data types can be distinguished (figure 2):

- **Observational data**: data which are tied to time and place. This data type is often irreplaceable, e.g., field observations of animals or plants, sampling data, weather station readings, satellite data, camera trap data,...

- **Experimental data**: data generated in controlled or semi-controlled environments. This type of data could be reproduced although it may be expensive to do so, e.g., field plots, greenhouse or ecotron experiments, chemical analyses, DNA sequencing,...
- **Simulation data**: data generated from models, e.g., population growth, climate, interaction, or movement modelling.
- **Derived data**: data not collected directly, but generated from (an)other data file(s), e.g., the biomass of a population which is derived from population density and average biomass of organisms.
- **Metadata**: data about data, or a set of data that describes and gives information about other data. It documents mainly the content and origin of data files.
- **Protocols**: a research protocol is a set of predefined procedures and methods for designing and implementing experiments. Protocols are often used to standardize laboratory methods to ensure successful replication of results. For similar reasons, they are also applied to document experimental research.
- **Calibration documents**: calibration typically is the setting or correction of a measuring device by adjusting it to match to a dependably known and unvarying measure. In order to interpret the results measured with a calibrated device and to replicate the measurements, it is important to keep track of the calibration methods.

### 2.2.2. Manipulation

Depending on the nature and progress of a research project, different types of data are distinguished (figure 10):

- **Raw data**: In the first step raw data are collected through observations, measurements or experiments (observational data and experimental data), e.g., species occurrences, genetic data, morphometric measurements, experiment results, environmental data, geospatial data,... The most important guideline for raw data is to **never ever delete or modify the original raw data file** (Hart et al. 2016). In case errors occur in the later processing steps, it is of utmost importance to be able to repeat the process using the original raw data file.
- **Intermediate data**: Intermediate data files are derived products from the raw data files after a (or many) processing steps. An intermediate data product aims to provide a stable markpoint in the analysis resembling the result of a defined set of data processing steps. Although it is possible to repeat the production of intermediate data through scripting using raw data as input for more complex analyses or models, it might be useful to save intermediate data files to reduce calculation time. Some particular guidelines are:
    - For each of the intermediate data products, define the importance to store the data product. The importance is directly related to the difficulty (e.g. large computational power) of underlying processing steps. Important intermediate data products fall under the same data management guidelines as raw data.

- When the data product is used as a source for any other research, application, or external process, it falls under the same data management guidelines as raw data.
- Keep process settings (running parameters) and process execution separate to enable easier management, while keeping versioning possible.
- When running many scenarios, consider using a database or hierarchical structure to link input and output and scenario settings.
- Make a data dump as a table from processed data just before it is used to make a graph or figure (for reporting, paper,...). As such, this figure data can be reused independently from the other processing steps to adapt the look and feel of the graph *(e.g. when a graph illustrates the aggregated result of 100000 simulations for which the storage of the individual simulations is not possible, the storage of the aggregated data is crucial to overcome the necessity to rerun all the simulation).*

- In the last step **processed data** are generated, such as the model results, statistical outcome, phylogenies, calculated indices, aggregated data, maps as outcome of geospatial analyses,..

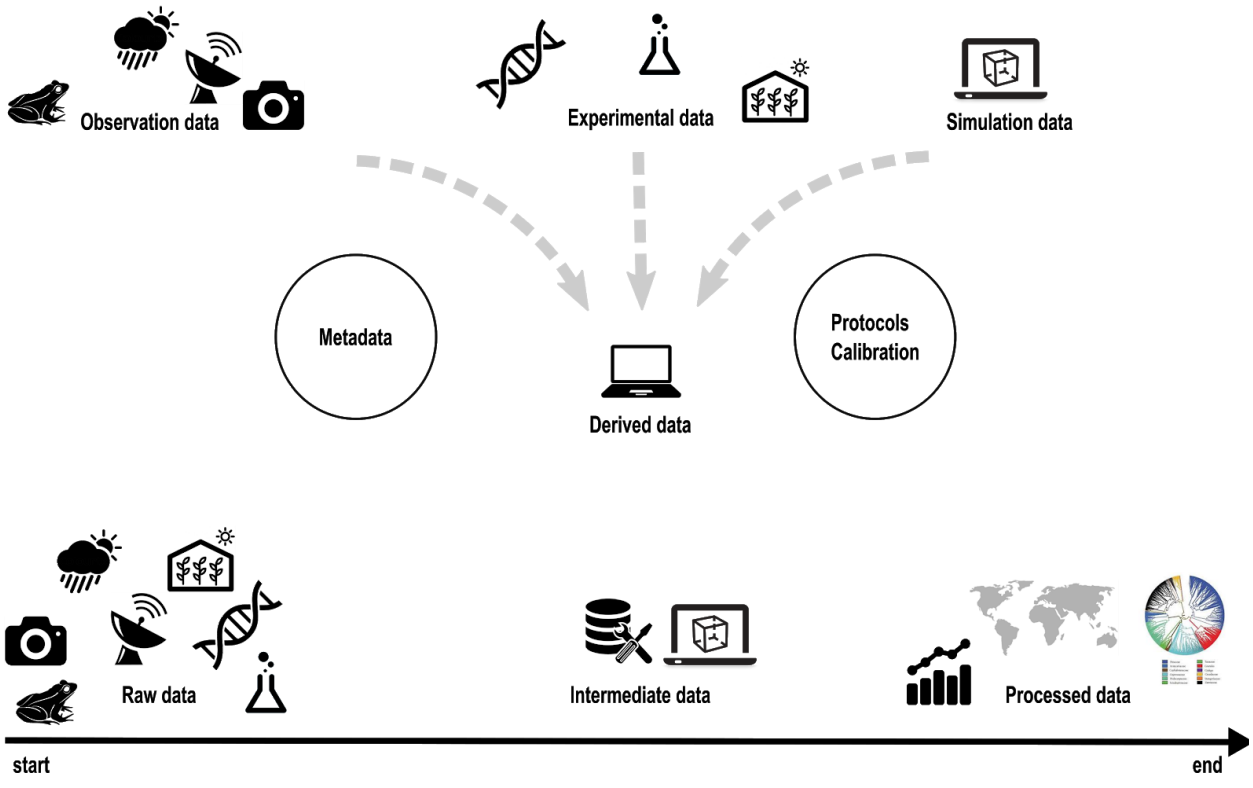



Figure 10 - Different types of research data based on collection methods (top) and degree of manipulation (bottom).

### 2.2.3. Physical nature

Based on the physical nature of data, one can distinguish two main data types:

- **Digital or digitized data**: often field observations are written down in a notebook and digitized in a later stage. In other cases, data are collected digitally (e.g., automated measurements, images, sounds,...). Often these data include observations, animal tracks, measurements, pictures, audio, models,...
- **Physical data**: in biology research often samples are taken from organisms, tissues, soils, water, etc. or collections of taxonomic groups are made. In most cases these physical data will result in digital data (e.g., the outcome of the analytical measurements on the samples, species occurrences,...), but it is important to keep in mind that these samples and collections are considered as data as well. Depending on the nature of research, it is important to keep a reference collection (or even the entire collection) for future reference.

## 2.3. Data organization

### 2.3.1 File formats

**DMP question**

*What file formats will the project generate?*

**Guidance**[2]

File formats that are non-proprietary (e.g. open source, de facto standards), and/or in widespread use are recommended (table 1). When it is necessary to save files in a proprietary format, include a README.txt file in your directory that documents the name and version of the software used to generate the file, as well as the company who made the software.

Table 1 - Some preferred file formats.

| Type | Format |
|---|---|
| Containers | TAR, GZIP, ZIP |
| Databases | XML, CSV |

[2] https://github.com/cengel/data-management/blob/master/datamanagement.Rmd

| | |
|---|---|
| Geospatial | geojson, geoTIFF, netCDF |
| Moving images | MOV, MPEG, AVI, MXF |
| Sounds | FLAC, WAVE, AIFF, MP3, MXF |
| Statistics | ASCII, CSV |
| Still images | TIFF, JPEG 2000, PDF, PNG, GIF, BMP |
| Tabular data | CSV, HDF5, netcdf |
| Text | TXT, XML, HTML, ASCII |
| Web archive | WARC |

### 2.3.2. What about spreadsheets (Excel, Access)?

In biological research, spreadsheets (e.g., Excel, Access) are widely used as a tool for digitizing and storing data. Although the use of simple spreadsheets does not require high level computer skills and might be adequate in small projects with a limited number of collaborators, a **non-negligible risk of creating incorrect data or even losing data** is associated with the use of these programs. Therefore, if you decide to use spreadsheets for managing your project's data, it is important to use them with caution and in a well-organised way.

**Guidance**

- Keep data (raw and intermediate level data) in a tidy format. A dataset is said to be tidy if it fulfills the following conditions:
    - individual observations are in rows
    - variables are in columns
    - each type of observational unit is contained in a single dataset (sheet or file)
- When splitting data in between different campaigns/years, make sure the individual data files are consistent which would support concatenation of the individual data tables
    - Similarly, making clearly structured templates for data import is useful in case collaborators are supposed to make measurements and digitize data (e.g., when working with students, volunteers,...).
- When using spreadsheet software:
    - **NEVER work in a spreadsheet with 'raw' data**, always use separate files for raw data and data analysis. Make links between the analysis files with raw data files on order to work with the most recent version of the dataset.
    - Put different datasets on different sheets.
    - Make sure each data table has column headers and row indices.

- Always keep **graphs in separate sheets.**
- **Document** the individual processing steps on a separate sheet.

### Examples

Data tables are often presented like shown in table 2, which is a good format for presenting data, but it is hard to run analyses in this format. In order to keep some overview in your data files and facilitate statistical analysis, it is better to put each individual observation in a different row (tidy data format, table 3).

Table 2 - Untidy table format, often used for presenting data summaries.

| **Species** | **Location A** | **Location B** |
|---|---|---|
| Red fox | 1 | 5 |
| Common toad | 55 | 12 |
| Common wood pigeon | 4 | 22 |
| European hedgehog | 5 | 11 |

Table 3 - Tidy version of table 2, suitable for calculations, long-term storage and data sharing.

| **Species** | **Location** | **Roadkill count** |
|---|---|---|
| Red fox | A | 1 |
| Red fox | B | 5 |
| Common toad | A | 55 |
| Common toad | B | 12 |
| Common wood pigeon | A | 4 |
| Common wood pigeon | B | 22 |
| European hedgehog | A | 5 |
| European hedgehog | B | 11 |

### 2.3.3. Data standards for biodiversity research

Structure your occurrence and environmental data as **Event-core**[3] oriented data tables:

- The *Event Core* brings together data from the same sampling area, or from an atlas, connected via their 'parent' event.
- Occurrences are provided in the *occurrence.txt* data table.
- Additional measurements (e.g. abiotic factors) are stored in the *measurementorfact.txt* table and linked to the proper (Parent)EventID.
- The events can also be used to relate data in time, for instance, from in situ measurements that are repeated every year.
- Sampling protocol can be documented with respect to the defined (Parent)EventIDs.

### 2.3.4. Naming conventions

**DMP question**

*What naming conventions do you follow?*

**Guidance**

- Do not use generic file names that may conflict when moved from one location to another. If you work on more than one computer ensure that your files are continuously synchronized.
- Files should be distinguishable from each other within their containing folder. Make file names **unique**, if possible.
- File names should outlast the file creator who originally named the file.
- Consider how scalable your file naming policy needs to be e.g. if you want to include the campaign number, do not limit that number to one digit, or you can only have 9 campaigns. Rather use 007 instead of 7, which leaves the opportunity to name up to 999 files instead of 9.
- Keep file names **short and relevant** - generally about 25 characters is a sufficient length to capture enough descriptive information for naming a data file.
- Do **not use special characters** in a filename such as : & * % $ £ ] { ! @ as these are often used for specific tasks in different operating systems.
- **Use underscores (_)** instead of full-stops or spaces because, like special characters, these are parsed differently on different systems.
- Try to find a naming convention where files can be **sorted in a logical sequence** (for example by adding the date at the beginning as YYYY-MM-DD).
- If including **dates**, format them consistently according to the proper standard (ISO 8601), e.g., `2010-08-11_interview_Jane_Doe`.

[3] http://www.geobon.org/Downloads/brochures/2016/The%20EventCore-brochure-2016.pdf

- Assume that `FILENAME`, `filename` and `Filename` are the same, even though some file systems consider them as different. Usage of capital letters should not differentiate file names.
- Where possible, use **file extensions** (often defaults) to accurately reflect the software environment in which the file was created and the physical format of the file. E.g., use `.xls` or `.xlsx` for Excel files, `.txt` for text files, etc.

**Examples**

2017-05-09_IN_Prj_Loc_Var.csv

Which includes the capture date (2017-05-09), initials of the investigator (IN), project acronym (Prj), location of sampling (Loc), and measured variable (Var)

### 2.3.5. Folder organisation

**DMP question**

*How will you structure and name your folders and files?*

**Guidance**

- Use a **well-designed folder structure**[4] for a project, which is consistent to a predefined general project folder layout at your research group.
- Use relative file names to the *root* of the project folder in scripts, programs,...
- Keep **raw data** files **read-only** when possible, to prevent accidentally unintended manipulations.
- For non-digital objects create an index (e.g., stored soil samples).

**Examples**

As an example, consider the following basic folder structure:

```
|- LICENSE
|- README.md <- The top-level README describing the general layout of the project.
|- data
|    |- raw  <- The original, read-only acquired raw data.
|    |- interim <- Intermediate data that has been transformed.
|    |- processed <- Final data products, used in the report/paper/graphs.
```

[4] https://drivendata.github.io/cookiecutter-data-science/#directory-structure

```
|     |_ external <- Used additional third party data resources (e.g. vector maps).
|- docs     <- Project specific literature.
|- reports  <- Reported outcome of the analysis as LaTeX, Word, markdown,...
|     |_ figures  <- Generated graphics and figures to be used in reporting.
|- src    <- Set of analysis scripts used in the analysis.
|       |_ readme.md  <- Explanation of the logical structure of the scripts and their purpose.
```

### 2.3.6. Software and tools

**DMP question**

*Which software is needed to open the data files?*

**Guidance**

- When possible use open source software. Doing so, anyone will be able to open the files.
- If this is not applicable because you need specific software to use lab machines or for analyses, give a clear overview of the software you used to produce the data, including the version of the software and the software provider.
- If applicable, export the data created in closed source software to a simple tabular data format (e.g., csv or txt files).

### 2.3.7. Versioning

**DMP question**

*How will you manage different versions of a data file?*

**Guidance and examples**

- **General guidelines**
    - Backup changes as soon as possible.
    - Keep changes small.
    - Share changes frequently with collaborators in the project.

- **Versioning option 1: manually track changes**
    - Add a CHANGELOG.txt file to the folder where you keep the data. In this log-file, you briefly document the changes that you made.
    - Copy the entire project after large changes.
- **Versioning option 2: Use a version control system (preferred)**
    - Git (https://git-scm.com)
    - GitHub, Bitbucket, Gitlab,...

# 3. Data quality

## 3.1. Data utility

**DMP question**

*Will the data be useful for other users?*

**Guidance**

Especially if you expect the data to be useful for other researchers, you should aim to organize your data following international standards from the start of the project. Make an estimate to whom (parts of) your data could be useful.

**Examples**

In biodiversity research, simple occurrence data are highly reusable in other contexts. In case you are collecting species occurrence data, it is advisable use the Darwin core standard as it facilitates later publication at GBIF.org and data sharing with other researchers.

## 3.2. Data quality documentation

**DMP question**

*Are data quality assurance processes described?*

**Guidance**

Explain how the consistency and quality of data collection will be controlled and documented.

**Examples**

Provide an overview of the methods used to check data quality, e.g.,

- calibration methods
- repeated sampling or measurements
- standardised data capture or recording
- data entry validation,
- peer review of data
- representation with controlled vocabularies

**Data quality and error** are often neglected in environmental sciences. A loss of data quality at any stage of the data manipulation process reduces the applicability and uses to which the data can be adequately put (Chapman 2005). Therefore, it is crucial to document the following processes and, when possible, to keep the data files generated in each step:

- Data capture and recording at the time of gathering.
- Data manipulation prior to digitisation (label preparation, etc.).
- Identification of the collection (specimen, observation) and its recording.
- Digitisation of the data.
- Documentation of the data (capturing and recording the metadata).
- Data storage and archiving.
- Data presentation and dissemination (paper and electronic publications, web-enabled databases, etc.).
- Using the data (analysis and manipulation).

**Taxonomic datasets** should consist of the following items (although this information is not always available, especially in older datasets):

- name (scientific, common, hierarchy, rank)
- nomenclatural status (synonym, accepted, typification)
- reference (author, place and date of publication)
- determination (by whom and when the record was identified)
- quality fields (accuracy of determination, qualifiers)

# 4. Data description

## 4.1. Metadata

**DMP question**

*Are the data produced and/or used in the project discoverable with metadata?*

*Describe how and when metadata will be recorded. Which standards will be used?*

**Guidance**

Metadata[5] explains the origin, purpose, time, geographic location, creator, access, and terms of use of the data. Typically it is used for resource discovery, providing searchable information that helps users to easily find existing data and as a bibliographic record for citation.

In general, metadata is a subset of core standardized and structured data documentation that contains the following components:

Table 4 - Components of a metadata file.

| | |
|---|---|
| **Title** | Name of the project and the dataset |
| **Creator (*who?*)** | Names and institutions of the people who created the data |
| **Date (*when?*)** | Key dates associated with the data, such as dates covered by the data or date of creation |
| **Content (*what?*)** | Description of the dataset |
| **Coverage (*where?*)** | Geographic coverage |
| **Purpose (*why?*)** | Description of the project scope |
| **Origin (*how?*)** | Methods used for generating the data |
| **Keywords or subjects** | Keywords or subjects describing the content of the data |
| **Identifier** | Unique number or alphanumeric string used to identify the data (e.g., DOI, see standard identification systems) |
| **Language** | Language of the resource |
| **Publisher** | Entity responsible for making the dataset available |
| **Funding agencies** | Organization or agency who funded the research |

[5] http://data.library.arizona.edu/data-management-tips/data-documentation-and-metadata

| | |
|---|---|
| **Access restrictions** | Where and how are your data accessible by other researchers? |
| **Copyright** | Terms of use of a data collection |
| **Format** | What format is your data in? (see file formats) |

**Examples**

- For many data types metadata standards exist, e.g., Darwin Core for biodiversity data which is used by GBIF.org.
- For geospatial data the ISO19115 standard is a commonly used metadata standard.
- For biology research, a list of accepted metadata standards is available on the DCC website.
- Writing a **data paper** could also help to get your data discovered and cited by other users. As such a paper practically includes all metadata, the reader will fully understand your data. Examples of journals that publish biodiversity data papers are ZooKeys and PhytoKeys. Or you could even publish your data management plan as such.

## 4.2. Documentation

**DMP question**

*How will data and the project be documented for future reference?*

**Guidance**

A good documentation ensures that data will be understood and interpreted by any user. Depending on the scope and size of the project, this can be achieved at project level and at the level of individual datasets.

### 4.2.1. Project documentation

Add a README.txt to your project folder which includes the main information of the DMP:

- Context of data collection
- Data collection methodology
- Structure and organization of data files
- Data validation and quality assurance
- Data manipulations through data analysis from raw data
- Data access and use conditions

### 4.2.2. Data documentation

Generate a README.txt for each dataset including:

- Variable names and descriptions
- Definition of codes and classification schemes
- Codes of, and reasons for, missing values
- Derived data created after collection
- Definitions of specialty terminology and acronyms
- Algorithms used to transform data
- File format and software used
- Data standards applied, exchange standards, use of community specific control vocabulary,...?

# 5. Data use

## 5.1. Purpose

**DMP question**

*What is the purpose of data collection? What are the research questions in the project?*

**Guidance**

An abstract of the project, mentioning the research questions gives an overview of the concepts of the project to anyone interested.

## 5.2. Analyses

**DMP question**

*What statistical analyses or models will be built?*

**Guidance**

Briefly summarize the analyses, models and scripts you expect to be generated in the project.

## 5.3. Expected output

**DMP question**

*Which output will be generated during the project?*

**Guidance**

Make a list of all expected outputs from the project. Output types are:

- Datasets
- Data papers
- Scientific papers
- Reports, books,...

# 6. Data storage and archiving

## 6.1. Storage

**DMP question**

*How will you provide secure storage of data during the project?*

**Guidance**

### 6.1.1. Storage guidelines

- Accessibility of any data is dependent on the quality of the storage medium and availability of software to view the data.
- Store data in a non-proprietary or open standard format for long-term readability (see file formats).
- Check the data integrity of stored data files at regular intervals.

### 6.1.2. Storage devices

Storage devices vary in their suitability for long-term storage and the security of the stored data (figure 11).

- Personal devices as laptops, computers or tablets are convenient for short-term and temporary storage but should not be used as the sole storage location of master files. These devices are at high risk of being stolen, lost or damaged.
- Network drives managed by IT staff (e.g., university servers) with a regular backup scheme ensure secure storage of data.
- External devices (e.g., hard drives, USB sticks, CDs, DVDs) are cost effective and portable. However, they are not fit for long-term preservation due to physical degradation and can be lost, stolen or damaged. If they are used for temporary or short-term storage, high-quality devices are recommended.
- Online storage services (e.g., Dropbox, Google drive, OneDrive) provide cloud technology which allows users to synchronize files across different computers.
- Online repositories ensure the safe storage of final research outcomes. The use of subject-specific repositories (e.g., GBIF for species occurrence data, or GenBank for genomic data) is recommended.

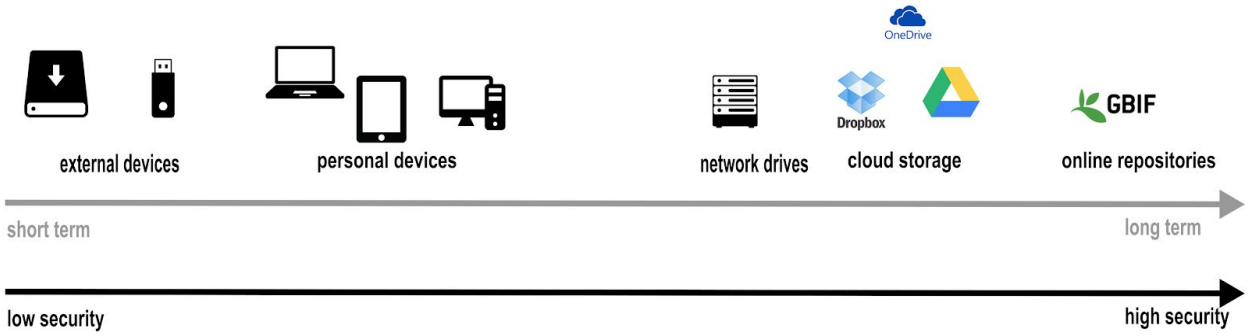



Figure 11 - Overview of available storage media, their suitability for long-term preservation and their reliability.

## 6.2. Back-up

**DMP question**

*How will data be backed-up in the project?*

**Guidance**

- Follow the 3 - 2 - 1 rule: 3 copies - 2 different types of media - 1 offsite.
- Create a back-up schedule.
- Use reliable back-up media.
- Test your back-up system by testing file restores.
- **Do not exclusively use (archival) CDs or DVDs** since these media are subject to physical degradation and can easily get lost.

**Examples**

The following storage media are some reliable locations for backing up your data:

- Personal Computer (short term)
- Departmental or University Server (medium - long term)
- Tape Backups (medium - long term)
- Subject archive (online repository) (long term)
- External Hard Drives (short term)
- Cloud Storage (medium - long term)

## 6.3. Preservation

### 6.3.1. Preservation methods

**DMP question**

*How will you safeguard your research data after finishing the project?*

**Guidance**

After finishing a short-term project or after a couple of years in long-term projects (e.g., continuous monitoring projects), it is recommended to archive data in an **online repository** where the data will be safeguarded and preserved on a long-term.

### 6.3.2. Preservation period

**DMP question**

*How long is it intended that the data remains reusable?*

**Guidance**

Select data for long-term archiving and preservation after finishing the study.

**Examples**

Data which cannot be remeasured and which have a high value for the scientific community should be preserved on a long-term (or even indefinitely).

# 7. Data dissemination

## 7.1. Publication

### 7.1.1. Open data

**DMP question**

*Which data produced and/or used in the project will be made openly available as the default?*

**Guidance**

Follows the principle **"as open as possible, as closed as necessary"** which focuses on encouraging sound data management as an essential part of research best practice.

If certain datasets cannot be published (or need to be shared under restrictions), explain why, clearly separating legal (e.g., privacy, copyright) and contractual reasons from voluntary restrictions.

Describe default options of open data publication and possible restrictions in the DMP. How will you ensure that data publication is agreed between the partners covering the funder's requirements?

**Examples**

The publication of biodiversity data as open data is most often not restricted by law. You could also consider to use an embargo period on publishing your data and make your data open after e.g. the acceptance of a research paper.

### 7.1.2. Data publishing

**DMP question**

*How will the data be made accessible?*

**Guidance**

Consider where, how, and to whom the data should be made available. Most research funders recommend the use of established data repositories, community databases and related initiatives to aid data preservation, sharing and reuse.

**Examples**

- For species occurrence data in biodiversity research, you can upload data to the Global Biodiversity Facility (GBIF.org) network and its national and thematic nodes including the Freshwater Information Platform.
- All kinds of datasets can be uploaded in general-purpose open access online repositories such as zenodo.org and datadryad.org, although these repositories do not require the use of standards which reduces the interoperability and reuse of data.

### 7.1.3. Interoperable data

**DMP question**

*Are the data produced in the project interoperable? What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?*

**Guidance**

Data are interoperable in case data exchange and reuse is facilitated between researchers, institutions, etc. To achieve this, the metadata need to be defined. In case data and metadata are produced in a standard format which is compliant with available (open) software applications, the recombination of different datasets is facilitated.

**Examples**

In biodiversity research, the Darwin core standard is used for occurrence data, while EML (ecological metadata language) is a standard specifically developed for ecological metadata. More biology related metadata standards can be found at the DCC website.

### 7.1.4. Restrictions and usability of data

**DMP question**

*Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the reuse of some data is restricted, explain why.*

**Guidance**

Restrictions to data sharing may be due to participant confidentiality, consent agreements or intellectual property rights. Strategies to limit restrictions may include: anonymising or

aggregating data, gaining participant consent for data sharing, gaining copyright permissions, and agreeing a limited embargo period.

**Examples**

In most cases, legal restrictions are not applicable for biodiversity data.

### 7.1.5. Embargo periods

**DMP question**

*When will the data be made available for reuse? If applicable, specify why and for what period a data embargo is needed.*

**Guidance**

In some cases, an embargo period is useful, e.g., when waiting for patents, or getting research published.

**Examples**

Data could be published at the moment a research paper is accepted.

## 7.2. FAIR data

The concept of FAIR data involves that data are **f**indable, **a**ccessible, **i**nteroperable and **r**eusable. It encompasses a list of concise and measurable principles. The intent is to enhance the reusability of research findings and to extract maximum benefits from research investments (Wilkinson et al. 2016). International funding programs often build their DMP template upon these principles (e.g., H2020). In this document, the FAIR principles are covered in various sections, therefore an overview is given in box 2, including links to the respective sections.

Box 2 - The FAIR Guiding principles (Wilkinson et al, 2016).

### The FAIR data principles

**Findable data**:
F1. (meta)data are assigned a globally **unique and persistent identifier**
F2. data are described with rich **metadata**
F3. metadata clearly and explicitly include the **identifier** of the data it describes
F4. (meta)data are **registered or indexed** in a searchable resource

**Accessible data**:
A1. (meta)data are **retrievable** by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are **accessible**, even when the data are no longer available

**Interoperable data**:
I1. (meta)data use a **formal, accessible, shared, and broadly applicable language** for knowledge representation.
I2. (meta)data use **vocabularies** that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

**Reusable data**:
R1. meta(data) are **richly described** with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible **data usage license**
R1.2. (meta)data are associated with detailed **provenance**
R1.3. (meta)data meet **domain-relevant community standards**

## 7.3. Citation

**DMP question**

*How should the data be cited?*

**Guidance**

The use of persistent and unique identifiers such as **DOIs** (digital object identifiers) for datasets is recommended as helps:

- to overcome confusion over multiple versions of a given resource
- to improve the ease of locating resources
- to promote interoperability
- it is a reference to the authenticity of resources on a longer term,
- it can be used to cite datasets if they are used in new publications

**Examples**

Open access online data repositories often used in biodiversity research, such as GBIF.org, zenodo.org, datadryad.org provide a persistent DOI for uploaded datasets.

# 8. Ethics and legal compliances

## 8.1. Ethical aspects

**DMP question**

*Are there any ethical or legal issues that can have an impact on data sharing?*

**Guidance**

- Make sure you reach a consent with your collaborators about data preservation and sharing.
- Be careful when using (commercial) cloud services such as Dropbox, as the security of your files is not always guaranteed. Especially when dealing with sensitive, restricted or confidential data, the use of cloud storage is not recommended.

## 8.2. Data policy

**DMP question**

*Is there a data policy in your institution? How will you comply with these regulations?*

**Guidance**

- If there is a data policy at your institution, refer to it in the DMP.
- If the regulations in your institution's data policy conflict with the funder's regulations, explain how you will overcome this.

**Example**

Your institution's data policy only foresees in open data publication 5 years after data collection, while the funder requires open data publication after 2 years.

## 8.3. Licences

**DMP question**

*How will the data be licensed to permit the widest reuse possible?*

### Guidance

Specifically for H2020 projects, the EUDAT B2SHARE tool has been developed for the storage of research data. It also contains a built-in license wizard that facilitates the selection of a license for research data. Furthermore, seven license types are regularly used (table 5). See also the creative commons website for more information on licensing.

Table 5 - License types[6] for data publishing.

| Icon | Description | Acronym | Allows Remix culture | Allows commercial use | Allows Free Cultural Works | Meets 'Open Definition' |
|---|---|---|---|---|---|---|
|  | Free content globally without restrictions | CC0 | Yes | Yes | Yes | Yes |
|  | Attribution alone | BY | Yes | Yes | Yes | Yes |
|  | Attribution + ShareAlike | BY-SA | Yes | Yes | Yes | Yes |
|  | Attribution + Noncommercial | BY-NC | Yes | No | No | No |
|  | Attribution + NoDerivatives | BY-ND | No | Yes | No | No |
|  | Attribution + Noncommercial + ShareAlike | BY-NC-SA | Yes | No | No | No |
|  | Attribution + Noncommercial + NoDerivatives | BY-NC-ND | No | No | No | No |

### Examples

For biodiversity data, the best license is often the Creative Commons Zero (CC0) license under which data are made available for any use without restriction or particular requirements on the part of users. For example, the Research Institute for Nature and Forest (INBO) uses the CC0 license as a default for open data publication.

[6] See https://en.wikipedia.org/wiki/Creative_Commons_license

## 8.4. Legislation

**DMP questions**

*Are there any legal restrictions on the use or publication of data?*

**Guidance**

In certain cases, specific legislations apply to collection data (e.g., when sampling protected species, or when using vertebrate animals in experiments) and to the publication of data (e.g., the privacy legislation in case of medical records).

## 8.5. Privacy

**DMP questions**

*Have you gained consent for data preservation and sharing? How will you protect the identity of participants if required? How will personal data be handled to ensure it is stored and transferred securely?*

**Guidance**

Ethical issues affect how you store data, who can see/use it and how long it is kept. Managing ethical concerns may include: anonymisation of data, referral to departmental or institutional ethics committees, and formal consent agreements. You should show that you are aware of any issues and have planned accordingly. If you are carrying out research involving human participants, you must also ensure that consent is requested to allow data to be shared and reused.

If you are collecting personal data (including email addresses) check whether your methods are GDPR compliant.

# PART III: Generic DMP template

*To be uploaded in the next version*

# References

Chapman, A. D. 2005. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Gibney, E., & Van Noorden, R. (2013). Scientists losing data at a rapid rate. Nature News. doi:10.1038/nature.2013.14416

Hart, Edmund M., et al. "Ten simple rules for digital data storage." PLoS computational biology 12.10 (2016): e1005097.

Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11(10): e1004525. doi:10.1371/journal.pcbi.1004525

Vines, Timothy H., et al. "The availability of research data declines rapidly with article age." Current biology 24.1 (2014): 94-97.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*.

Wolkovich, Elizabeth M., James Regetz, and Mary I. O'connor. "Advances in global change research require open science by individual researchers." Global Change Biology 18.7 (2012): 2102-2110.