SWISS POLAR INSTITUTE

# Data management in the field

Jen Thomas, Swiss Polar Institute

September 2021

DOI: 10.5281/zenodo.5531876

Rue de l'Industrie 17
1950 Sion

Switzerland
+41 21 693 76 74

secretariat@swisspolar.ch
www.swisspolar.ch

# Contents

# Chapter 1

# Introduction

## 1.1 Benefits of good data management

Investment of time in good preparation and management of data files, software, documentation and associated files is an investment for the future. Setting time aside and having a routine for preparing detailed documentation, organising files and doing backups, will make life a lot easier for yourself and others in the future.

## 1.2 What this guide covers

In this guide we cover good research data management practice with a focus on field work. The first part of the guide is can be applied throughout all stages of the research life cycle. The second part is specifically about data management in the field and is broken into three sections: planning before you go, managing data whilst in the field and a set of priorities for when you get back.

We regularly refer back to the key sections of the first part of the document and recommend that you are familiar with this before moving on to the second part.

### 1.2.1 Scope

Research data includes raw data directly recorded by hand or from an instrument, processed data which are any files that have been modified, documentation, metadata, supporting data files, code, plots, photographs and any other files associated with the research.

This guide is applicable to all file types and "kinds" of data.

We also consider good practice of sample collection from a data management point of view. This section has a particular focus on labelling and collection of metadata about samples.

Finally, we consider field work, and more precisely, polar field work (although the same principles apply anywhere) in the broad sense of not working at your institution. It could be at a field camp, remote operation of an instrument, on a well-established base, on a ship or any other platform, or wherever else you might be taken by your research.

## 1.3 Further development of this guide

Ongoing development of this document is done in a GitHub repository: https://github.com/Swiss-Polar-Institute/data-management-guidance and the latest PDF version is published online: https://doi.org/10.5281/zenodo.4185227. Text is written in Markdown, then Pandoc is used to convert it to LaTeX and then PDF format.

We would greatly welcome feedback by way of comments, suggestions, corrections, questions or ideas for further development of this guide. Feel free to get in touch using GitHub issues or email jen.thomas@swisspolar.ch.

# Part I

# Data management - good practice

# Chapter 2

# Metadata

We begin with metadata because without this essential information, data loses a huge amount of its value and understanding.

## 2.1 What is metadata and why is it important?

Metadata is a description of data (McCarthy, 1982; Gray *et al.*, 2005; Michener, 2005), giving information about the "who, what, when, where, why and how", it was collected (Recknagel and Michener, 2018). Put another way, it is "data about data" (Fegraus *et al.*, 2005).

Information about a dataset quickly deteriorates over time (Michener *et al.*, 1997). This information, metadata, is documentation that is a reminder to your future self (and others). Without information about how data and samples were collected, final results cannot be interpreted correctly (Smith *et al.*, 2015).

The following questions are just some examples of things that might affect data or sample collection.

- Will an acronym that was used in a data file, mean something to someone in ten years?
- Why was a sample collected in an unexpected location?
- Why does a particular sample contain double the material than the other samples from the same location? (This sample wasn't collected on time because of access problems due to poor weather).
- Which instrument models were used to collect the data? Could there have been a difference in the data collected because one of the instruments was an older model?

## 2.2 How is metadata used?

Metadata serves as a **useful reminder** from the inception of a project, through its lifetime and well into the future, of how, why, when, where and by whom, data or samples were collected, how they have been transformed, and any information that may affect their state.

One source of metadata provided alongside the data avoids multiple explanations of the same points, confusion or misinterpretation when **collaborating** with others. Having metadata readily available ensures they are able to interpret and use the data in a suitable way.

Think also of **reproducibility**: for publications it is likely a set of detailed methods of how data were produced and results derived, will be needed. **Validation** of datasets as part of the publication review process is becoming the norm and with that comes detailed metadata, with thorough documentation of how the data have been produced and processed. Keeping these notes will save a lot of time and avoid forgotten details when writing up a project, as well as being useful for others who would like to do a **similar study** or try to **replicate results**.

Detailed metadata allows others to **interpret and understand** data and results.

It is often a requirement of funders, and is good practice, to **publish data** openly so that it is available for other researchers and the general public. Many platforms or repositories that provide these publishing services require a certain amount of metadata to ensure that a dataset is discoverable. It is also in the dataset producer's interest to give as much information as possible so that a future user properly understands the suitability of the dataset for their purposes, as well as its limitations. Providing a good description of a dataset means a potential user is much more likely to use it in the future. In addition to citations arising from sharing datasets, research has shown that citing datasets clearly in a publication leads to more citations and goes some way to giving further credibility to a publication because it is more reproducible (Colavizza *et al.*, 2020).

## 2.3 What should be recorded?

Metadata should be captured about different aspects of a project and at all stages.

### 2.3.1 Samples

It should be possible to follow the life history of a sample from its collection to the final dataset, then be able to find the sample in its final resting place (or know that it no longer exists).

**Collection:** include where, when, by whom, methods and particular conditions that might affect sample collection or resulting data. Also note specific details of any instruments (include manufacturer, model, serial number and set-up) used for sample collection.

**Storage conditions:** record how the samples were stored and what happened to them between collection and processing. Information could include mode of travel, storage temperature and conditions, and how they were prepared for processing.

**Storage location:** record where samples are currently stored (or if they were destroyed as part of the processing) and how they can be found.

**Processing:** include where and when the processing was done, as well as who did it, techniques, protocols, guides and which standards were used. Also include specific details of any instruments that were used (manufacturer, model, serial number and set-up).

**Supporting datasets:** include information about any supporting datasets (e.g. meteorological conditions) from the time of sample collection.

**Resulting datasets:** describe where datasets resulting from processed samples, can be found and how they are linked to a particular sample.

### 2.3.2 Data files

Keep a README file alongside the data. Information within the README file will depend on the complexity of the data and how they are stored. It might include some or all of the following information about the files and data itself:

**File description:** explain what the files are and their purpose. Include each parameter name, a full description and units. Describe in full any acronyms or shorthand that are used within the data files.

**File access:** describe who has access to the files and if there are any limitations on what should happen to them.

**Data collection:** describe by whom, where, when and how the data were collected. The conditions and instrument (manufacturer, model, serial number and set-up) used for collection are particularly important. Describe any problems that occurred during data collection and any known problems with the data files. This might include

times when the instrument malfunctioned, particular conditions that might affect the parameters of interest, periods of missing data or anomalies that were noted.

**Data file transfer:** keep full records of how data files were saved, backed up and transferred between different storage types during collection and afterwards. Where possible, use scripts to do this and keep them as a record.

**Data file manipulation and processing:** wherever possible, use scripts to manipulate data files as this creates a record of what has been done whilst working on the data. Keep and backup these scripts. It is also good practice to keep human-readable notes alongside this, particularly about which decisions were taken and why. When describing data files, if no processing has been done to the files, also make this clear.

**Dataset rights holders:** note details of the rights holders of the dataset (person(s) and / or organisation(s)), who contributed to the production of the dataset, as well as what they did. Full details of authors, contributors and funders are normally asked for when publishing a dataset.

### 2.3.3 Project

It is easy to lose information about the context of data and samples when projects come to an end. More detailed information, particularly about methods, might be lost because it cannot all fit in a journal article.

**Project description:** maintain a good description of the project to which datasets and samples contribute. Include what the study set out to do, how it contributes new knowledge to the field, what the research questions/hypotheses were, and which methodologies, instruments and measures were used (The University of Edinburgh, 2021).

## 2.4 README.txt

**Always** include a README file within a directory to describe its contents. This will help anyone looking at the files in the future.

The Gurdon Institute (Downie, 2019) provides a very useful list of what to include here, summarised as follows:

- summarise what is in the directory
- use keywords for the project, data type or parameters so that they can be searched in the future
- include the name of the person(s) who created the directory and their contact details
- describe any changes made to the directory and when
- explain file naming conventions
- details of backups: how often and where to, how they can be accessed
- make sure the file is written and saved in plain text format so that it can easily be read in the future.

---

*Example*

– Data collection –

An instrument (name, manufacturer, model, serial number) was installed on the southern side of a building at a station in Antarctica (name, coordinates, altitude). The instrument was installed from 20th December 2020 until 20th December 2021. Testing was done from 20th - 31st December 2020 and 24-hour data collection began on 1st January 2021.

The instrument operated 24-hours a day, saving data directly onto a connected laptop which ran Windows 10. Software (name, version number, reference) saved the files containing 24 hours of data with one-second resolution. Data were saved into daily files, named in the format instrumentName_antarctica_YYYYMMDD.csv where YYYYMMDD is the date (UTC) on which the data were recorded. No processing of the data was done in the

field; these data are as-recorded. Data are in ASCII format and can be read by text editors, spreadsheet software and other software. Parameters follow definitions of the Climate and Forecast (CF) Metadata Conventions.

– Data backup in the field –

Data were backed up automatically onto a portable hard drive once a day at 20:00 using rclone (version 0.3). Hard drives were rotated every week and each time they were swapped, data were copied to a third hard drive.

– Instrument calibrations –

Instrument calibrations were done once a week on a Monday. During the calibration, the instrument did not operate normally, but instead data for the calibrations were saved. These are part of the normal data files. A record of calibration dates and times can be found in the file instrument_calibrations.csv.

– Data transfer –

After the expedition, data were copied to network attached storage at Institution X using rclone (version 0.3). Another backup of the raw data was transferred to Amazon Deep Glacier storage using rclone (version 3.0) straight after the expedition. Both copies of the data are read-only for all users. They should be copied elsewhere for further processing.

– Log –

The laptop crashed on 30th March 2021 and the data recorded before 16:34 UTC on this day, were lost. Bad weather meant the hard drives couldn't be changed on 6th June 2021 so one remained with an extra 4 days of backups. The calibration was not done on 7th June 2021 because the instrument could not be accessed due to bad weather. An interruption in the power supply to the instrument occurred on 12th July 2021 and the instrument could not be accessed and the power supply fixed until 19th July 2021. There are no data from this period.

– Project –

Data were collected as part of the project, "Meteorology in Antarctica", funded by the Antarctic Funding Organisation (grant number ABC1234, PI: name). Data are owned by PI and PhD student (names). Data were collected by PhD student A, assistant B and assistant C.

---

## 2.5 Machine-readable metadata

It is important for humans to be able to understand if a dataset is fit for their purpose, something normally achieved by reading documentation. However, if metadata are structured and documented correctly, they can be easily read by machines (software). The advantages of machine-readable metadata are to ensure that datasets are discoverable in searches, easier to combine with other datasets and meet community standards, which in turn allows comparison of similar datasets.

Some machine-readable metadata schemas to consider are:

- schema.org (https://schema.org) - good for describing any digital object
- Frictionless Data (https://frictionlessdata.io/) - good for describing data files in a tabular format, as well as the dataset as a whole
- DataCite Metadata Schema (https://schema.datacite.org) (DataCite Metadata Working Group, 2019) - often used for dataset Digital Object Identifiers (DOIs)

## 2.6 Controlled vocabularies

A controlled vocabulary normally provides a set of keywords or parameter names that have a specific definition and unique identifier. Using these vocabularies where they exist and are applicable to a particular dataset, makes datasets more comparable and understandable, particularly when working across domains.

There are many controlled vocabularies, so it pays to look into this closely within a particular discipline or ask for assistance as necessary.

Some commonly used vocabularies for environmental datasets are:

- SeaDataNet - oceanography terms at many levels
- Global Change Master Directory (GCMD) Keywords - environmental data keywords, instrument names, geographical names and more
- Climate and Forecast (CF) Convention - parameter names with definitions and units, mostly related to climate, forecasting and atmospheric science
- Ecological Metadata Language (EML) - vocabularies and syntax for documenting earth and environmental science data.

It is very difficult to have complete and exhaustive lists of controlled vocabularies as science is constantly evolving, so many of these providers of controlled vocabularies allow users to request additions. Do consider doing this if there is a term that would be of use.

# Chapter 3

# Storing data

In this section, we focus on types of data storage. Suggestions are *not* ordered by preference.

Local IT and data management teams will be able to advise on infrastructure that is available at your institution, so this would always be a good place to start.

## 3.1 What to consider when deciding where to store data

Think carefully about the data that need to be stored, how they will be used and for how long. Consider the following questions:

- What do you want to do with the data?
- Do others need access to the data?
- What is the volume of the data?
- How long do the data need to be kept for?
- What budget is available for data storage?
- How much does the data storage cost?
- Are the data sensitive or personal?
- Does the data owner have any requirements about where the data are stored?

## 3.2 Types of storage media

Different types of media offer solutions to different problems and therefore will need to be considered accordingly. It is important to have backups of data in multiple places and on different types of media, so more than one media type will likely be necessary.

### 3.2.1 Laptops or desktop computers

**Laptops or desktop computers should not be used to store the primary copy of a dataset.** Failure of the hard disk, theft or damage could mean loss of all the data and associated files.

If using such a computer to work on data, always work on a *copy* of the raw data, ensuring that a master copy is backed up elsewhere and any work done is backed up on a regular basis.

### 3.2.2 Networked data storage

Networked data storage, such as an institution file server, is often provided as standard by institutions and can be a useful and reliable place to store data. There are some important things to understand before placing data (particularly sensitive, personal or large data) on institutional file servers.

**Get familiar with it:** it is important to properly understand the details of the storage, such as how to access it, how it is backed up and what the backup schedule is.

**File recovery:** is this possible and if so, how? Check if the recovery of previous versions of files or deleted files is possible, or if it is only possible to recover data if a hard drive is corrupt, for example. Find out if it is possible to recover files yourself or if the help of a systems administrator is required, how to do the recovery and how long it takes.

**Check access rights:** are personal areas of data storage available, or is access shared amongst lab personnel or a wider group? Consider if there should be restricted access to the data, particularly if it is sensitive (e.g. information on protected species) or contains personal data. Access rights should also be limited to avoid changing or deleting files accidentally. Ensure that primary (raw) data are read-only for everyone. Check if external collaborators can be given access to the data if this is an important consideration.

**Storage volume:** check if there are limits on the volume of storage that is permitted. If the volume of data is hundreds of GB or some TBs, check that accessing the files on this kind of storage is reliable and suitable for your needs.

**Off-site access:** access to data may be required from off-site from time to time. Find out early on how to do this. Make sure access to the data works and is appropriate for your particular needs.

### 3.2.3 Portable media

Portable hard drives are commonly used for temporarily backing up data in the field, but their use-case should be carefully considered. Some portable media types are not useful for long-term storage because they quickly degrade or become obsolete.

**Buying something suitable:** buy reputable makes of portable hard drives. If several are required (for example, during a particular field season), buy at least two different makes to avoid buying a "bad batch". Consider what kind of USB-port will work with computers in the field (USB-A or USB-C).

**How many?** Consider buying several medium-sized hard drives rather than fewer large ones. Label them well and always have at least two backup copies of the data (original, plus two backups) on separate disks that are stored separately. Consider how long it takes to back up your data and bear this in mind for planning purposes. If using an older device, it may have USB 2.0 only which is considerably slower than USB 3.0.

**Set them up:** A portable hard drive may need to be formatted before use, depending on which operating system(s) it will be used with. Before being taken into the field, the hard drive should be used and checked. Check also that the cable is compatible with the computer with which it will be used in the field.

**Take care of hard drives:** they are susceptible to physical damage and depending on how many times they are written to, may only last a few years.

**Pen drives** are easy to lose and should not be considered as reliable for data backup.

**Consider encrypting hard drives:** if a hard drive is lost, it can be easily read by anyone. The hard drive should be encrypted if it holds any personal or sensitive data.

**Regularly check data on hard drives:** always have other backups. Check hard drives can still be accessed on a regular basis. They should only be used as a temporary or additional backup.

### 3.2.4 Cloud storage

Cloud storage is becoming the norm in many cases. There are many types of cloud storage that can be set up or bought as a managed or unmanaged service. Some institutions may have cloud storage that is available.

**Physical server location:** institutions or data owners may also have requirements about which countries or regions data should be stored in. Physical servers of the cloud storage should meet the needs of these regulations.

**Privacy policy:** check the privacy policy carefully to make sure the files cannot be used in any way by the cloud storage company.

**Access rights:** consider who might have or need access to the data and how access rights are administered.

**Costs:** some cloud storage providers charge not only for the data storage, but also for the number of files, copying data to and from the storage, as well as listing files. Consider how this could impact on costs.

**Storage type:** some providers offer different levels of storage which provide faster or slower access to data. Standard storage often has immediate access to data; a different option, sometimes known as "cold storage", might have access which takes 12-24 hours. The latter is often a low-cost option and good for long-term backups of data that are not actively being worked on.

**Managed or unmanaged systems?** As with any system, make sure that you properly understand how it works and what interaction you need to have with it for it to meet your needs. Particularly for unmanaged systems, some understanding of the system will be needed to copy and access files, change access rights and verify data. Many systems require the use of command-line tools for moving data around, such as `rclone` (Craig-Wood, 2014–2020). Always ensure it is easy to move data to another provider at a reasonable cost and in a reasonable time frame, for instance if there are changes to the services that are provided. Note that some providers may use proprietary formats which might make moving the data very hard, or the same provider may be required to access the data: in this case, it is not a recommended solution to data storage. Ensure the subscription is always maintained (if paid for), otherwise data may be deleted.

Wikipedia has a very handy comparison of online backup options. These would normally be used for medium to long-term storage, or backups.

#### 3.2.4.1 Note about object storage

Many cloud service providers (as well as institutions) are now using object storage to store data, rather than file storage, particularly for large data volumes or large numbers of files. Object storage has various idiosyncrasies in terms of differences to file systems that are useful to be aware of. For example, files are known as objects and there is no concept of directories (Chan, 2020a, 2020b). It is essential to understand how to work with the files if this kind of storage is being considered.

### 3.2.5 Data publication

When work on datasets in both the raw and finalised stages, has been completed, consider publishing them with a persistent identifier. Some scientific disciplines or institutions provide specialist repositories and assistance to do this, but more generic online repositories are available if this is not the case for your particular dataset. The registry for research data repositories (re3data) is a good place to start looking for a suitable place to make data available.

# Chapter 4

# File organisation

It is worth investing time and effort in ensuring a coherent directory structure and understandable file names.

## 4.1 Directory structure

Consider a set of higher level directories that can be used consistently across projects.

Raw or finalised data files should not be worked on directly, to avoid modifying them accidentally (see the section about working on data). For this reason, consider creating a work-in-progress ("wip") directory which can have different permissions and a different backup schedule.

Directories containing data may be organised differently according to the project. It is important to consider the number of files in a directory and size of files to make them easier to work with.

---

*Example*

```
projectName
|-- documentation
|-- plots
|-- processedData
|-- rawData
|   |-- fieldSiteA
|   |   |-- 2018
|   |   |-- 2019
|   |   |-- 2020
|   |-- fieldSiteB
|   |   |-- 2018
|   |   |-- 2019
|   |   |-- 2020
|   |-- fieldSiteC
|   |   |-- 2018
|   |   |-- 2019
|   |   |-- 2020
|-- README.txt
|-- wip
```

---

### 4.1.1 File size

Many small files will take longer to copy and be harder to work with than a single file of the same total size. Copying lots of smaller files to cloud storage can also increase the cost, which may be a factor to consider.

Avoid creating files that are more than 1 GB in size because in some cases, they can be difficult to read into memory.

### 4.1.2 Number of files in a directory

In addition to the size of files, it is important to consider the number of files in a directory. Whilst in the majority of modern file systems the number of files within a directory is not technically limited, having more than a few thousand in one directory can make them more tricky to work with. As a guide, around 10,000 files could be considered as a sensible maximum number within one directory but it will greatly depend on the technologies being used. If data collection will produce around this number, consider splitting them into sub-directories.

---

*Examples*

A directory structure for data collected in hourly files to avoid too many files in the same directory:

`YYYY/MM/DD/HH/YYYY-MM-DD-HHmmSS-waves.bin`

If there were files with different names another possibility is to put files in sub-directories, such as:

```
A/m/Amie.txt
J/o/John.txt
```

A more complex example would be to store files with the hash as a filename and a database table linking the logical filename with the hashed name. The first characters of the hash would be used for the directory structure. The database table would show:

```
1, Amie.txt, a/9/a9564ebc3289b7a14551baf8ad5ec60a.txt
2, John.txt, 0/5/056a3c5c319c5288dba5f48ac619ab70.txt
```

---

## 4.2 File and directory naming

Some key points adapted from Borer *et al.* (2009):

**Reflect contents:** use a file or directory name which accurately reflects what is contained within it. Splitting it into separate parts, such as project, title, year or location of collection, year of collection, data type, version number and the file type can help to have a hierarchical name and standard naming procedure.

**Acronyms:** use sparingly and only if necessary. Always explain them fully in the README file.

**Be concise:** some file systems have a limited number of characters that can be used in the full file path (directory path plus the filename) so keep names concise.

**Stick to letters and numbers:** special characters (non-ASCII characters) are unfortunately not well-supported by some software and can cause problems. Underscores ( _ ) and dashes ( - ) are conventionally used for separation of different parts of a filename. Avoid using spaces because they can cause some problems with different file systems. camelCaps (starting each word with a capital letter) can be used to separate words within filename sections.

**Versioning:** this can be done using the date in the format YYYYMMDD. Placing the date at the start of the filename can be useful. If more granularity is useful, then version numbering such as v01_01, v01_02 can be used as well. Placing this at the end of the filename is useful, although if detailed versioning such as this is required, consider using a versioning tool.

Files and directories will most likely be listed in alphabetical order. Prefixes, such as numbers or letters, used to order them are not helpful if they do not mean anything. If using numbers for versioning, always use leading zeros, e.g. for sites that are numbered from one to ten, 01, 02, . . . , 10 should be used.

**Describe naming:** describe how files are named within a README file.

**Be consistent!**

---

*Example*

`ace_meteorology_data_20170130-120000.csv`

- `ace` is the overarching project - the acronym should be described in the README file, which should be stored with the files
- `meteorology` is the sub-project
- `data` shows that this is data rather than documentation
- `20170130-120000` is the first timestamp of data in the file (the meaning of a date/time should always be clarified in the README file)
- `.csv` is the file type (comma-separated values)

---

*Example*

`ace_meteorology_processedWindData_201701.csv`

- `ace` is the overarching project - the acronym should be described in the README file, which should be stored with the files
- `meteorology` is the sub-project
- `processedWindData` is information about the data contained in the file
- `201701` is the subset of data in the file (data from January 2017. The meaning of a date/time should always be clarified in the README file)
- `.csv` is the file type (comma-separated values)

---

# Chapter 5

# File formats

The format of data and accompanying documentation files will directly impact how this information can be used in the future (Stanford University Libraries, no date).

Using open data file formats helps to ensure the longevity of datasets (Borer *et al.*, 2009). Open file formats are well-documented, easy-to-read by a variety of software and are more future-proof. Using file formats that are closed and specific to a certain piece of software (proprietary), have a higher probability of becoming unreadable in the future. As software versions change, they are not always backwards compatible, meaning that a file produced ten years ago may no longer be readable. Trends also change and now-common software applications may not be widely used in the future.

If converting data from a proprietary format to an open format, ensure that no data or meaningful information is lost. It is always important to keep both copies of files and thoroughly document the proprietary software needed to create and read the proprietary files (name, version, operating system (**???**)).

## 5.1 How to choose a file format

Depending on the type of files under consideration, there will be many choices of file format to use, but for the reasons described above, file formats should be (**???**; The libraries of the Massachusetts Institute of Technology, no date):

- an open and documented standard
- commonly used by the research community (where possible)
- unencrypted
- uncompressed
- use standard representation (ASCII, Unicode).

### 5.1.1 For tabular data

- CSV files are an easy solution if minimal metadata is contained within the file. Otherwise HDF5 is a good option.
- NetCDF is a common choice in some domains, such as for climate and some oceanography data.
- Consider if a database is more suitable for working with data.

### 5.1.2 For other data types

For other types of data (such as media or geospatial) EPFL's Research Data Management Fast Guide (Blumer *et al.*, 2019) provides a number of suggestions.

### 5.1.3 For documentation

Consider a combination of the following for recording documentation about data:

- plain text (TXT) files are simple and easily read by humans;
- for formatted information, consider PDF, LaTeX, reStructuredText or Markdown;
- tools such as Frictionless Data are very useful for creating machine-readable metadata about your datasets.

### 5.1.4 For metadata

As for the documentation section, a plain text file is useful for written metadata because this can be read by different software and is more future-proof that proprietary file formats.

Where the metadata is structured in some way, for example, sampling locations and times, consider if this would be better stored in a database. Particularly where queries of this information will be needed and there are hundreds or thousands of records, it is likely a database would be more suitable. Advantages of using a database are, i) queries can be used to subset the information, for example, to find a list of samples that contain a certain type of material; ii) the metadata can be constrained on entry by permitting only valid data, leading to cleaner and more correct information on entry; iii) metadata can be combined with other data in an efficient way. One disadvantage is that more specialist knowledge may be needed to set the database up and maintain it.

# Chapter 6

# Data backup

Ensuring you have several reliable copies of data during and after fieldwork, helps to avoids data loss.

Set-up and schedules will differ when working in the field or at an institution. This is covered in more detail in the planning and working in the field sections of the guide.

## 6.1 Creating a backup schedule

### 6.1.1 How many and how often

Think carefully about how often data should be backed up and if full or partial backups for files that have changed, are needed. Documentation, code, plots and other associated files should also be backed up alongside the data.

Automating backups will make everything much simpler and helps to avoid mistakes.

Arguably it is much simpler to do a full backup of files each time and retain these for a certain period of time. If data and associated files are not being worked on any longer, then as long as the backups are secure and regularly checked, these could be backed up less often.

If however files are being worked on on a daily basis, daily backups should be done. In this case, a backup of a subset of the files may be considered.

### 6.1.2 Size of backup

It is important to consider how much space each backup will take and therefore how much total space is required for all backups. Data volume and the number of files will both affect how long the backup takes to complete.

### 6.1.3 Retention of backups

If using managed data storage, be sure to understand how often backups are done and for how long these are retained.

For regular backups, a backup cycle should be considered, where each backup is retained for a certain period of time before being deleted. For example, if ten backups of files are retained and these backups are done once per week, when a new backup is done (and verified), the oldest would then be deleted. Cycling of backups should be automated.

The long-term preservation of data should always be planned in a data management plan at the outset of a project (sometimes this is required as part of a project proposal). When planning this long-term storage, take into

account how the data have been collected, if they have been published openly anywhere and their importance for future work.

## 6.2 Documenting backups

It is important to document where the files have been backed up, when, how often, as well as how they can be accessed and restored if necessary. Whilst restoring a backup should be done on a regular basis to ensure it is still available, there are times (such as after a long field season) that this information will not be as fresh, so good documentation is vital. A short README file with this information should allow anyone to be able to restore backed-up data straight away.

## 6.3 Verifying a backup

It is important to verify a backup when it has finished to ensure it contains the files as expected. This verification should check not only the number of files and that the correct files are present in the backup, but also that the files have not been changed in any way.

### 6.3.1 Checksums

A Checksums is like a fingerprint of a file: if its content changes in any way, then its checksum also changes. Comparing checksums of the original and backed-up files is one method to verify the backed-up files are as expected.

md5sum and sha1sum are two examples of computer programs that compute checksums of files and would be suitable for this purpose.

## 6.4 Backup restoration

There can be many possible ways in which file backups are lost: storage media become obsolete, file permissions and access can be changed accidentally, and subscriptions to services are sometimes not renewed. Regularly checking that files can still be read and restored is important to ensure that there are no problems.

Choosing a simple, well-understood, transparent and multi-platform tool to do backups will often make file restoration much simpler, but this should be chosen with care.

If any data were produced using proprietary software it is particularly important to ensure that they can be read on a regular basis. Consider outputting data into an open, documented format such as TXT or CSV - be aware that in this process some information or data may be lost, so it is always good practice to keep both sets of files.

### 6.4.1 Backup tools

Many tools are available to help create backups. It is worth spending time finding one, taking into account future and current needs. It is essential to fully understand how the data are being saved and how they can be restored.

Using a multi-platform tool (such as for Windows, Mac, Linux and other operating system users) offers higher resilience, ensuring more possibilities for accessing the data in the future. Some institutions may have tools or recommendations that can help.

## 6.5 Summary of key points

Make sure:

- at least two, preferably three or more copies of data are kept;
- data are backed up on at least two different types of media, particularly for preservation purposes, such as institution storage, cloud storage, external hard drives;
- as far as possible, backups are automated. This avoids potential mistakes, minimises the chances of data loss, makes it much easier to do (it is less of a chore) and ensures the backups are always done in the same way;
- data are backed up on a regular basis, but particularly during collection and after making any changes;
- backed up versions of data are identical to the primary copy (i.e. verified). Whilst checking that files have been copied, even if they are listed in the secondary location, using checksums will confirm they have been copied correctly;
- backups can be easily restored;
- decide on a directory structure and file naming convention and stick to it. Making changes to these (unless absolutely necessary) can create problems with backups because it is easy to lose track of what has been copied and what has not, which is the latest version, and so on;
- bear in mind how long a backup will take and consider this when deciding how backups will be scheduled.

# Chapter 7

# Working on your data

In this chapter we focus on ensuring data security (not losing data), integrity (not changing raw data files) and capturing the provenance of datasets whilst working on data.

Careful recording of how a dataset has been produced is important so that someone else can understand what has been done. It is also a useful reminder for when it comes to writing up work for publication and for when getting back to work after a long field season away. Making this information clearly available alongside the dataset would allow someone else to reproduce the work. Reproducibility is becoming ever more important to ensure scientific validity (Peng, 2011).

## 7.1 Raw data

The original copy of the data, known as "raw" data, is that which comes directly from the instrument, sample analysis, or primary observations. This raw data should be saved as it is, backed up and **should never be modified**. The primary copy of the raw data should have read-only access so it can never be altered inadvertently. Never reorganise or alter the primary copy of raw data files.

## 7.2 Work on a copy of raw data

In preparation for working on the data, performing quality checking, applying calibrations, or indeed making any change whatsoever to the data, make a copy of the raw data and ensure all work is done on the copy. Ensuring a "pristine" version of the raw data is maintained is imperative: throughout the analysis process it is possible that inconsistencies and errors will be discovered, that may mean going back to the raw data files in order to find their origins. Any changes that have subsequently been made may mean that it is then impossible to track down the origins of these problems.

## 7.3 Versions of files

When working on data processing, quality checking and making other changes to produce the preliminary dataset, creating different versions of the files after each step rather than overwriting files, is a valuable way of ensuring all steps can be followed and help to get back to a certain point to investigate errors or make other changes.

File versions can be named using the date (in the format YYYYMMDD) or version numbers, such as `v01_01`, `v01_02`. Including the date in this format, or version numbers that have leading zeros, ensures that files are listed in order when viewing them. Alternatively, versioning tools should be used.

Backups of any edited files should be done on a regular basis.

## 7.4 Recording the provenance of data

It is often impossible to repeat data collection, which in polar research is also extremely expensive. This "original" data is considered as raw and should be kept as such. But any work that is done on a copy of this raw data should be carefully recorded. It is natural to keep notes of what has been done throughout the process of collecting and working on data so that when writing up publications, methods can be written more easily.

Additionally, journals often ask that all supporting documentation, code, data and information about how plots and figures were generated, be available. Spending time organising how to capture the full provenance of datasets and research paper will save a lot of time when ready to publish.

Using scripted languages or tools that assist with transforming and recording the transformation of data to do any file manipulation is the ideal way to record what has been done to data files and demonstrate how a dataset has been created in a reproducible manner (Borer *et al.*, 2009). Scripts can be modified easily if a mistake is spotted and then re-run, rather than having to run through a set of manual steps again. Steps can easily get lost, forgotten or critical errors introduced without noticing, if using spreadsheets or making direct changes (eg. Ziemann, Eren and El-Osta (2016)). Alternatively, detailed written notes alongside a flow chart outlining the steps can be a good start. Whichever method is used, maintaining full documentation about every step of the processing and outputting different versions of the dataset at crucial stages is a way to demonstrate the provenance of a dataset.

Refer to the metadata section for full details of what information should be captured to properly describe the provenance of your data and samples.

## 7.5 Tools

### 7.5.1 Versioning tools

Versioning tools capture any changes that are made between files, allowing a user to go back to and look at differences between previous versions. There are many different tools available but we mention two of the most commonly used ones:

- Git - commonly used for software versioning
- Git-LFS - used for data file and software versioning

Github, which is based on Git, is commonly used as a platform for software versioning.

### 7.5.2 Tools for capturing data provenance

Versioning tools are very handy for making incremental improvements to code and even versions of data or documentation. Some tools also exist which record what has happened to a dataset, in other words, recording its provenance.

- RENKU: https://datascience.ch/renku/
- Whole Tale: https://wholetale.org/

## 7.6 Summary of the main points

- Never modify the raw data files.

- Wherever possible, use a scripting language or data transformation tool that records processes, to do any manipulation of data, for applying algorithms, quality-checking and any other processes that work towards the final, output dataset.
- Back up and keep different versions of any code so that it is possible to see where errors are introduced into data or processing.
- Maintain clear documentation: clearly state references to algorithms, which software and packages (including the version) as well as which the decisions taken and why.
- Record the set-up of the computing environment that has been used to run the scripts: include details of the operating system, package names and versions.
- Always keep data and code that produces plots: this could not only save time if a mistake needs correcting but it could also be required for publication in a journal.

# Chapter 8

# Collecting samples

Obtaining accurate results from an experiment when analysing samples, depends on "proper collection, processing and handling of samples" (Smith *et al.*, 2015), planning for which begins before sample collection.

Our main aims when considering good practice for collection of samples are:

- accurate labelling of samples;
- quickly identifying or finding a sample;
- matching a sample to any documentation or metadata about how it was collected.

## 8.1 Labelling samples

Carefully planned sample labels will make it much easier to meet our aims above. Keep labels as simple as possible to avoid mistakes and confusion.

### 8.1.1 Requirements from other partners

If samples belong to a larger project, find out about any wider project requirements that should be met. All projects will need to ensure that sample labels are unique, therefore it might be necessary to add a specific prefix or suffix to differentiate samples from those of another part of the project.

The laboratory or organisation where the samples are going to be stored and / or analysed may also have their own requirements for labelling.

Awareness of these requirements in advance will hopefully avoid last-minute amendments not only to the labels written on the samples themselves, but documentation and metadata as well.

### 8.1.2 Unique identification labels for each sample

Each sample should have a unique identification label written on it: avoid a short version of a longer sample label that could easily be confused if more than one sample from a field campaign (or the wider project) end up with the same label.

This unique identification label will then "link" that physical sample to the information that is recorded about its collection (metadata).

### 8.1.3 Duplicate samples

In some experiments, more than one sample will be taken at the same time and under the same conditions in order to corroborate results. These samples may be known as "duplicates" (or "triplicates" in the case of three, and so on). It is important to be able to distinguish between these samples and each should still have a unique label.

### 8.1.4 Further considerations

- Space on the sample container may be limited.
- Be very aware of mistakes that can be made by labelling in advance. This could be if a particular location ID is going to be used in the sample label, for example, and it may lead to samples being held in an incorrectly labelled container.
- Bar coding systems can be used to record and label samples. There are various costs associated with setting up and maintaining such a system, but it may be worth consideration and the investment if large numbers of samples are collected (Copp, Kennedy and Muehlbauer, 2014).

## 8.2 Labelling

Label containers directly using permanent waterproof marker pens (or suitable alternative depending on the storage conditions), or on an adhesive label where necessary. If working with alcohol for sample preservation, use pencil to label samples. Never solely label the cap of a container: the body of the container should always be the primary label (Smith *et al.*, 2015).

For very small containers, it might be necessary to place the label on a small piece of appropriate material (to ensure neither the sample nor the label are damaged) inside the container.

Carefully consider how samples are going to be stored and transported to make sure they will not come into contact with any chemicals or conditions that may damage the labelling. If in doubt, label the sample multiple times on the outside and inside of the container.

## 8.3 Recording information about sample collection

Without information about how, where and when samples are collected, stored, processed and curated, it is not possible to interpret results correctly. Recording this information, known as metadata, is essential and should be considered as important as the samples themselves. Often, a particular sample is linked to its metadata and the final data which arises from it, using its label.

Recording sample contents (and often treatments) whilst in the field is essential: this information may be required by permitting authorities and border control when transporting samples for analysis.

# Part II

# Field guide

# Chapter 9

# Introduction

This section provides a guide to good data management whilst in the field. It is split into three distinct sections: planning beforehand, data management in the field, and things to consider on return.

Within in each section, we consider whether data are being collected from an instrument or by hand, and the collection of samples.

The field guide often refers to the first part of this document for information or guidance on a particular topic. We assume the reader is familiar with these earlier sections and recommend returning to the relevant sections to have a full grasp of the context.

# Chapter 10

# Planning

In polar and high-altitude environments, data and sample collection can be complicated by difficult-to-reach instrumentation, harsh weather conditions and remote field sites. Careful planning helps to minimise the risk of data and sample loss when it is hard to troubleshoot problems in these environments.

The main points for planning are the same, whether collecting samples, data by hand, using an instrument to gather data in an automated manner, or a mixture of all of these:

- find out **what is available** at the field work location;
- carefully plan how saved data and recording of samples will be **organised**;
- plan for enough **data storage** or notebooks to record data;
- ensure data can be **backed-up** in the field;
- plan which **metadata** will be recorded about the data or sample collection;
- understand what is needed to get the data and samples home safely.

## 10.1 What is available on-site?

Resources and support that are and are not available at the fieldwork site can impact on how data are saved, backed-up and accessed whilst on site. It is useful to get advice from others that have previously visited the site if this is possible, but always double-check details in case changes have been made since previous years.

### 10.1.1 Network

A connection to the local IT network can be useful for automated, secure backups (if there is also storage), collaborating more easily with others, viewing data without having to access the instrument (very useful in bad weather or if the instrument's location makes it tricky to access) and depending on the site, might mean data can be backed-up to a remote location.

**Questions to ask**:

- Do you have access to a local network? If so, is this from a computer based at the site, or can you connect your own computer?
- Can your instrument be connected to the network? Is there network access where the instrument will be based, or do you need to think about taking longer network cables? Is it too far away from a network point?
- What is the speed of the network?
- Are network cables provided?

- Is there someone who can assist in connecting your computer or instrumentation to the network, or do you need to know how to do this?

### 10.1.2 Data storage

Network-attached storage may be provided and therefore would offer another means of managed data backup. It is likely to be more secure than a portable hard disk if the infrastructure is well-managed.

**Questions to ask**:

- Are you able to backup data to the network-attached storage? If so, what capacity would be available to you?
- Do you have read-access to look at your data?
- Is access to read data available any time? Is this possible through your own computer or is it through a computer based on the site?
- Is there someone who can assist in setting up your access to the data storage, or do you need to know how to do this?

### 10.1.3 Power and electricity supply

In addition to the power supply needed for instrumentation, consider what is available to power extra data storage and associated UPSs (uninterrupted power supply; keeps instrumentation running using battery power for a short period if the main supply fails).

### 10.1.4 Support and restrictions

Some camps or bases might be able to provide some IT support or have strict rules about connecting computers to local networks. It helps to understand these limitations properly before you arrive so planning can be done accordingly.

## 10.2 Preparing for data collection from an instrument

Whilst planning, think carefully about the following questions:

- How much data are you planning to collect?
- Do you have enough data storage for the planned data collection, plus extra for unforeseen circumstances?
- How will you organise the data files?
- How will the data be backed-up in the field?
- Will you be able to access the data during collection?
- What do you need to know about how your data were collected?

### 10.2.1 How much data are you planning to collect?

It is important to have a good understanding of any instrumentation, associated software and how the files are saved.

**Which data do I need to save?**

Some software automatically writes “false” data from variables which are not actually being recorded or are not of interest to the particular experiment. These should not be recorded in the data file because they can cause confusion for others looking at the data files in the future. Check how to “select” which variables are saved into the data files.

If data storage is limited and 24-hour recording is not possible, consider which periods of time are more crucial (i.e. data collection only at night, or for five minutes every hour, one-minute resolution instead of one-second). All of this should be considered with your experiment in mind to ensure it is not compromised. Do not forget to take into account local sunrise and sunset times if this is an important factor, particularly if the instrument is going to be installed on a moving platform such as a ship.

**Do I understand how big each data file will be?**

File sizes vary depending on the configuration of the instrument, the number of variables recorded, and importantly, the data themselves (i.e. more background noise can produce higher figures and therefore more bytes). Ensure the instrument is configured for the particular experiment. A proper test-run with the correct settings, will allow all of these points to be checked thoroughly before going into the field. Verify that the size of the test files is as expected.

**How do I calculate how much data I am going to collect?**

Following the verification of expected file size after a test-run, calculate how many data files will be produced (one per hour, one per day?) and use this to calculate an estimate of daily storage that will be produced. Multiply that calculation by the number of days of data collection.

Always round estimates up. It is better to overestimate.

---

*Example*

An instrument produces an 8 MB file each day. Data will be recorded for a planned field season of 8 weeks (56 days), but there is a potential that it could be for up to 16 weeks (112 days) if it is not possible to make it back to the instrument on the first occasion.

8 weeks: $8 * 56 = 448MB$

16 weeks: $16 * 56 = 896MB$

On the second instrument, six different files are produced when a certain event occurs. When the event occurs, data are recorded for a period of 36 hours and each file contains an hours worth of data. That means $6 * 36 = 216$ files are produced during an event. Each file can contain around 600 MB of data.

For an event: $216 * 600 = 129600MB = 129600/1024GB = 126.56GB$

During a period of 8 weeks, we expect an event to take place once a week, but to be certain about data storage, we plan for two events a week. This means $8 * 2 = 16$ events, or $16 * 2 = 32$ events if the instruments are left for the full 16-week period.

8 weeks: $126.56 * 16 = 2025GB = 2025/1024TB = 1.98TB$

16 weeks: $126.56 * 32 = 4050GB = 4050/1024TB = 3.96TB$

---

**Do I have enough data storage for the planned data collection, plus extra for unforeseen circumstances?**

Where possible, do some initial set-up tests in the field before or during deployment but remember that this will use up data storage. This is important to do, so budget space accordingly. Pre-departure testing should give a good idea of how much data to expect.

In the event of bad weather and not being able to access the instrument, or other unforeseen circumstances such as the field season being extended, data may be collected over a longer period of time. Don't miss out on

the opportunity for additional or opportunistic data collection if it becomes available, just because there is not enough data storage is available!

Primary storage, that is where the raw copy of the data will be saved initially, should be of a volume that more than covers planned data collection. Always ensure a buffer of at least 20 %, preferably more, and test how the files are stored thoroughly beforehand. If in doubt, have more storage rather than less.

NOTE: in the example above, a 5TB hard drive would probably be a good choice in case it is not possible to get to the instrument to change it. Additional 5TB disks would be needed for backups.

**Metadata and documentation**

Ensure enough data storage has been budgeted for these important aspects of data collection as well. This could include spreadsheets containing notes and supplementary data, photographs of experiment setups, digitised hand-written notes or anything else that could be useful. Photographs and video could be particularly large in terms of data storage and backups, so bear this in mind.

### 10.2.2 Data storage media

Carefully consider the hardware on which data will be stored, ensuring that it can withstand the conditions of the field site.

Many permanent field sites have some kind of provision for computing but details should be checked carefully:

Data storage might or might not be available, but portable media (or a good internet connection) will still be needed to take the data home and do backups.

If instrumentation and data storage are connected to an electrical supply, consider an external power supply such as a UPS, to keep them running in case of power loss. In this case, find out details of the electricity supply to ensure compatibility.

### 10.2.3 Organising data files

Think carefully about the directory structure and filenames that are used, particularly if you are collecting data automatically. Refer to the relevant sections of this guide. Where possible, set up these details beforehand.

### 10.2.4 Backing up data in the field

Always ensure data and metadata can be backed-up whilst in the field. Plan carefully to make sure these backups are automated as far as possible, making it much less of a chore and harder to make a mistake. Test out each method of backup carefully before leaving to ensure the method and the hardware (if applicable) work properly. Don't forget to verify and test recovery of backups as well.

Depending on the circumstances, backups could be:

- a number of portable hard drives, held by different members of the team or stored in different locations and rotated;
- if others are coming and going to a field site during the season, consider asking a responsible person to carry a copy of the data back to the institution. This provides a copy in a different location and means it could also be placed on secure networked storage away from the field site as an extra precaution. It would be particularly useful if conditions at the field site make it difficult to keep portable media safe and in good condition;
- on network-attached storage if accessible;
- by sending files using a mobile or satellite connection from the instrument (this will depend on situation and cost) to cloud storage.

Using on-site options such as network-attached storage or sending files via the Internet are really bonus options, so always have a backup plan in place, in case this doesn't work out.

### 10.2.5 Accessing data in the field

Being able to access data in the field during collection is extremely useful and cannot be underestimated. In particular it allows:

- checks that the instrument is running as expected;
- confirmation that data files are being saved as expected;
- observation of interesting features in the data that might indicate problems. Knowing about them in near-real time can be hugely advantageous when quality-checking and processing data.

Setting up quick visualisations of data files saves a lot of time and provides a lot of information with a quick glance.

If the instrumentation is nearby and easy to access, checking the data files periodically helps to spot obvious issues with the instrument early on, and ensures data are being saved as expected (check parameters, file format, frequency of records).

If instrumentation is going to be left for a period of time, it is worth considering what access will be possible. Running initial tests whilst still with the instrument in the field is essential. Once data collection has been confirmed, then it is still useful to be able to access data periodically. This could be across a network (for example if on a ship or at a base) or even using mobile or satellite communications if the instrument is isolated. Even if only a subset of the data can be received, a small daily file with a subset of the data could be enough to check that everything is going well or flag up issues. Of course, this is of more use if someone is then able to go and fix the problem.

For instruments that are very isolated and there is no possibility of being able to access them, consider if there is a way to remotely connect to the instrument. This might offer the possibility to restart it, for example, or change crucial settings. Set this up and test it thoroughly beforehand.

## 10.3 Preparing for recording data by hand

If recording data by hand directly into a notebook, think carefully what information should be recorded. Keep separate notes (documentation and metadata) about how measurements will be recorded, units and any parameter abbreviations that will be used in the field.

### 10.3.1 Data backup and digitisation

Scan, photograph or type-up hand-collected data and notes as soon as possible in the field as a form of backup but also to make it easier to check data collection.

If using a spreadsheet to transcribe hand-recorded notes, prepare the file template in advance and have a test-run of data collection and digitising of data. Where possible and appropriate, use drop-down lists of specified terms within the spreadsheet (example for Excel) to keep data entry consistent. This will allow quick data validation and save a lot of time cleaning data. As when collecting data from an instrument, early digitisation also offers the opportunity to produce some quick visualisations or numerical checks of data.

If it will not be possible to use a laptop in the field, then consider other methods such as taking a camera and carefully photographing hand-written notes and data. Photographs or scans of hand-written notes are very useful to verify data entry, as well as being a vital backup.

Even though data may not being collected directly from an instrument, calculate a data storage budget for the digitised data, not forgetting separate backups.

## 10.4 Preparing for sample collection

The following should be considered when planning sample collection in the field:

**Labelling scheme:** the scheme should be meaningful and well-described.

**Plan metadata collection:** decide what information is needed to fully document the sampling. Create templates of data collection sheets and spreadsheets for transcribing the metadata. Plan how additional documentation such as notes and photographs will be recorded.

**Plan storage backups of metadata:** metadata files will need to be stored and backed-up securely whilst in the field (calculate data storage budget).

**Check permit requirements:** permit authorities may insist on certain information being kept about the samples and it is likely this will need to be reported. Be aware of this beforehand to avoid any doubt in the field or on the return journey and plan to record it as part of regular metadata entry.

**Check customs, border crossing and entry requirements for the transport of samples:** it is likely a list of samples with information about what they contain, will need to be presented. Carefully check and understand exactly what is required before travelling to the field site where it might not be able to access this information. Plan to record this alongside other metadata on a regular basis.

## 10.5 Metadata

Prepare files or a notebook template to record metadata before leaving to avoid forgetting anything. Do not be afraid to add extra information though whilst in the field.

### 10.5.1 Key points: "where"

**Sample and data collection location recording:** if the instrument will always be in the same place, a hand-held GPS can be used to record its location. If collecting data or samples in several distinct locations, a hand-held GPS will also be useful to do this. If the platform is moving, for example a ship or an aircraft, it is important to have more than one device that is constantly recording the location of the platform.

**Understand the position of the instrument in relation to the geolocation device:** particularly on a large ship or in relation to a specific location, it might be important to know exactly how far away the instrument or collection site is, if very accurate positions are required.

**More than one source of location data:** as with any other instrument, navigation and geolocation devices can fail. At least two sources of this data is imperative to ensure that the position of the data or sample collection can be recorded. This is even more important on moving platforms such as a ship or a plane, where a short period of no data can mean there are large gaps in location information.

**Location accuracy:** understand how accurate the location needs to be, to be meaningful for the experiment. Be aware that local conditions such as mountainous or tree-covered terrain can affect the accuracy of this data.

**Source and accuracy of location data:** record all metadata from the navigation device to have information about the accuracy of the position (position to nearest x metres, number of satellites; device name, type, manufacturer, version and serial number; see instrumentation section below for details).

### 10.5.2 Key points: "when"

**UTC:** it is good practice to record all scientific work in UTC to avoid confusion with time zones.

**Set the time:** ensure the timing device is set accurately. Record if the time is recorded to the nearest day, hour, minute or second. If using networked computers or instruments, ensure that they are all synced. Watches should also be set to the same time.

**Time source:** if working as part of a larger project, particularly where data and sample collection is simultaneous among teams, ensure everyone is working from the same time source. Note details about the source of time (device name, type, manufacturer, version and serial number).

**Time offsets:** find out if there is any offset between the instruments / devices recording location, time and the data itself because if working on a moving platform, it is likely that data points will need to be "matched" to the location using time.

### 10.5.3 Instrumentation and computers

Keep a detailed record of instrumentation that is used for primary and secondary datasets, sample collection and saving / backing up data. If at all possible, record this information in advance before going to the field and remember to take it into the field.

For **instrumentation**, record:

- name
- type
- manufacturer
- version / model
- serial number

Good documentation of instrumentation is important, particularly in case of issues. Keep the details to hand in the field and with anyone who is able to offer support back at the institution: this makes it much easier to contact the manufacturer for support whilst in the field, or afterwards when trying to solve problems.

During an expedition, it is possible that part of the instrumentation might change: for instance, a new sensor could be added, or swapped if one fails. Always record the following details about **sensors**:

- date installed
- date removed
- details of problems (if there were any)
- location of installation on the platform or on parent instrument (e.g. side of ship, height on mast, which part of CTD rosette)
- data files to which each instrument corresponds.

For **computers and software**, record:

- operating system name and version
- software name and version, including any packages
- any particular set-up and computing environment.

Prepare file templates (spreadsheets or plain text files are ideal) for collecting this information before going into the field so nothing is forgotten.

## 10.6 Documentation

Much of the specific metadata to be recorded has been described in detail above. Carefully prepare template documents that can be taken and completed in the field.

For further documentation about methods, problems encountered and anything else, this can be a simple README file with headers as reminders of information should be recorded.

Notes, diagrams, photographs and any other forms of documentation should be backed up alongside your data.

## 10.7 Travel and customs

Plan how to take data storage (and related) hardware to and from the field location with good time. Carefully check cargo restrictions to ensure that hardware meets the requirements of carriage. Batteries that are in UPSs or other instrumentation are particularly regulated. Import regulations should also be carefully checked. Finally, do not forget to consider the regulations of countries that will be transited, as well as modes of transport.

When returning with portable media devices holding carefully-collected data, think about how they will be transported. Consider encryption of the device if it holds personal or sensitive data and if someone else is travelling back to the same location, ask them to carry a separate copy of the data.

As previously discussed in the section about preparation for sample collection, it is important to understand the requirements of customs and other permits that are required so this information (metadata) can be recorded whilst in the field.

# Chapter 11

# In the field

Careful preparation and a good routine in the field can help to manage data well and avoid frustration through lack of documentation and bad organisation of files. A careful routine backing up data, transcribing hand-written notes or data, and checking records of samples at the end of each day, is vital.

## 11.1 Data collection from an instrument

Automation, careful setup and testing are key. Much of the setup for data storage should already be planned and in place before beginning work in the field. Refer to the planning section for what to prepare before arriving in the field.

### 11.1.1 Initial setup and testing

Retaining data and keeping notes (metadata) about initial setup and tests in the field is essential in case any problems are noticed later on.

Firstly, check that data files can be read. If files are saved in proprietary formats, then make sure the software that is required to read them, is available.

Check that data files contain data for the expected parameters in the correct units. Ensure the files are being saved in the directory structure and with the filenames that are expected. Now is the time to make any changes, before "real" data collection begins. Test doing a backup and recovering (opening and checking) the files.

### 11.1.2 Periodic checking of data

If it is possible to access the data whilst in the field, check that files are being saved correctly and that they can be read. Consult the software documentation for how is best to do this, to avoid problems. Make a copy of the primary data files to avoid interfering with the file-writing process; check copies of files rather than the files that are being written.

For data that is being recorded continuously, it is good practice to check files at least once a day, or more often if time allows. Events such as bad weather may mean data need to be checked more regularly.

If it is possible to visualise the data, this helps to spot anomalies which may in turn help to spot an instrument problem which could subsequently be fixed and get the data collection back on track.

Do not change parameters that are recorded as part of a dataset, part way through a field campaign unless a parameter has been forgotten. This adds much unneeded complexity and confusion when trying to read the data files during the post-processing stages. In particular, do not change the format of fields or parameter names

during data collection. Testing data collection before starting data collection in the field should help to prevent this kind of problem. Any changes that are made should be clearly documented, including details of which files are affected.

### 11.1.3 Backups

Backups should be automated wherever possible. If this is not possible, make sure they are done on at least a daily basis and are stored in more than one place. At least two backups of the data should be kept, preferably more.

Check the integrity of the backups on a regular basis: make sure the files that are backed up are the same as the originals and that they can be read correctly.

## 11.2 Recording data by hand

If collecting data in a notebook by hand, it is good practice to digitise the data as soon as possible. Digitised data files, metadata and documentation should be considered as valuable as any other data file: consider the file name, directory structure and backups.

### 11.2.1 Backing up or digitising your data

Hand-written data should be digitised as soon as possible after data collection and at least once a day. A simple first backup can be done by photographing notes in case the unthinkable happens and the notebook is lost (keep the camera and notebook separately). Digitising hand-written notes in a structured manner such as in a spreadsheet, cannot be underestimated and this should be done as regularly as possible (at least daily). This allows embellishment of any shorthand that has been used before it is forgotten, or any queries to be followed-up. In particular, when recording species, it is possible to follow up on missing identifications before key details are forgotten. See the previous chapter about what to prepare beforehand to save time in the field.

If adding to a data file with new observations each day, using the date to denote different versions of the file is helpful e.g., `project_datatype_observations_YYYYMMDD.csv`. This acts as an additional backup: if the last day's file is lost, at least only the latest data has been lost.

Do not forget to backup digitised data as well.

## 11.3 Sample collection

Record metadata about sample collection as accurately and as soon as possible. Cross-checking this information is useful to spot mistakes.

Ensure samples can be easily found at the field site from the records that are kept, particularly if participating in a long field campaign. Recording their storage location and box / crate number is also useful in case they need to be found, such as on arrival at customs. Be sure to note if they are destroyed during sample analysis that is done whilst in the field, to avoid wasting time searching for a sample that no longer exists!

Finally, keep an accurate record of what is contained in each sample according to what is required for permits, border entry requirements and where the sample is being sent for analysis following field work.

## 11.4 Data from sample analysis

In some cases, samples may be collected and processed in the field. The field setup will very much dictate how data from sample analysis can be recorded. Consider if it will be possible to have a laptop or computer in the lab,

or if values will be noted by hand. If the former, then treat this as data collection from an instrument, taking good care to regularly save and backup data files. If the latter, then refer to the section recording data by hand. This data should be digitised as soon as possible and backed up securely.

If recording further information about the samples such as identifying different species of plant, then refer to the section, recording data by hand.

## 11.5 Recording metadata

Ensure the metadata section of this guide is followed, paying particular attention to recording where and when samples and data were recorded. Metadata is most accurate when recorded as soon as possible - don't leave it until the end of the field campaign to make notes.

Remember to digitise this information if it would ordinarily be recorded in a notebook and treat these files like any other data, using secure storage and regular backups.

Take photographs of as many things as possible! The instrument set-up, makes, models and serial numbers of equipment equipment, notebooks, field sites, laboratory setting, notebooks and anything that could affect the experiment directly or indirectly. This is hugely valuable metadata and should be carefully backed-up.

## 11.6 Verifying metadata

We have already discussed checking data whilst you are in the field to ensure that instrumentation or measurements are going to plan. It is just as important to make sure that metadata are correct.

**Plot coordinates on a map:** prepare and take a very simple map background which allows you to see the context of your study area. Plot any data or sample collection points, routes or important locations on the map as a way to verify that recorded coordinates are correct. This could help to spot errors in navigation devices (particularly if they are recording constantly), errors in transcribing positions or simple confusion of latitude and longitude.

**Bounding box for coordinates:** verification of latitudes and longitudes can also be done on data entry, by restricting them to a bounding box in which you are working. Identifying the minimum and maximum permitted latitude and longitude can help to have a quick check in a recording sheet.

**Check dates and times:** depending on your field site, you may be working in different time zones, recording information in UTC or even changing time zone regularly. This all adds to confusion in the field. Always record time zone changes and keep a record of how many hours ahead or behind UTC you are. It is worth regularly checking dates and times of data and sample collection in your records. Do they make sense? Do they follow on from each other in the correct way? Were you in the field at that time? Plots can help, as can simple addition and subtraction of times in a spreadsheet.

**Sets of possible values:** for many parameters that are recorded by hand or as metadata, it is often possible to restrict them to a set of possible values. For example, it may be that you are collecting on three types of material: water, sediment and vegetation. Or perhaps the temperature is being recorded each time you collect a sample; given the location and time of year, you know that it would be very unlikely that the temperature would be outside of the range of -5 to +15 degrees C. In common spreadsheet software it is possible to restrict values entered into a column. Otherwise, simple checks or plots can be done on a regular basis to make sure these limits are not violated. Be strict with yourself; this will save a lot of time later on.

Whilst it is very common to record this kind of metadata in a spreadsheet, investment in putting together a simple database which includes parameter value constraints as mentioned above, can really be worth it. Resulting data will be cleaner and contain less mistakes which are easily made at the point of data entry.

# Chapter 12

# Upon return

When returning from the field, data management **priorities** are to:

- backup data (data, metadata and documentation) to networked (or other suitable) storage: ensure the primary copy is read-only;
- ensure safe storage of samples;
- ensure safe storage of hand-written notes. Some institutions may have an archiving service in which these could be placed.

**Further considerations** will then be to:

- check metadata and documentation and add any extra information whilst it is still fresh;
- document clearly where data are backed up and how they can be accessed;
- organise a **copy** of the raw data where it can be worked on.

Follow the guide to working on data to follow best practice.

# Chapter 13

# Document revisions

## 13.1 Version 1.0

- First version

## 13.2 Version 1.1

- Update title page: date, version and DOI
- Update introduction: what this guide covers, contact email address
- Update metadata: refine text, add README section, improve example
- Update storing data: add section “data publication” and “set them up” within portable media, refine text, reformat
- Update file organisation: move README section, refine text
- Update file formats: add section “for metadata”, refine text
- Update backing up data: refine text
- Update sample collection: refine text
- Update field guide introduction: refine text
- Update planning before you go: refine text, add example data storage calculation, reformat
- Update in the field: add section “verifying metadata”, refine text
- Update upon return: refine text

Git commit: ab51ba7

# Bibliography

Blumer, E.N., Chaptinel, J.J., Masson, A., Reichler, F. and Samath, S. (2019) 'EPFL Library Research Data Management Fastguides'. Zenodo. Available at: https://doi.org/10.5281/zenodo.3327830.

Borer, E.T., Seabloom, E.W., Jones, M.B. and Schildhauer, M. (2009) 'Some Simple Guidelines for Effective Data Management', *Bulletin of the Ecological Society of America*, 90(2), pp. 205–214. doi:10.1890/0012-9623-90.2.205.

Chan, A.W.L. (2020a) *S3 keys are not file paths*. alexwlchan. Available at: https://alexwlchan.net/2020/08/s3-keys-are-not-file-paths/ (Accessed: 15 September 2020).

Chan, A.W.L. (2020b) *S3 prefixes are not directories*. alexwlchan. Available at: https://alexwlchan.net/2020/08/s3-prefixes-are-not-directories/ (Accessed: 15 September 2020).

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. and McGillivray, B. (2020) 'The citation advantage of linking publications to research data', *PLOS ONE*. Edited by J.M. Wicherts, 15(4), p. e0230416. doi:10.1371/journal.pone.0230416.

Cook, R.B., Olsen, R.J., Kanciruk, P. and Hook, L.A. (2001) 'Best Practices for Preparing Ecological Data Sets to Share and Archive', *Bulletin of the Ecological Society of America*, 82(2), pp. 138–141. doi:10.1890/0012-9623(2001)082[0136:C]2.0.CO;2.

Copp, A.J., Kennedy, T.A. and Muehlbauer, J.D. (2014) 'Barcodes Are a Useful Tool for Labeling and Tracking Ecological Samples', *Bulletin of the Ecological Society of America*, 95(3), pp. 293–300. doi:10.1890/0012-9623-95.3.293.

Craig-Wood, N. (2014–2020) *Rclone*. RCLONE. Available at: https://rclone.org/ (Accessed: 19 October 2020).

DataCite Metadata Working Group (2019) 'DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.3'. DataCite. doi:10.14454/7XQ3-ZF69.

Downie, A. (2019) *Bite-sized RDM #5 - the readme file*. IT and Research Data Management at the Gurdon Institute. Available at: https://gurdoncomputing.blog/2019/12/02/bite-sized-research-data-management-5-the-readme-file (Accessed: 27 February 2020).

Fegraus, E.H., Andelman, S., Jones, M.B. and Schildhauer, M. (2005) 'Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation', *Bulletin of the Ecological Society of America*, 86(3), pp. 158–168. doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.

GitHub, Inc. (2020) *GitHub: Where the world builds software · GitHub*. Available at: https://github.com/ (Accessed: 19 October 2020).

Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J. and Heber, G. (2005) 'Scientific data management in the coming decade', *ACM SIGMOD Record*, 34(4), pp. 34–41. doi:10.1145/1107499.1107503.

Greenbelt, MD: Earth Science Data and Information System, Earth Science Projects Division, Goddard Space Flight Center (GSFC) National Aeronautics and Space Administration (NASA) (2020) *Global Change Master Directory (GCMD) Keywords | Earthdata*. Available at: https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords/ (Accessed: 19 October 2020).

McCarthy, J.L. (1982) 'Metadata management for large statistical databases', in *Proceedings of the Eighth International Conference on Very Large Data Bases. Eighth International Conference on Very Large Data Bases, September 8-10, 1982; and published in the Proceedings*, Mexico City, Mexico, pp. 234–243. Available at: https://escholarship.org/content/qt5cc031cm/qt5cc031cm.pdf.

Michener, W.K. (2005) 'Meta-information concepts for ecological data management', *Ecological Informatics*, 1(1), pp. 3–7. doi:10.1016/j.ecoinf.2005.08.004.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B. and Stafford, S.G. (1997) 'Nongeospatial metadata for the ecological sciences', *Ecological Applications*, 7(1), p. 13. doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

Peng, R.D. (2011) 'Reproducible Research in Computational Science', *Science*, 334(6060), pp. 1226–1227. doi:10.1126/science.1213847.

Recknagel, F. and Michener, W.K. (eds) (2018) *Ecological Informatics*. 3rd edition. Cham: Springer International Publishing. doi:10.1007/978-3-319-59928-1.

Smith, P.G., Morrow, R.H., Ross, D.A., Association, I.E. and Wellcome Trust (London, E. (eds) (2015) *Field trials of health interventions: A toolbox*. 3rd edition. Oxford: Oxford University Press.

Stanford University Libraries (no date) *Best practices for file formats*. Stanford Libraries. Available at: https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats (Accessed: 24 July 2020).

The libraries of the Massachusetts Institute of Technology (no date) *File formats for long-term access | Data management*. Data management. Available at: https://libraries.mit.edu/data-management/store/formats/ (Accessed: 24 July 2020).

The University of Edinburgh (2021) *Documentation, metadata, citation*. MANTRA Research Data Management Training. Available at: https://mantra.edina.ac.uk/documentation_metadata_citation/ (Accessed: 31 July 2020).

Wholetale (2019) *The Whole Tale*. Available at: https://wholetale.org/ (Accessed: 19 October 2020).

Wikipedia contributors (2020a) 'Checksum', *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia. Available at: https://en.wikipedia.org/w/index.php?title=Checksum&oldid=962858813 (Accessed: 28 July 2020).

Wikipedia contributors (2020b) 'Comparison of online backup services', *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia. Available at: https://en.wikipedia.org/w/index.php?title=Comparison_of_online_backup_services&oldid=984305279 (Accessed: 28 July 2020).

Wikipedia contributors (2020c) 'Md5sum', *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia. Available at: https://en.wikipedia.org/w/index.php?title=Md5sum&oldid=961659444 (Accessed: 15 September 2020).

Wikipedia contributors (2020d) 'Sha1sum', *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia. Available at: https://en.wikipedia.org/w/index.php?title=Sha1sum&oldid=965259914 (Accessed: 15 September 2020).

Zenodo (no date) *Zenodo - Research. Shared.* Zenodo. Available at: https://zenodo.org/ (Accessed: 15 September 2020).

Ziemann, M., Eren, Y. and El-Osta, A. (2016) 'Gene name errors are widespread in the scientific literature', *Genome Biology*, 17(1), p. 177. doi:10.1186/s13059-016-1044-7.