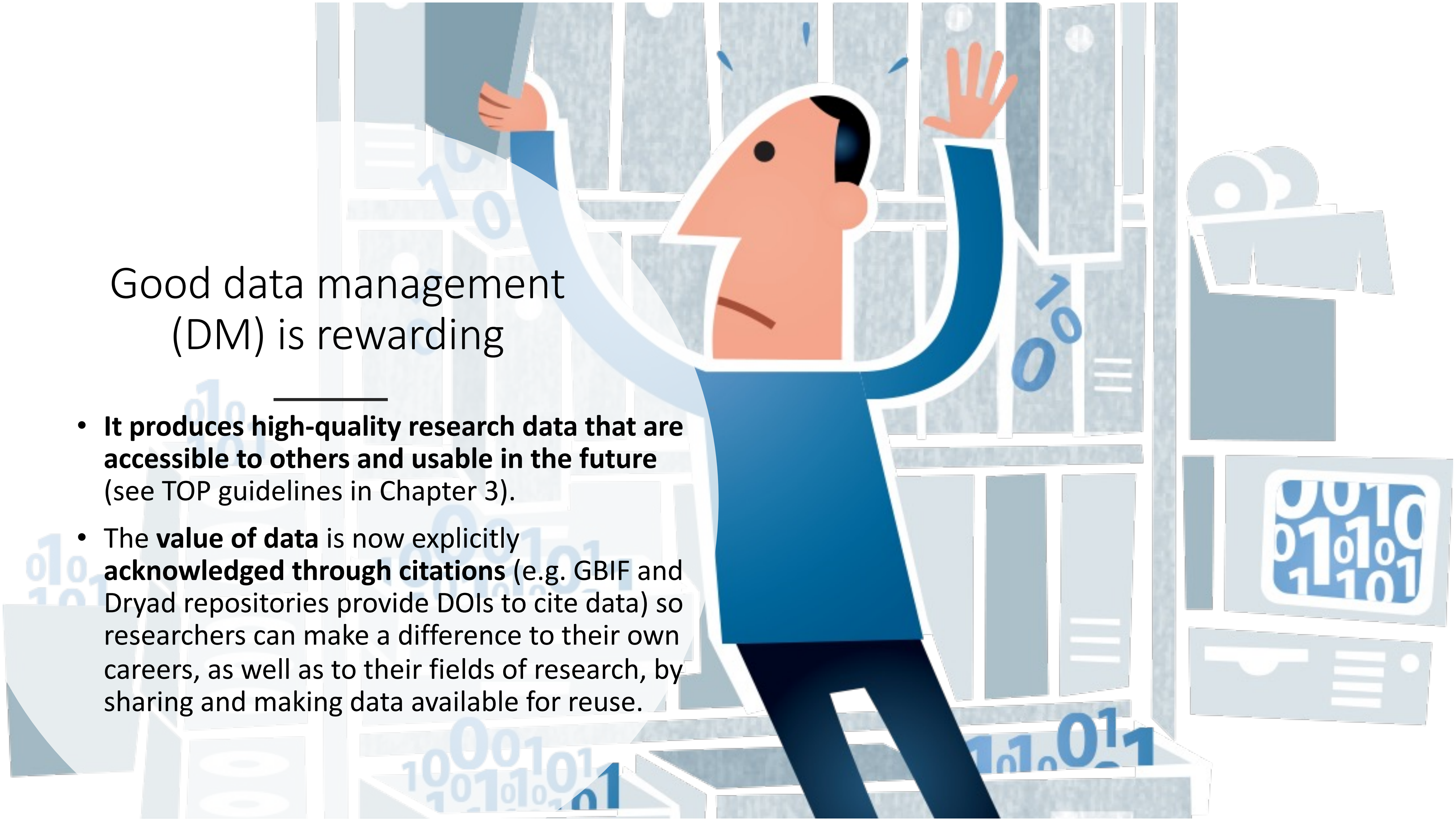
EEB 603 – Chapter 5: Data management

BOISE STATE UNIVERSITY

Learning outcomes

- Learn what are the different types of research data and their specificity.
- Provide an overview of the data life-cycle.
- Highlight the benefits of good data management.
- Study best-practices to planning data management.

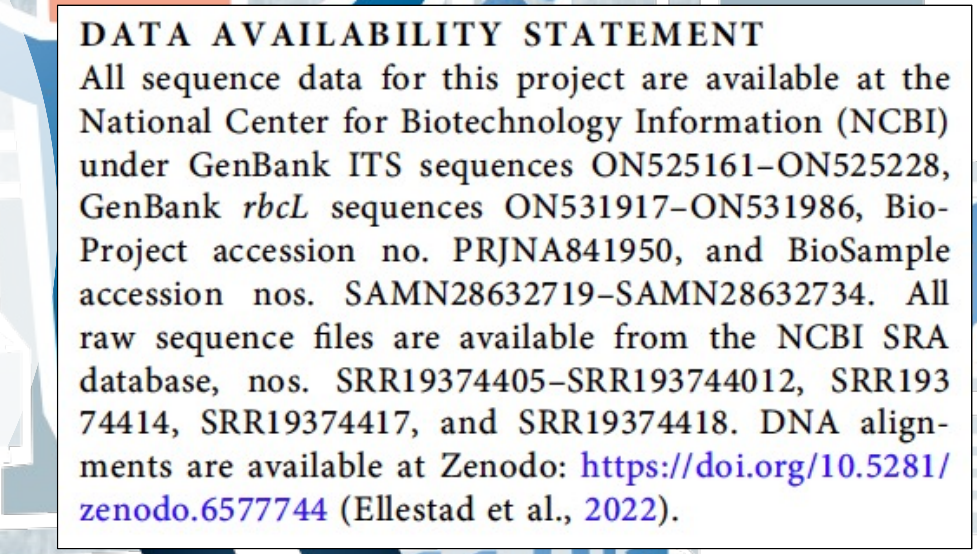


Good data management (DM) is rewarding

- **It produces high-quality research data that are accessible to others and usable in the future** (see TOP guidelines in Chapter 3).
- The **value of data** is now explicitly **acknowledged through citations** (e.g. GBIF and Dryad repositories provide DOIs to cite data) so researchers can make a difference to their own careers, as well as to their fields of research, by sharing and making data available for reuse.

Data Availability Statement

- **All data citations have to be included in publications in dedicated section**

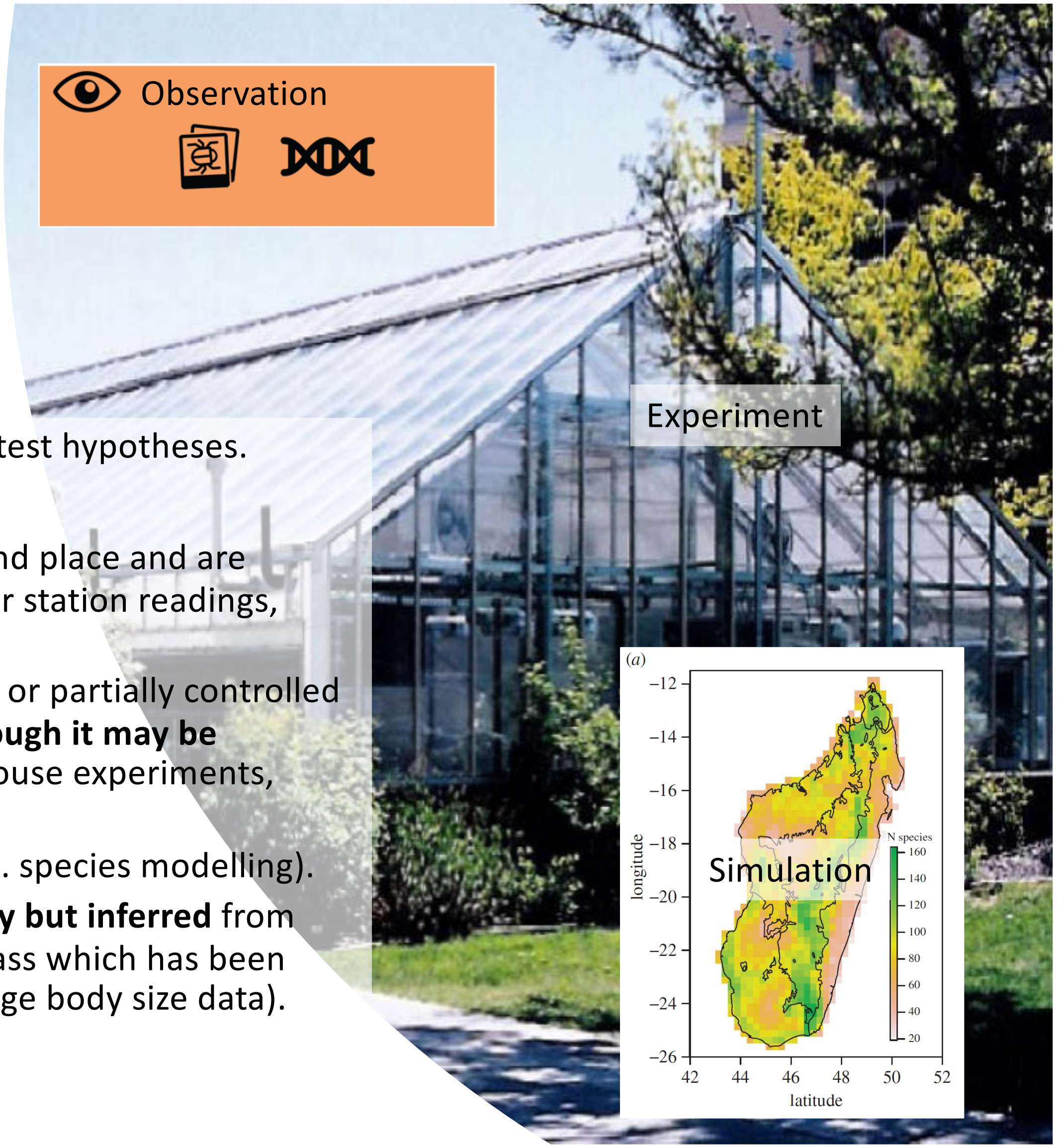


DATA AVAILABILITY STATEMENT

All sequence data for this project are available at the National Center for Biotechnology Information (NCBI) under GenBank ITS sequences ON525161–ON525228, GenBank *rbcL* sequences ON531917–ON531986, BioProject accession no. PRJNA841950, and BioSample accession nos. SAMN28632719–SAMN28632734. All raw sequence files are available from the NCBI SRA database, nos. SRR19374405–SRR193744012, SRR19374414, SRR19374417, and SRR19374418. DNA alignments are available at Zenodo: https://doi.org/10.5281/zenodo.6577744 (Ellestad et al., 2022).

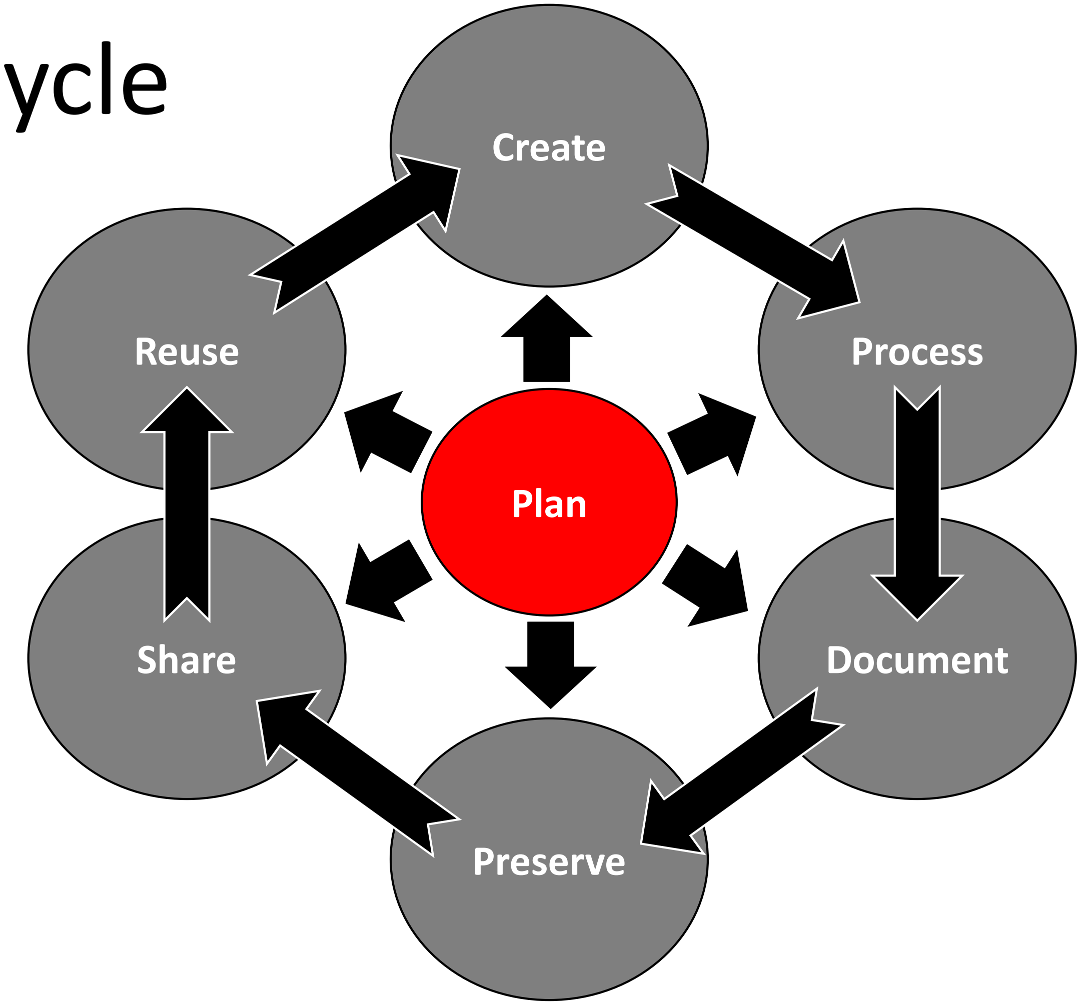
What are research data?

- Research data are the factual pieces of info. used to test hypotheses.
- Data can be classified into five categories:
  1. **Observational:** Data which are tied to time and place and are **irreplaceable** (e.g. field observations, weather station readings, satellite data).
  2. **Experimental:** Data generated in a controlled or partially controlled environment which **can be reproduced, although it may be expensive** to do so (e.g. field plots or greenhouse experiments, chemical analyses).
  3. **Simulation:** Data generated from models (e.g. species modelling).
  4. **Derived:** Data which **are not collected directly but inferred** from (an)other data file(s) (e.g. a population biomass which has been calculated from population density and average body size data).
  5. **Metadata:** Data about data.



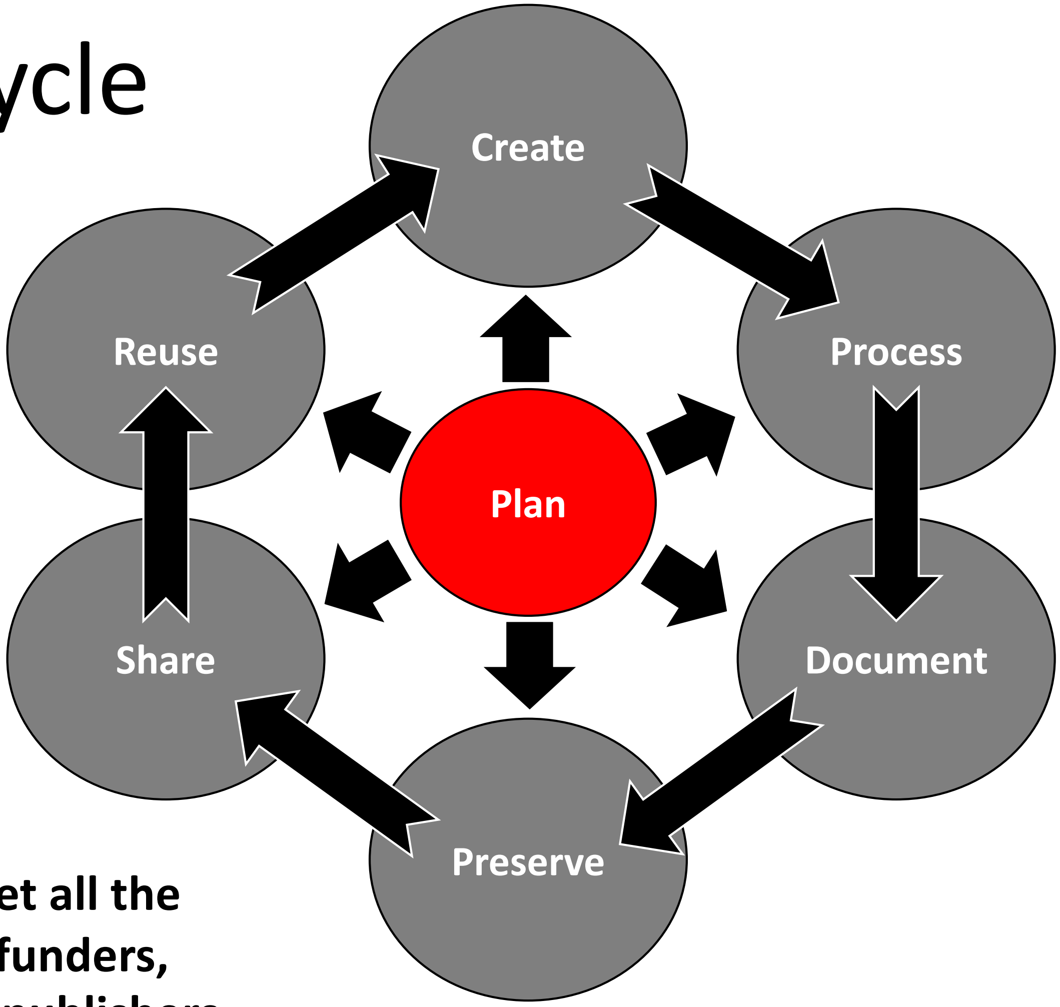
The data life-cycle

The data-life cycle mirrors the scientific process and provides evidence to support your research.



The data life-cycle

The Data Management (DM) plan concerns how you plan for all stages of the data life-cycle and implement this plan throughout the research project.



Make sure that your data meet all the expectations set by yourself, funders, institutions & legislation and publishers

Why should I manage data?

To bring some perspective on this topic, ask yourself this question:

Would a colleague be able to take over my project tomorrow if I disappeared, or make sense of the data without talking to me?

If you can answer **YES**, then you are managing your data well.

Benefits of good data management include

- Ensuring data are accurate, complete, authentic and reliable.
- Increasing research efficiency.
- Saving time and money in the long run – ‘undoing’ mistakes is frustrating and sometimes impossible.
- Meeting funder requirements.
- Minimizing the risk of data loss.
- Preventing duplication by others.
- Facilitating data sharing.

Why should I share my data?

- It is common for funders and publishers to **mandate data sharing**.
- Benefits of sharing data include:
  - Increasing the **impact and visibility** of research.
  - Encouraging **collaborations and partnerships** with other researchers.
  - Maximizing **transparency and accountability**.
  - Encouraging the **improvement and validation of research methods**.
  - Reducing **costs of duplicating data** collection.
  - Advancing science by **letting others use data in innovative ways**.

Sometimes data can't be shared!

- If the datasets contain sensitive information about endangered or threatened species.
- If the data contain personal information – sharing them may be a breach of protocol (and even break the law).
- If parts of the data are owned by others – you may not have the rights to share them.

During the planning stages of your project **determine which of your data can't and shouldn't be shared**. Journal data archiving policies recognize these reasons for not sharing.

Please consult information presented in Chapter 4 for more details on this topic

Planning data management

- Before you start planning:
  - ➢ Check funder specifications for DM plans.
  - ➢ Consult with your institution (especially regarding resources and policies).
  - ➢ Consider your budget.
  - ➢ Talk to your supervisor, colleagues and collaborators.

Key things to consider when planning?

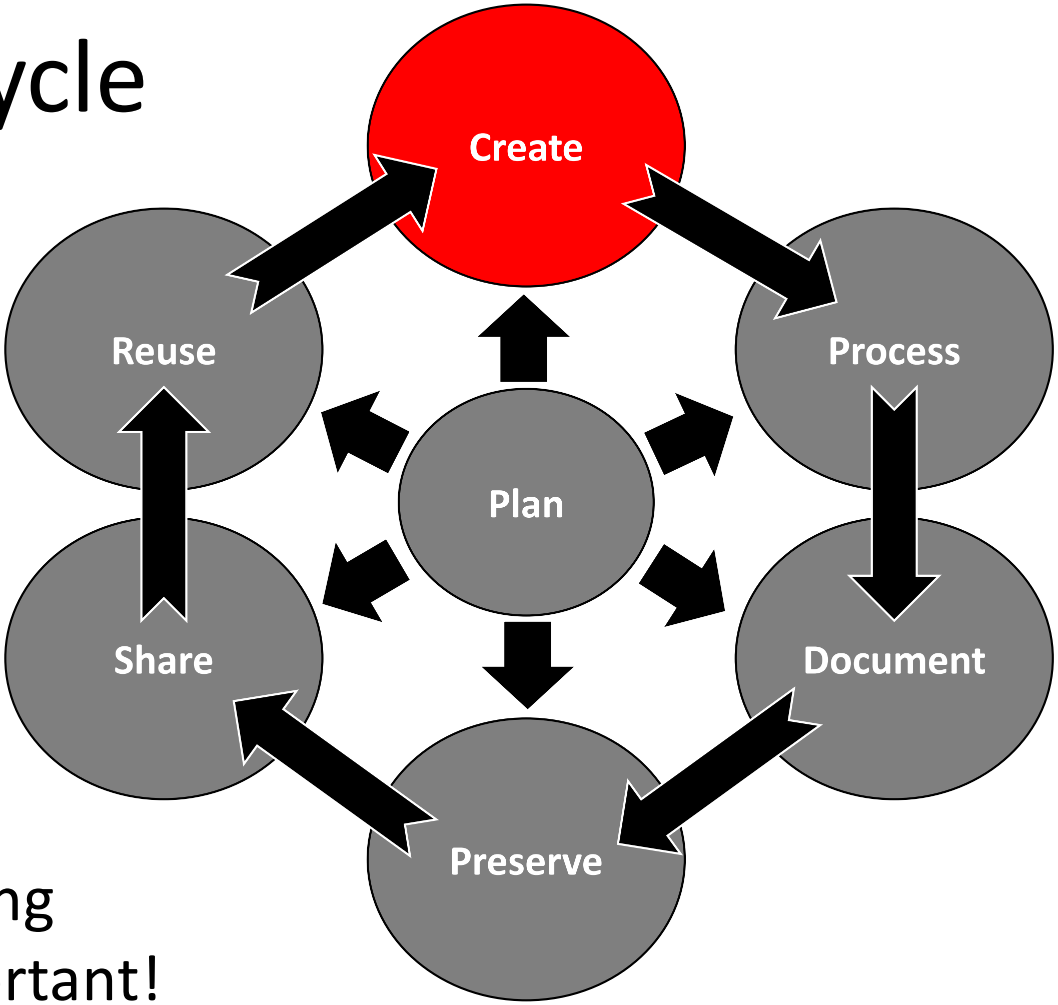
- **Time:** Writing a DM plan takes time. It is not as simple as filling out a template. Planning for DM should be thorough.
- **Design according to needs:** DM should be planned and implemented with the purpose of the research project in mind.
- **Roles & responsibilities:** Creating a DM plan may be the responsibility of one single person, but DM implementation may involve various people. One of the major uses of a DM plan is to enable coordinated working and communication among researchers on a project.
- **Review:** Plan how DM will be reviewed throughout the project and adapted if necessary. It helps integrating DM into the research process and ensure that best practices are implemented. Reviewing also help catching any issues early on.

The data life-cycle

Creating datasets

==

Collecting and digitizing data to generate raw dataset(s)



Quality control(s) during data collection is important!

Key things to consider during data collection

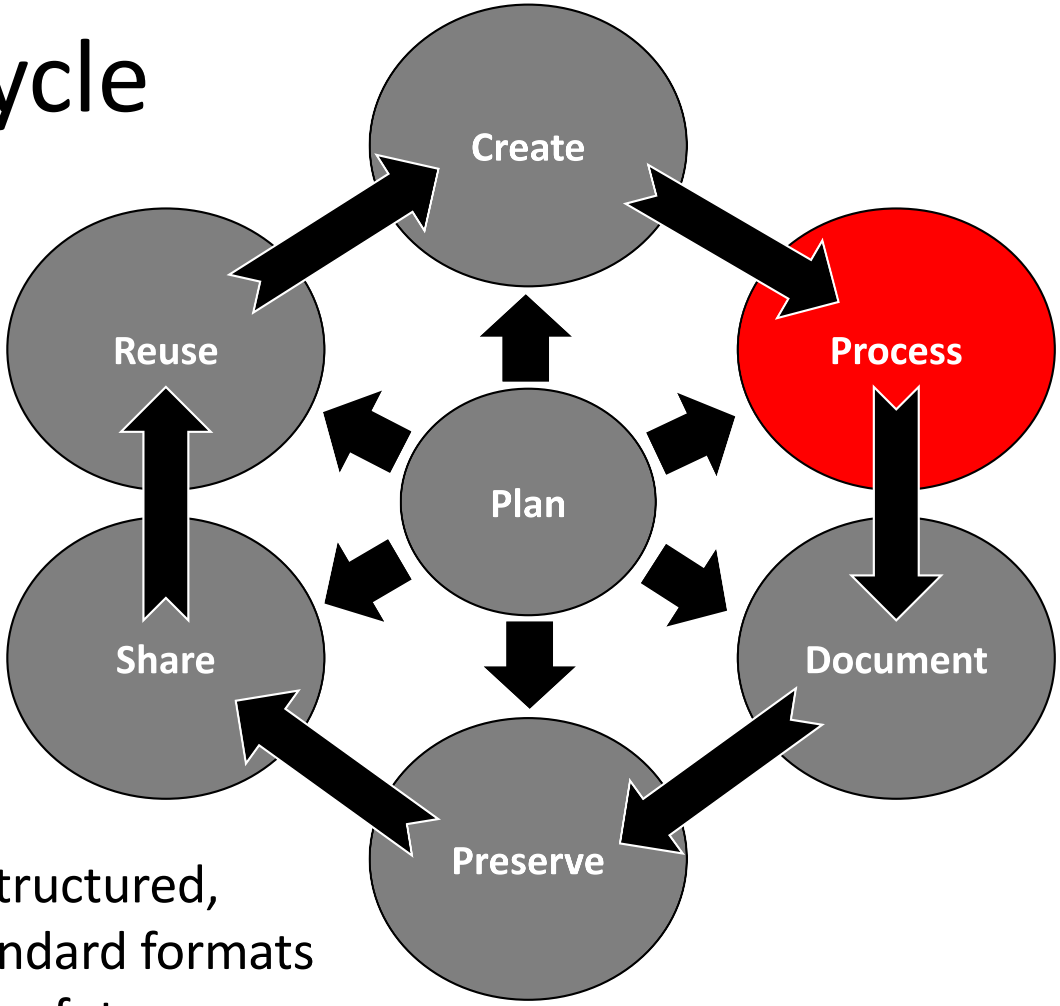
- **Logistical issues** in the field (challenges associated with your fieldwork; e.g. no power to re-charge your batteries).
- **Calibration of instruments**.
- **Taking multiple measurements/observations/samples** (to e.g. ensure statistical accuracy or cover pop. genetic variation).
- **Creating a template/protocol** for use during data collection to ensure that all information is collected consistently.
- Describing any **conditions** during data collection **affecting data quality** (e.g. weather conditions).

Key things to consider during data digitization

- **Designing a database structure** to organize data and data files (information must flow and shouldn't be duplicated).
- **Using a consistent format for each data file** – e.g. rows equal record and columns represent parameters (spreadsheet format).
- **Atomizing data** – make sure that only one piece of data is in each entry (this greatly helps analyses).
- **Using plain text characters** (e.g. ASCII) to ensure data are readable by a maximum number of software.
- **Using code** – coding assigns a numerical value to variables and allows for statistical analysis of data. Keep coding simple (and provide a key).
- **Describing the contents of your data files** in a *Readme.txt*.
- **Keeping raw data raw.**

The data life-cycle

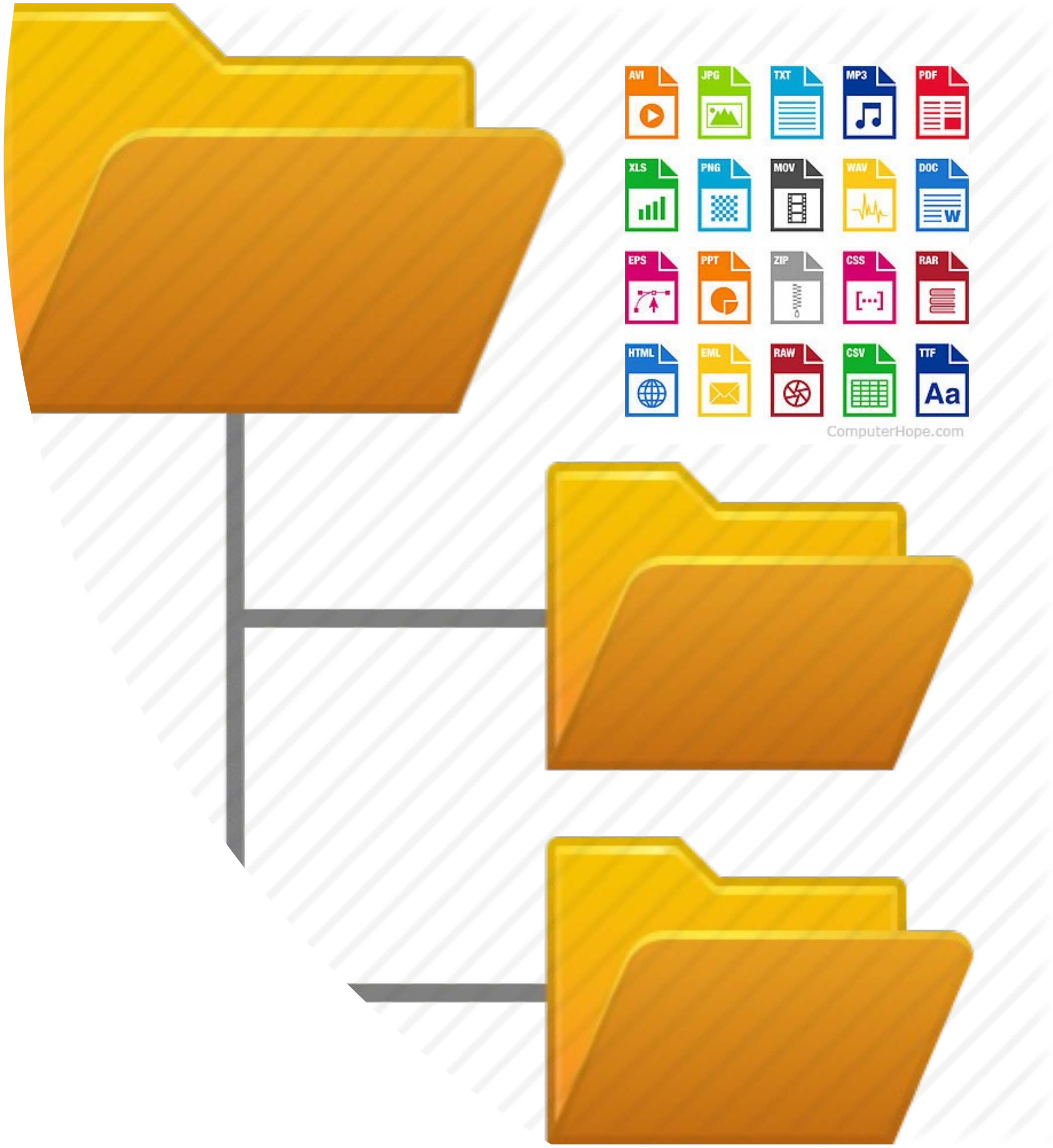
Data should be processed into a format that is suited to subsequent analyses and ensures long-term usability.



Data should be organized, structured, named and versioned in standard formats that can be interpreted in the future.

Guidelines to ensure best processing of data

- **File formats:** Data should be written in non-proprietary formats, also known as open standard formats (e.g. .csv, .txt, .jpeg).
- **File names and folders:** To keep track of data and know how to find them, digital files and folders should be structured and well organized. Use a folder hierarchy that fits the structure of the project and ensure that it is used consistently.
- **File names should be:**
  - ➢ Unique,
  - ➢ Descriptive,
  - ➢ Succinct,
  - ➢ Naturally ordered and consistent,
  - ➢ Describing the project, file contents, location, date, researcher’s initials and version.

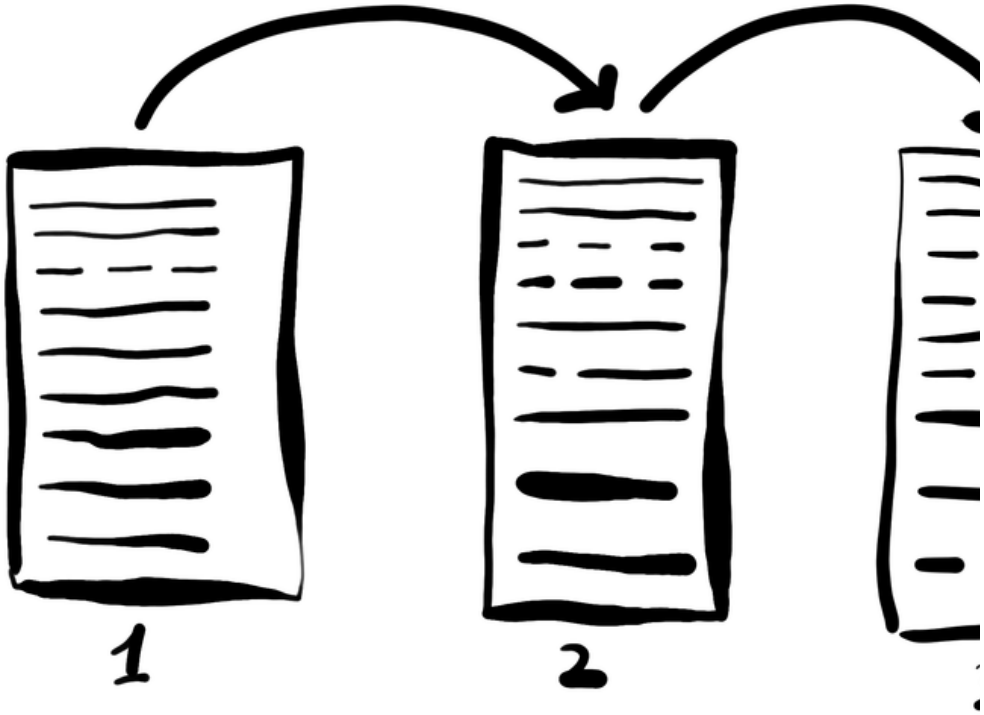


Guidelines to ensure best processing of data

- **File names should not include spaces** – these can cause problems with scripting and metadata.
- **Quality assurance:** Checking that data have been edited, cleaned, verified and validated to create a reliable masterfile, which will become the basis for further analyses
- **Assurance checks may include:**
  - ➢Identifying estimated values, missing values or double entries.
  - ➢Performing statistical analyses to check for questionable or impossible values and outliers (which may just be typos from data entry).
  - ➢Checking the format of the data for consistency across the dataset.
  - ➢Checking the data against similar data to identify potential problems.

Guidelines to ensure best processing of data

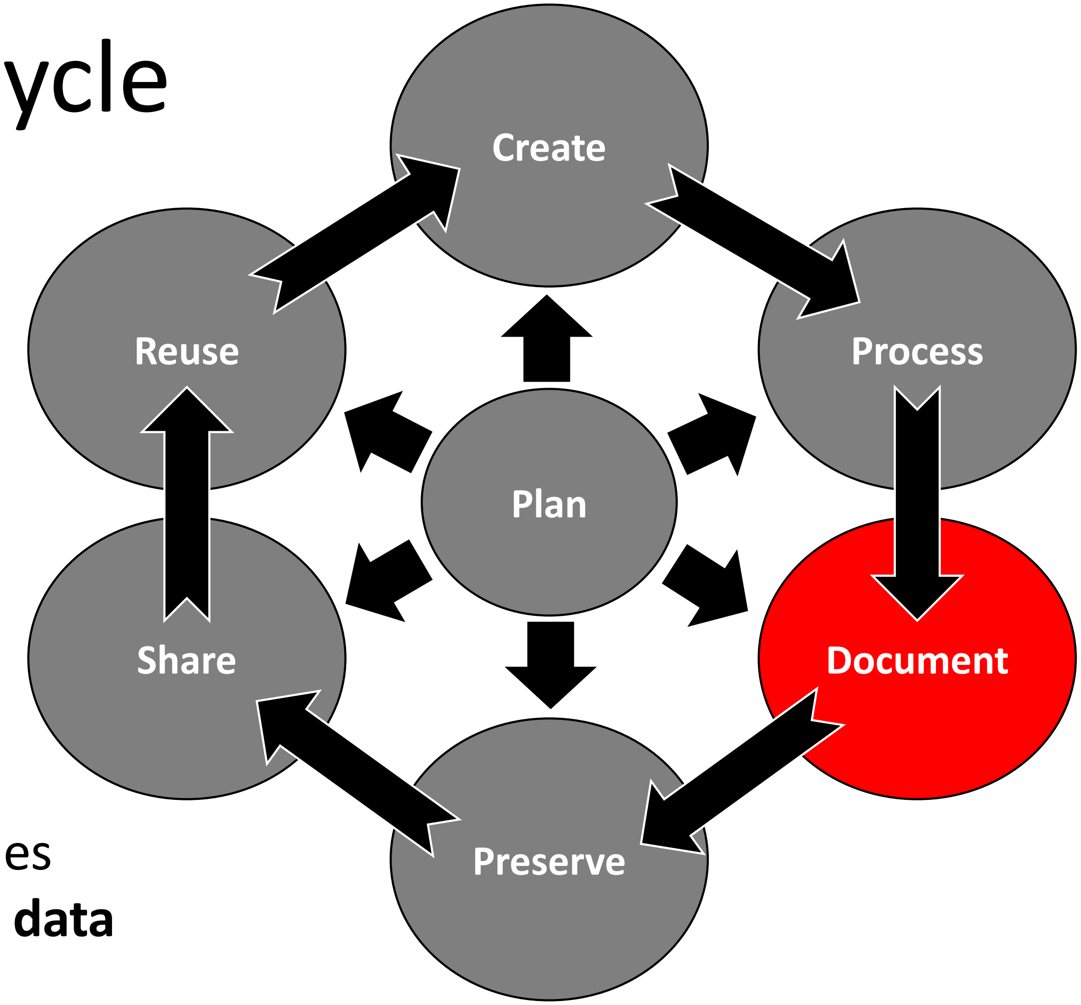
- **Version control:** Once masterfile has been finalized, keeping track of ensuing versions of this file can be challenging…
- **Version control best practice includes:**
    - ➢ Deciding how many and which versions to keep.
    - ➢ Using a systematic file naming convention, using filenames that include the version number and status of the file (e.g. v1_draft, v2_internal).
    - ➢ Record what changes took place to create the version in a separate file.
    - ➢ Mapping versions if they are stored in different locations.
    - ➢ Synchronizing versions across different locations.



The data life-cycle

Producing good documentation and metadata ensures that data can be understood and used in the long term.

Data documentation includes **information at project and data levels**.



Project level – Information to provide

- The project aim, objectives and hypotheses.
- Personnel involved throughout the project, including who to contact with questions.
- Details of sponsors.
- Data collection methods, including details of instrumentation and environmental conditions during collection, copies of collection instructions if applicable.
- Standards used.
- Data structure and organisation of files.

Project level – Information to provide

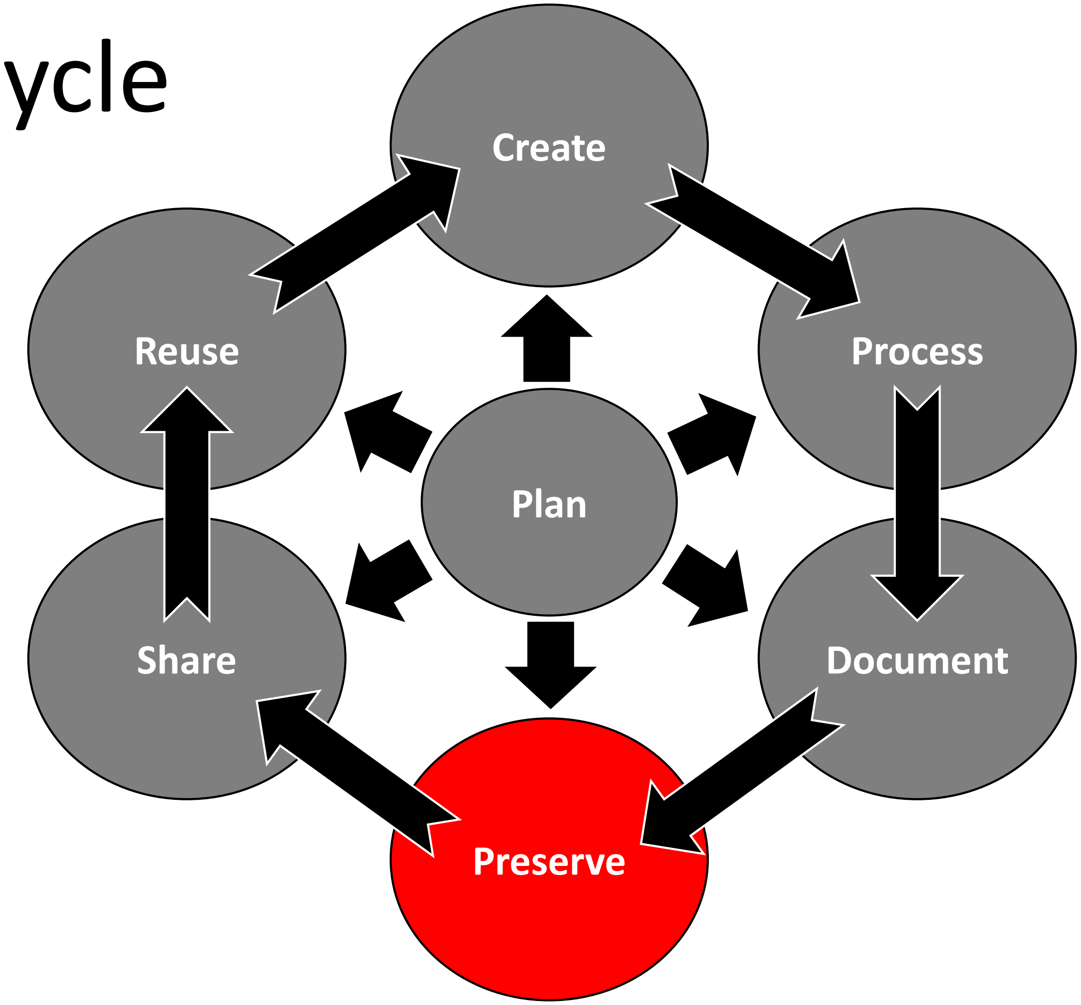
- Software used to prepare and read the data.
- Procedures used for data processing, including quality control and versioning and the dates these were carried out.
- Known problems that may limit data use.
- Instructions on how to cite the data.
- Intellectual property rights and other licensing considerations.

Data level – Information to provide

- Names, labels and descriptions for variables.
- Detailed explanation of codes used.
- Definitions of acronyms or specialist terminology.
- Reasons for missing values.
- Derived data created from the raw file, including the code or algorithm used to create them.

The data life-cycle

To protect data from loss and to make sure data are securely stored, good DM should **include a strategy for backing up and storing data** effectively

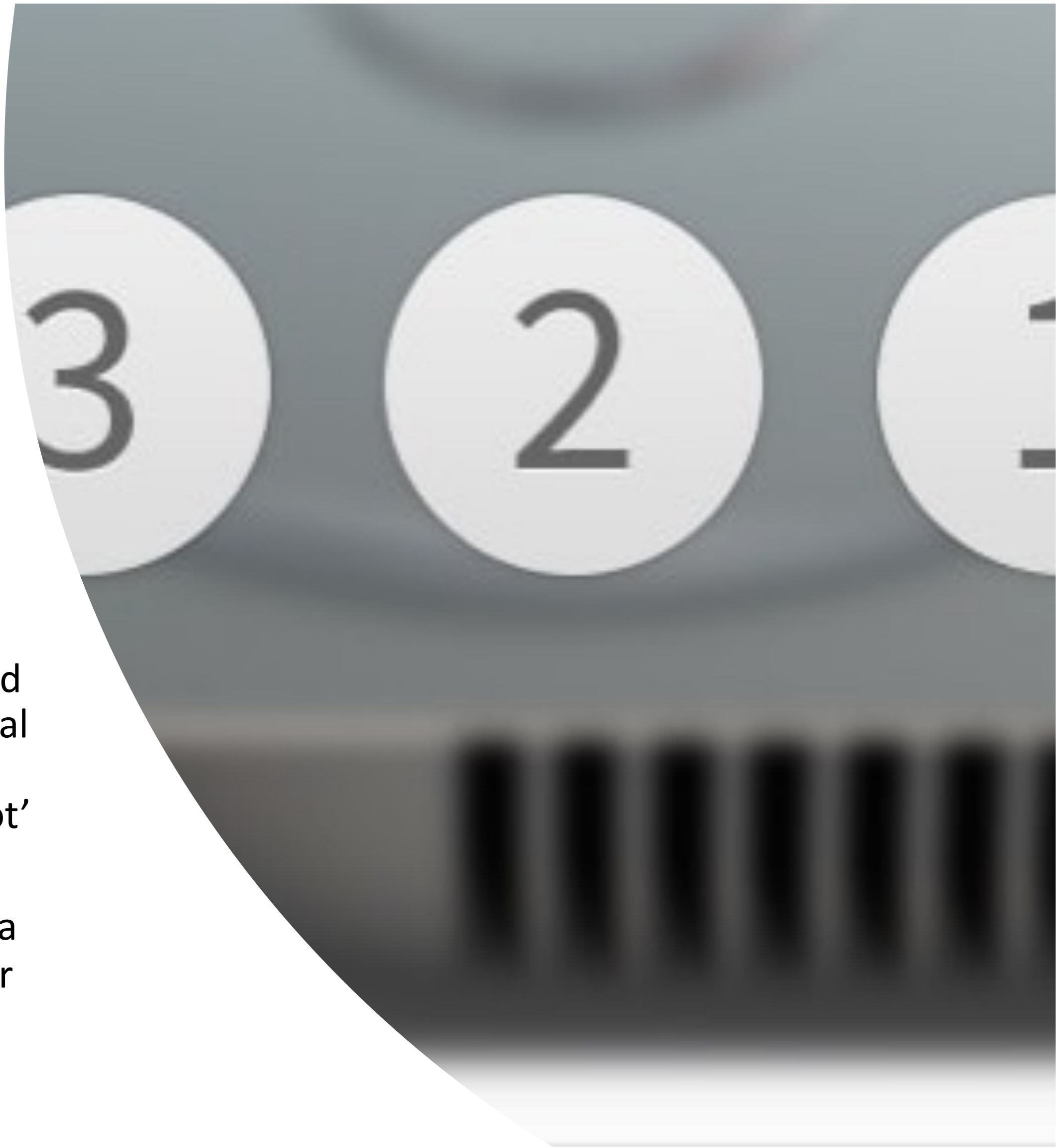


Data backup – How can data be lost?

- Hardware failure.
- Software faults.
- Virus infection or hacking.
- Power failure.
- Human error.
- Hardware theft or misplacement (especially common during fieldwork).
- Hardware damage (e.g. fire, flood. Again common in the field).
- Backups – good backups being overwritten with backups from corrupt data (this happen more often that you would imagine).

Things to consider to establish a backup plan

- Which files require backup.
- Who is responsible for backups.
- The frequency of backup needed, this will be affected by how regularly files are updated.
- Whether full or incremental backups are needed – consider running a mix of frequent incremental backups (capturing recent data changes) along with periodic full backups (capturing a ‘snapshot’ of the state of all files).
- Backup procedures for each location where data are held, e.g. tablets, home-based computers or remote drives.
- How to organize and label backup files.



Best practices for data storage

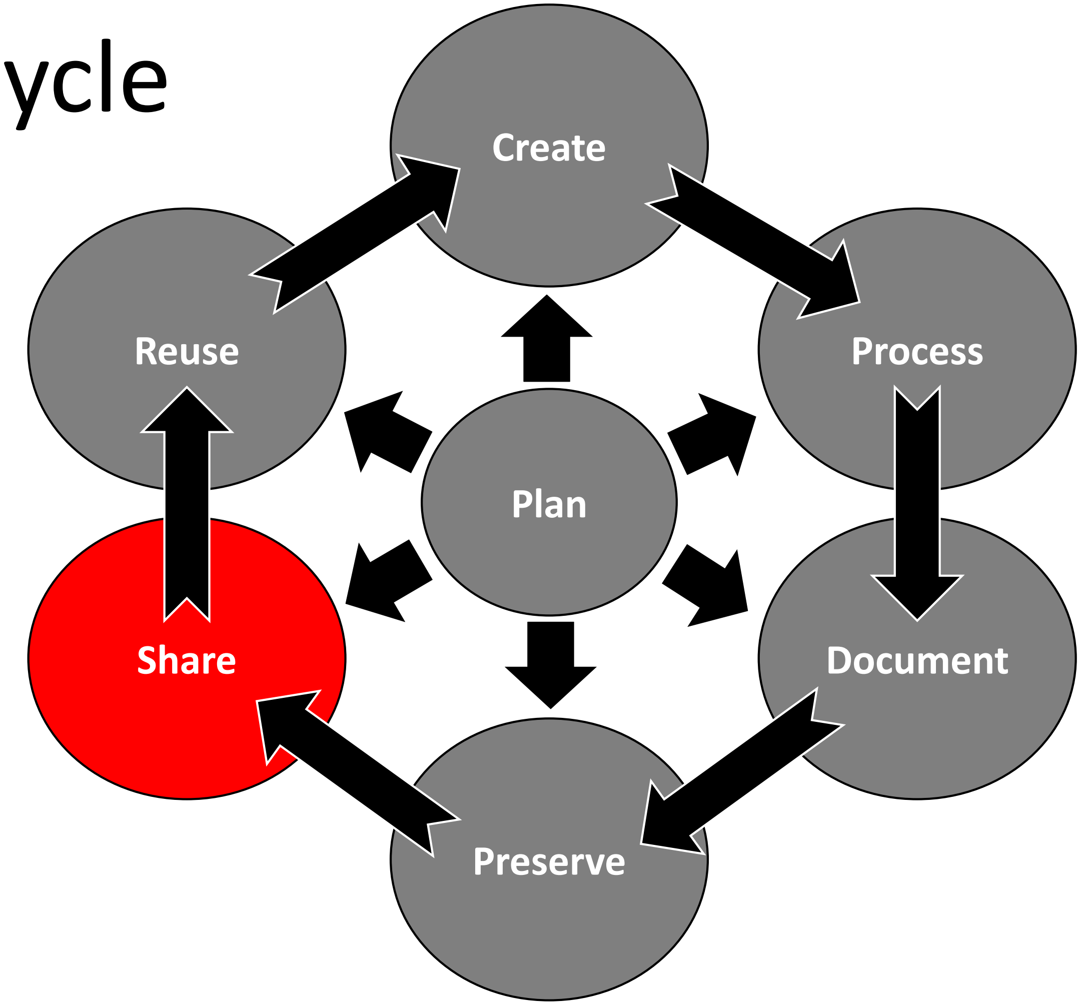
- Use high-quality storage systems (e.g. media, devices).
- Use non-proprietary formats for long-term software readability.
- Migrate data files every two to five years to new storage – storage media such as CDs, DVDs and hard drives can degrade over time.
- Check stored data regularly to make sure nothing has been lost.
- Use different forms of storage for the same data, this also acts as a method of backup, e.g. using remote storage, external hard drives and a network drive.
- Label and organize stored files logically to make them easy to locate and access.
- Think about encryption: sensitive data may need to be encrypted.

Storage and backing-up of data

- **Network drives** managed by IT staff and regularly backed up. They ensure secure storage and prohibit unauthorized access to files.
- **Personal devices** e.g. laptops are convenient for short-term, temporary storage but should not be used for storing master files.
- **External devices** such as hard drives, USB sticks, CDs and DVDs are often convenient because of their cost and portability. However, they do not guarantee long-term preservation and can be lost/stolen.
- **Remote or online services** such as Google Drive use cloud technology to allow users to synchronize files across different computers.
- **Paper!** If data files are not too big, do not overlook the idea of printing out a paper copy of important data files.

The data life-cycle

Institutions, funders and journals have specific policies associated with data sharing and you should read those prior to establishing a data sharing plan.



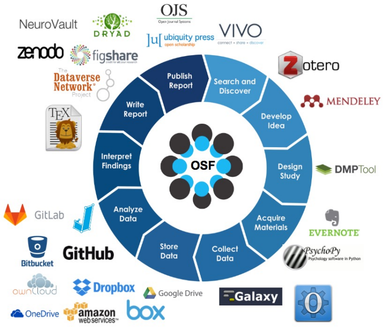
Sharing data – Guidelines

- Use a disciplinary data center such as Dryad and/or GenBank.
- Deposit data in your research funder's data center.
- Deposit data in university repositories.
- Make data available online via open notebooks or project websites (e.g. Open Science Framework from the Center of Open Science).
- Use virtual research environments such as OSF, SharePoint (a Microsoft web-based collaborative platform) and Sakai.

SharePoint

Sakai

https://sakaiproject.org/



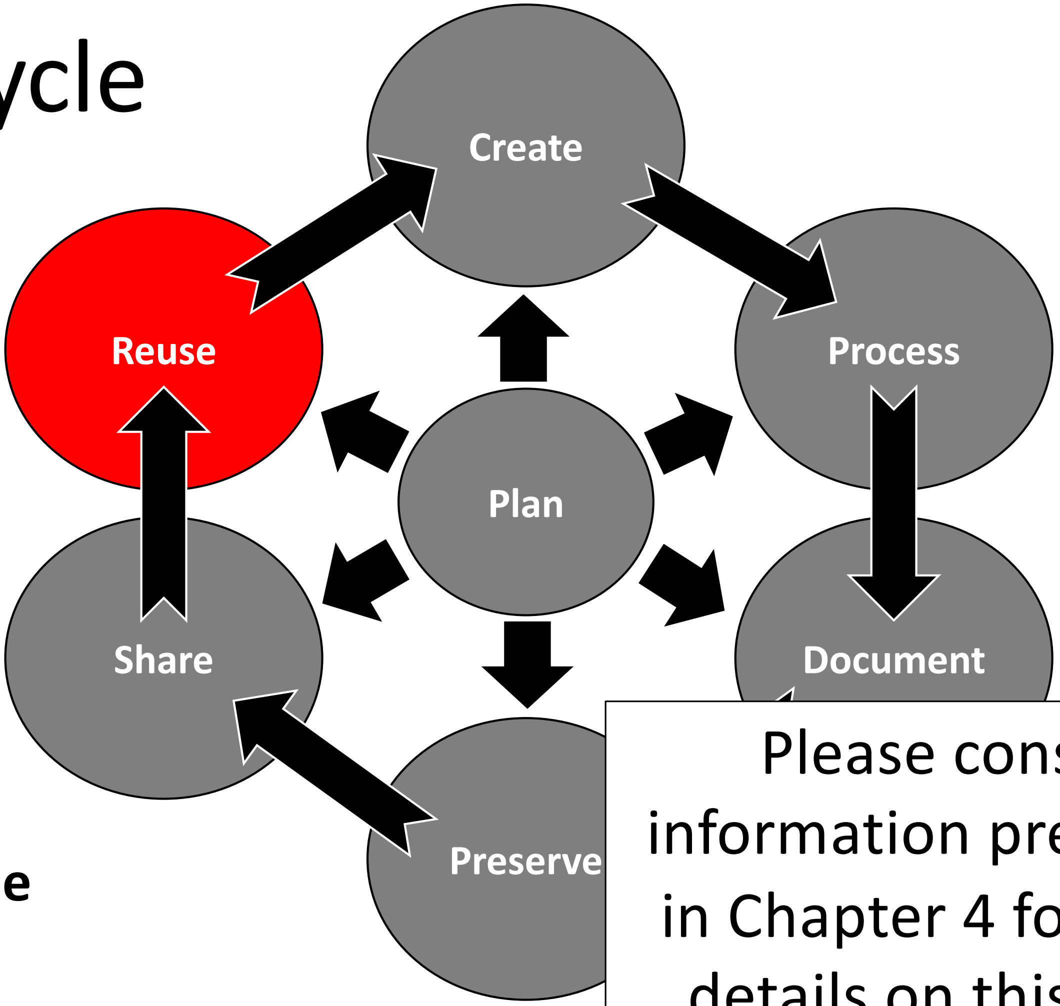
Data repositories

- Archiving your data in a repository is a reliable method of sharing data. Data submitted to repositories have to conform to submission guidelines, which restricts which data you share via the repository.
- Advantages of sharing data through centers include:
  - Assurance for others that the data meet quality standards.
  - Guaranteed long-term preservation.
  - Data are secure and access can be controlled.
  - Data are regularly backed up.
  - Chances of others discovering the data are improved.
  - Citation methods are specified.
  - Secondary usage of the data is monitored.

The data life-cycle

All aspects of data management lead up to data discovery and reuse by others!

Intellectual property rights, licenses and permissions, which concern reuse of data, should be explained in the data documentation and/or metadata.



Please consult information presented in Chapter 4 for more details on this topic

Reusing data

- **State your expectations for the reuse of your data**, e.g. terms of acknowledgement, citation and co-authorship. It becomes the **responsibility of others to reuse data effectively**.
- When requesting to use someone else's data clearly state the purpose of the request, including the idea you will be addressing and your expectations for co-authorship or acknowledgement. Co-authorship is a complex issue.
- **Increasing openness to data and ensuring long-term preservation of data fosters collaboration and transparency, furthering research that aims to answer the big questions in ecology and evolution**.
- By implementing good DM practices, researchers can ensure that high-quality data are preserved for the research community and will play a role in advancing science for future generations.