Data Management Planning

Dr. Sara El-Gebali

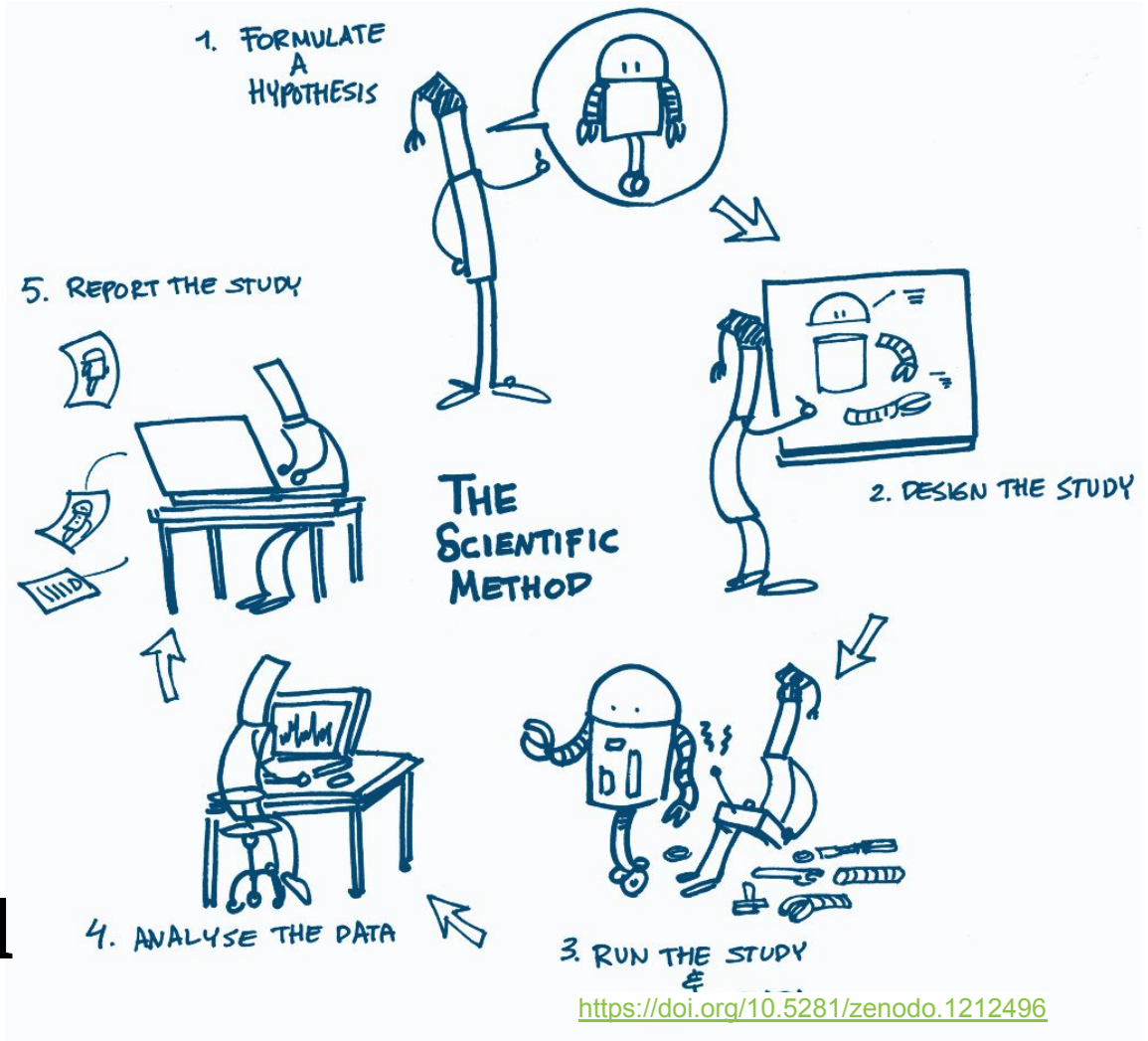
0000-0003-1378-5495
@yalahowy

19-July-2022

FAIRpoints

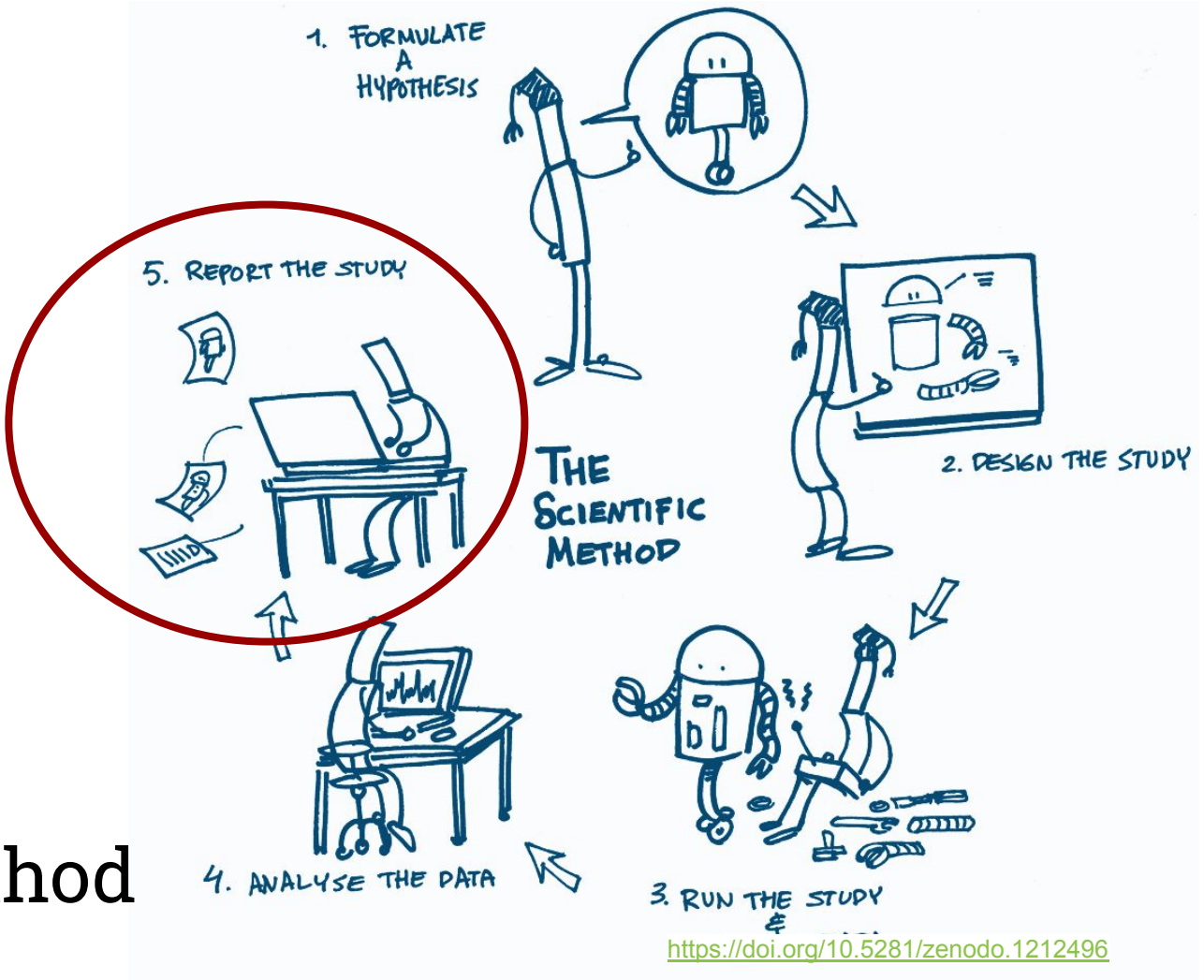
SciLifeLab

Scientific method



https://doi.org/10.5281/zenodo.1212496

Scientific method



https://doi.org/10.5281/zenodo.1212496

PLAN & DESIGN: Plan processes from onboarding to project closure and data resources

COLLECT & CREATE: Organization and integration of data sets and collection processes

ANALYZE & COLLABORATE: Processing and analyzing data should be collaborative and documented

STORE & MANAGE: Each stage of the Biomedical Data Lifecycle revolves around the management of data storage

EVALUATE & ARCHIVE: Identify essential research records and evaluate for retention

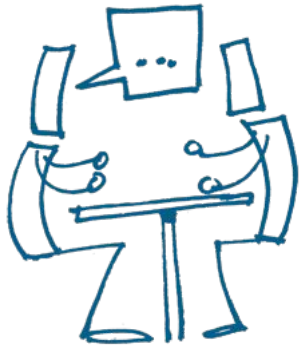
SHARE & DISSEMINATE: Establishing and supporting the reach and impact of your data

ACCESS & REUSE: Ensuring the broad utility of your research data efforts for other researchers

Harvard Biomedical Data Management

https://datamanagement.hms.harvard.edu/

Planning



doi.org/10.5281/zenodo.1212496

Data management planning (DMP)

Data description & metadata extraction

Data documentation

Choice of repositories

Choices of file formats

File naming

Data re-use

Funders requirements

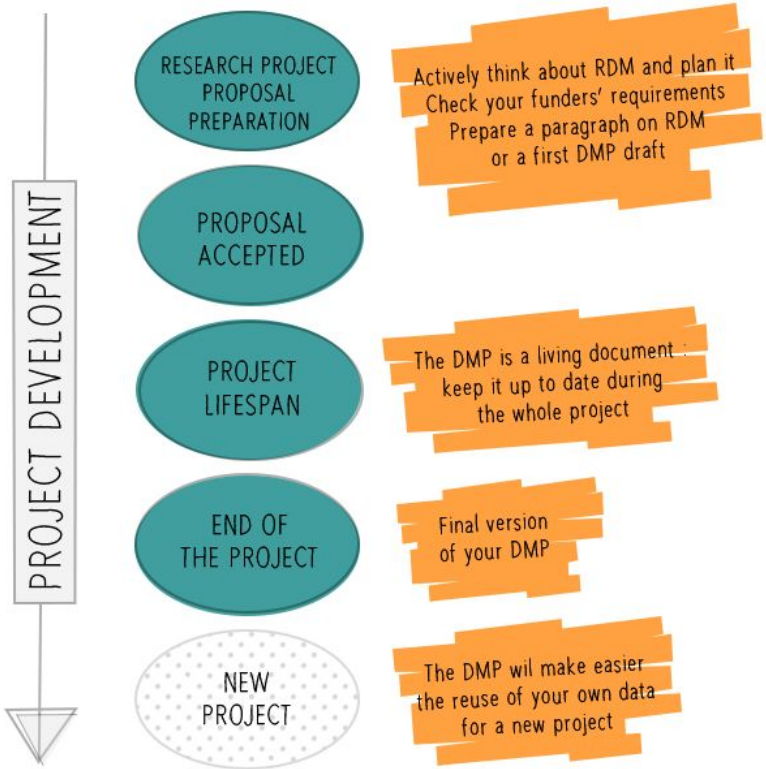
Ethics and Research conduct

Funding for RDM activities

Data Management Plans (DMPs)

A **data management plan** (**DMP**), is a formal document where you outline what you will do with your data **during** and **after** a research project.

DMPs are a 'living' document - should be updated during a research project as necessary.



https://infoscience.epfl.ch/record/294783/files/RDM_Walkthrough_Guide_20190930.pdf?ln=en

Why DMP?

Efficiency:
Reduces risk of data loss, quality control strategies, ensures sustainability and accessibility of data in the long-term, saves time and avoids problems

Reuse:
Provides quality assurance, outlines conditions for use/re-use, provides essential information e.g. data, software, protocols, file formats, etc.

Compliance:
Requested by funders, institutions, many organizations

Security:
Provides necessary information regarding access and security

Anatomy of a DMP

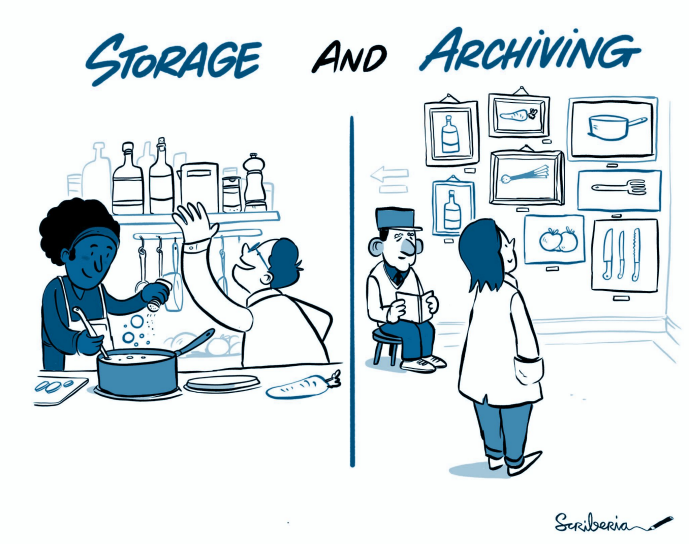
Administrative

Basic information about the project and those involved

- Project name
- Project description
- Contact details for primary investigator
- Organization details (unit, institute, etc..)
- Funding/grant details
- Details for contributors (those involved in the project)
- Roles and Responsibilities

Anatomy of a DMP- Storage

- What kind of data do you have? Digital or non-digital?
- Do you have sensitive or private data? What level of security does it require?
- Where will you store it?
- How will you transfer it?
- How will you store it?
- How long will you store it for?
- How many copies?
- Which version? How to keep track of different versions?
- How will the data be backed up?
- How often?
- Who has access? Including data during and after research?
- How will you safely delete or destroy the data?

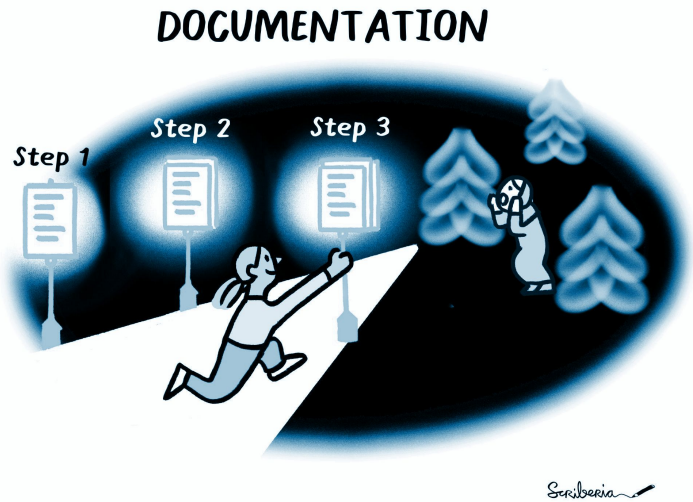


The Turing Way Community, & Scriberia. (2022). Illustrations from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.3332807

Anatomy of a DMP- Data Documentation & Metadata

What is necessary to document ? e.g. details on how the data will be collected or produced.

- Which type of data will be generated, or collected?
- How much data will be produced?
- What will you document? (e.g. protocols, variables, measurements)
- Which metadata will you document?
- Which quality control measures will you implement?
- Which file formats?
- Which equipment, software or tools are required?
- Where is the data coming from? (e.g. previous research, repository)



The Turing Way Community, & Scriberia. (2022). Illustrations from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.3332807

Anatomy of a DMP- Data Documentation & Metadata

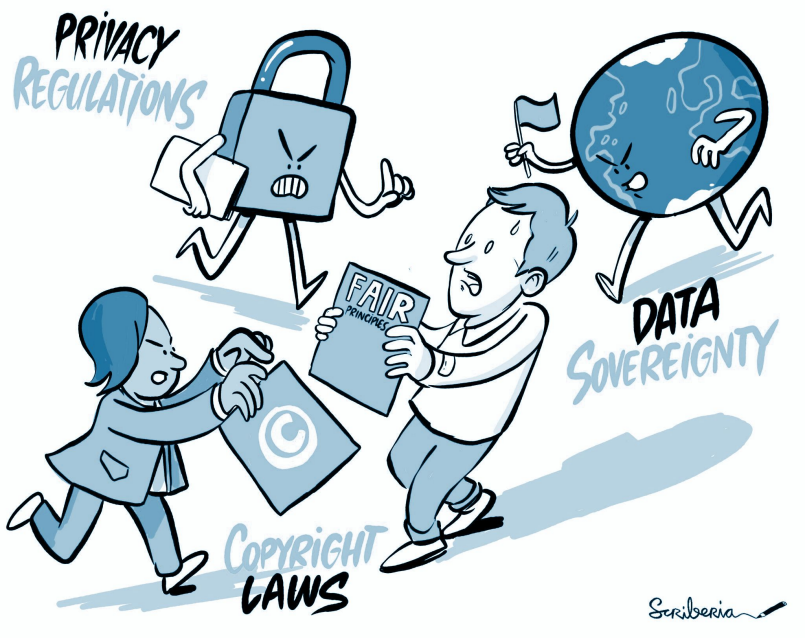
How will it be structured? e.g. how will you ensure consistency in your documentation?

- Which metadata standards will you use?
- Which terminology, ontologies, standards, controlled vocabularies and abbreviations will be used?
- How will it be documented? (e.g. ELNs, logbook)
- How often will you document it?
- How will you name your files?
- How will your organize your files?(folder structure)
- Which versioning strategies will you adopt?

The Turing Way Community, & Scriberia. (2020, March 3). Illustrations from the Turing Way book dashes. Zenodo. http://doi.org/10.5281/zenodo.3695300

Anatomy of a DMP- Access & Security

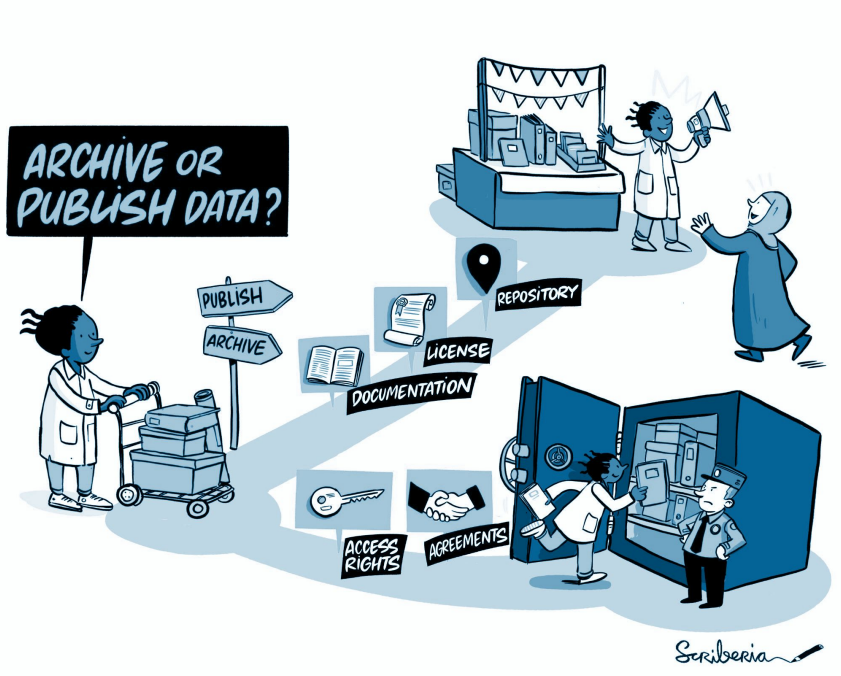
- Does this data require an ethics assessment? (e.g. ethical review)
- Is there any confidential data?
- What is the security classification for the data?
- Is there any processing of personal data?
- Are there any intellectual property rights?
- What are the access conditions? Who has access and when?
- Is there any embargo on the data/parts of it?
- Are there any other agreements or limitations that might be in place? (e.g. 3rd party agreements)
- Does this data require anonymization?



The Turing Way Community, & Scriberia. (2022). Illustrations from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.3332807

Anatomy of a DMP- Share and Reuse

- Which data will you preserve and which one will you share?
- Who owns the copyright in your data?
- When will you share your data?
- How are others allowed to use it? (e.g. license)
- Where will you deposit your data?



The Turing Way Community, & Scriberia. (2022). Illustrations from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.3332807

Software Management Plans (SMPs)

Software Management Plans

Why write a Software Management Plan?

Research software can take many guises. It can be a 50 line bash shell script for manipulating and filtering files, a collection of 100 line R scripts for running a bioinformatics analysis, 10,000 lines c Java for medical image analysis or 100,000 lines of Fortran for computational fluid dynamics. It may be written in scripting languages such as Unix shell, Python, R or MATLAB or in "traditional" programming languages such as C, C++, Fortran or Java. But, whatever guise it takes, research software is an integral part of the modern research ecosystem.

When developing research software, it is easy to focus only on goals and activities such as collaborating with other researchers, writing papers, attending conferences and applying for funding. Together, the demands of daily research practice can all conspire to prevent proper planning for the development of research software.

A Software Management Plan (SMP) can help you to define a set of structures and goals to understand your research software including what you are going to develop; who the software is for (even if it is just for yourself); how you will deliver your software to its intended users; how it will help them; and how you will assess whether it has helped them, and contributed to research, i the ways that you intended. An SMP also helps you to understand how you can support those wh wish to, or do, use your research software; how your software relates to other artefacts in your research ecosystem; and how you will ensure that your software remains available beyond the lifetime of your current project.

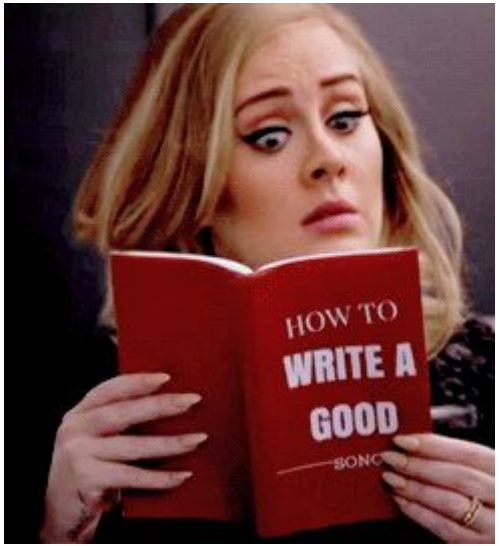
Though an SMP can be of most benefit when starting a project to develop research software, ther are benefits to adopting one on a project that is already underway. An SMP provides a way to dra

https://www.software.ac.uk/software-management-plans

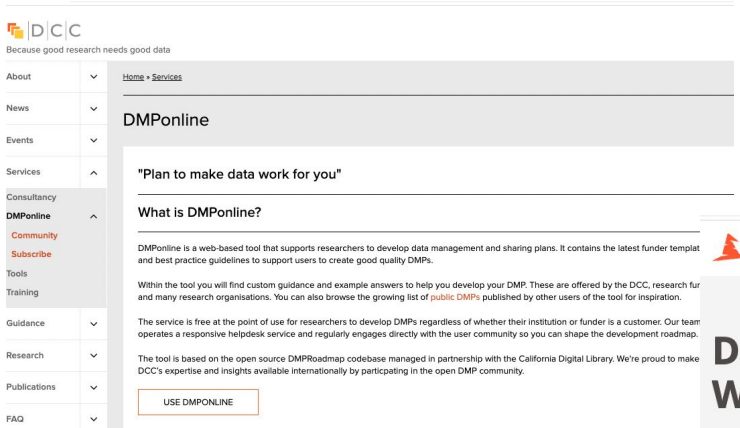
Checklist for a Software Management Plan

https://doi.org/10.5281/zenodo.1422656

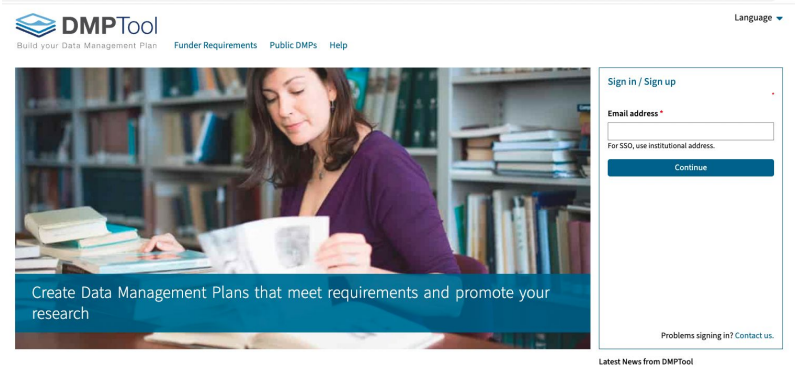
How to!



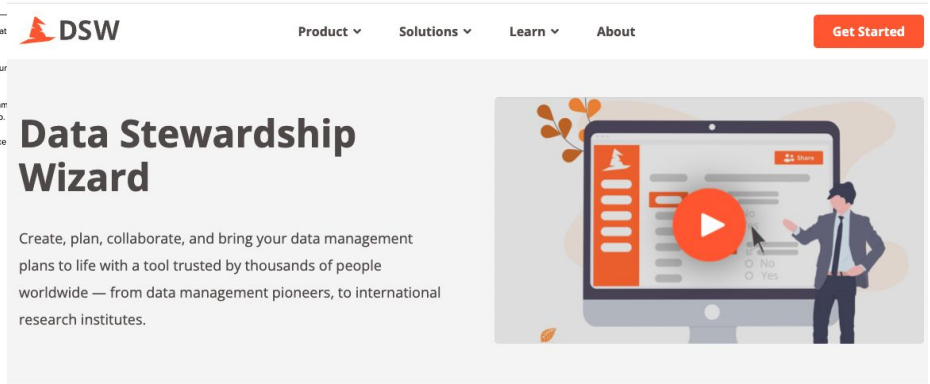
DMP tools



https://www.dcc.ac.uk/dmponline



https://dmptool.org/



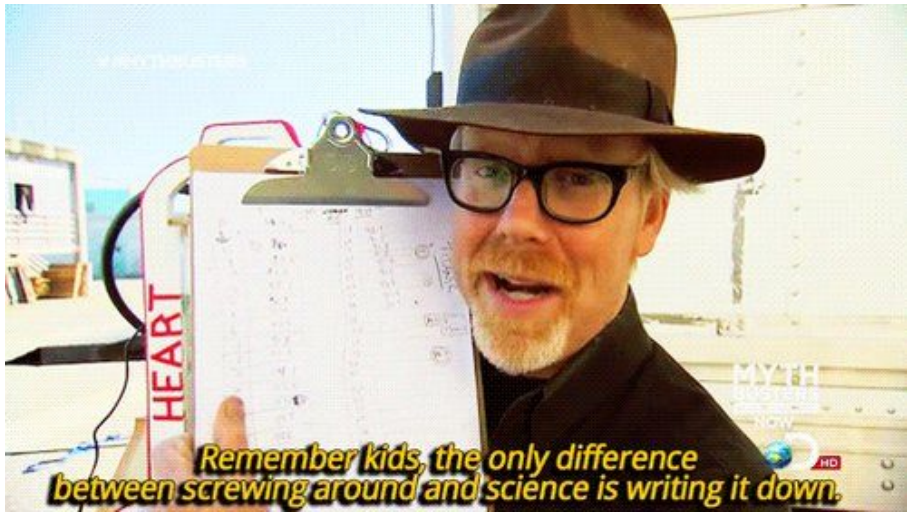
https://ds-wizard.org/

Checklists

- UK Data services: https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/checklist/
- Swedish National Data Archive https://snd.gu.se/en/manage-data/guides/dmp-checklist
- EPFL https://www.epfl.ch/campus/library/wp-content/uploads/2018/09/DMP-Checklist.pdf

Document everything!

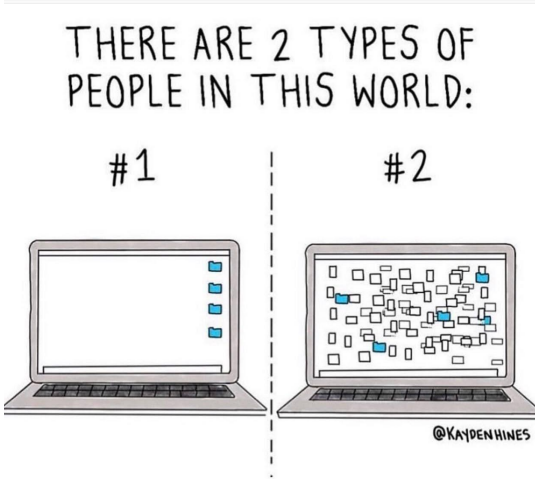
- **Who** created and owns this data?
- **What** are the contents?
- **What** output and results?
- **When** was this data created and last updated?
- **Where** is it stored and published?
- **Which** methods were used?
- **Which** instruments were used?
- **How** was the data created, controlled and analysed?
- **How** can I use this data i.e. license?



https://www.tested.com/making/557288-origin-only-difference-between-screwing-around-and-science-writing-it-down/

Build new habits

- Organize your data
- Name your files appropriately
- Choose file formats wisely
- Use versioning strategies
- Outline quality control strategies

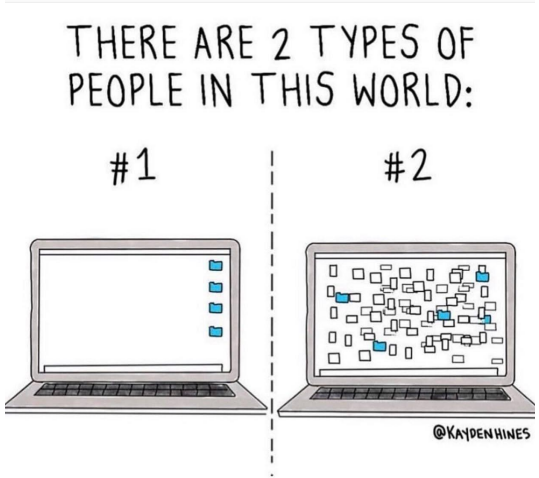


Organize your data

Organize your data in a logical manner

Separate the data according to type: i.e. raw data, analysis, code,

Use directories and folders hierarchy

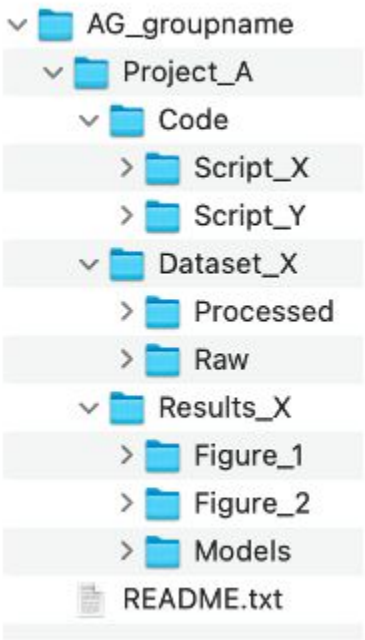


A clear directory structure will make it easier to locate files and versions and this is particularly important when collaborating with others.

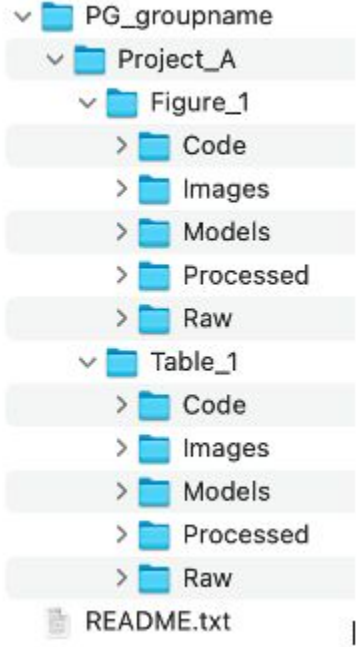
Directory structure guidelines

Consider a hierarchical file structure starting from broad topics to more specific ones nested inside, restricting the level of folders to 3 or 4 with a limited number of items (max. 50 items if possible) inside each folder.

A) Organized by file type



B) Organized by analysis



Name your files

Common guideline, you should know what your fil
is before you double click it!

A file name should be unique, consistent and descriptive.

This allows for increased visibility and discoverability and can be used to easily classify and sort files.

Do’s and Don’ts of file naming

Do’s

- names that are not too short or too long, i.e., no less than 12-14 characters exceptions;t generic i.e.README
- Use identifiers to classify types of files i.e., Int1 (interview 1)
- Preferably use underscores (_) or hyphens (-) as an element separator
- If applicable, include versioning within file names
- Ensure file format extension is present (e.g. .doc, .xls, .mov, .tif, .fasta, .html)
- For dates use the ISO 8601 standard: **YYYY-MM-DD** at the end of the file **UNLESS** you need to organize the files chronologically
- For experimental data files, use the project/experiment name and conditions as abbreviations
- Add a README file in your top directory, include naming convention, directory structure, and abbreviations

Do’s and Don’ts of file naming

Don’ts

- Avoid using capital letter to separate words such as CamelCase
- Avoid naming files/folders with individual names as it impedes handover and data sharing.
- Avoid long names. e.g., no longer than 35-40 characters.
- Avoid using spaces, dots, commas and special characters (e.g. " / \ ~ : ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “ |), or any foreign (Unicode) characters e.g. äöüß r カイダー字 .
- Avoid repetition for ex. Directory name Electron_Microscopy_Images, then you don’t need to name the files ELN_MI_Img_20200101.img.

README

The purpose of a README file is to give an overview of the content, aiding individuals in making sense of the data enclosed thereby persevering the long-term value of the data. This can be very helpful if you are sharing your data with others, or to keep track of content and edits or changes made in multiple projects, or revisiting data after some time has passed.

- A README file is better suited for a collection of data such as a directory for a specific project or experiment, software tool, or any data that is related to each other “logically”.
- Place the README file in a parent directory associated with the content described.
- Use plain markdown or a simple text editor to create the README file in either .md or .txt file format.
- For dates use the ISO 8601 standard: YYYY-MM-DD.
- Whenever possible use the standard vocabulary from your field, see metadata standards directory by RDA community.

Quality control

Quality control is a fundamental step in research, which ensures the integrity of the data and could affect its use and reuse and is required in order to identify potential problems.

It is therefore essential to outline how data collection will be controlled at various stages (data collection, digitisation or data entry, checking and analysis).

How quality control could save your science

It may not be sexy, but quality assurance is becoming a crucial part of lab life.

Monya Baker

27 January 2016

PDF | Rights & Permissions

Chris Ryan/Nature

There are at least six things in this picture that a quality-assurance manager would try to improve. Can you spot them? (Answers, below)

https://www.nature.com/news/how-quality-control-could-save-your-science-1.19223

Quality control- Data collection

Outline the number of measurements/samples/procedures repeated.

Outline instrument calibration tests & data set or samples used for calibration.

Outline standardized controls (e.g., sample controls).

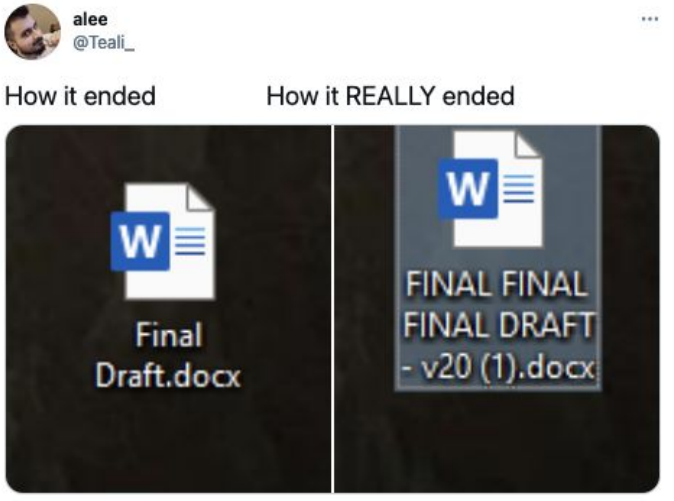
Use of standardized protocols and methods with clear instructions and documentation.

Quality control- Data entry

- Decide a method for documentation i.e., Electronic lab notebooks vs paper.
- Outline the non-digital data structure and strategy for digitization.
- Collect and create metadata throughout the data collection and handling process.
- Use controlled vocabularies.
- Outline how the data/samples/variables are labelled.
- Document terminology used.
- Describe how to flag/tag questionable data.
- Ensure data and time is represented in a machine-readable format and valid.
- Set up validation rules or input masks in data entry software.

Versioning

A version control strategy will allow you to easily detect the most current/final version, organize, manage and record any edits made while working on the document/data, drafting, editing and analysis.



Find out more

- RDM 1 day workshop https://zenodo.org/record/4562630
- FAIR4Software https://zenodo.org/record/6574092
- Turing way https://the-turing-way.netlify.app/welcome

Thank you!

Get in touch
Email: sara.elgebali@scilifelab.uu.se
Twitter: @yalahowy

Resources- Data Organization:

- Towards a Standardized Research Folder Structure
- Swedish National Data service
- DataOne research data management modules
- Imperial College Research data management guides
- King’s college- Managing your data
- UK Data Services