

Data Management Plan

University of Pisa - Phd course 2022

Gina Pavone, CNR  0000-0003-0087-2151

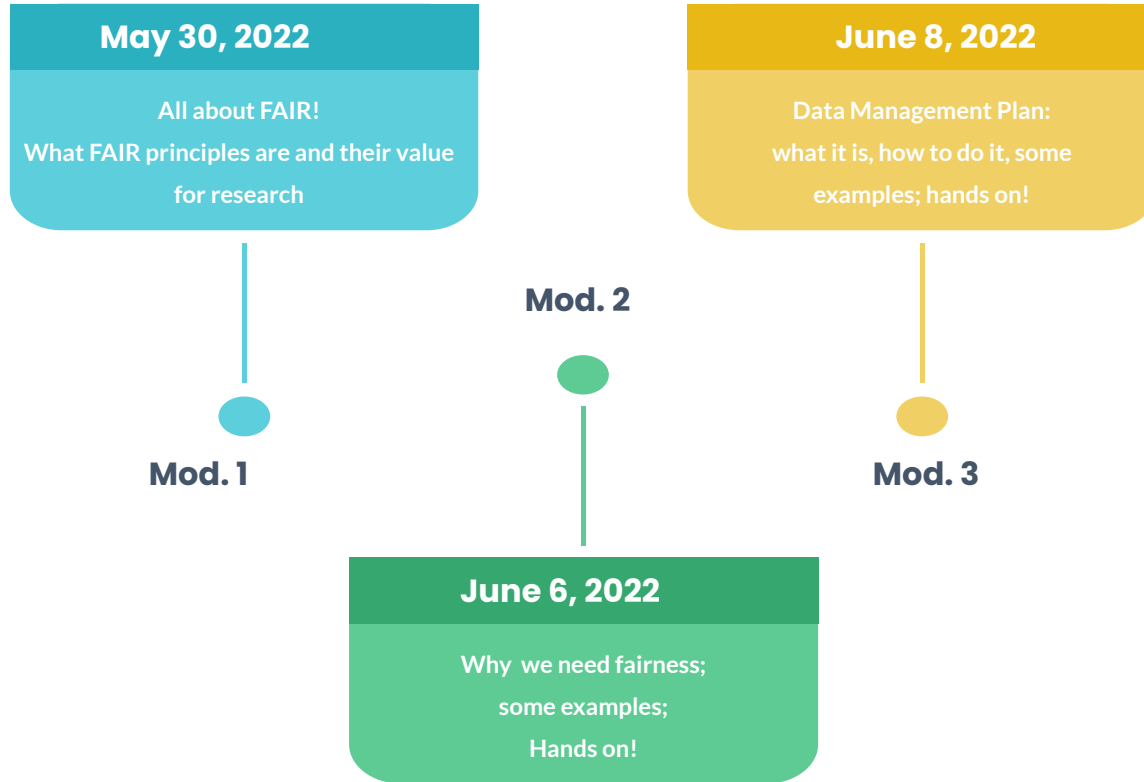
Day 3 - 8 June 2022

10.5281/zenodo.6597239



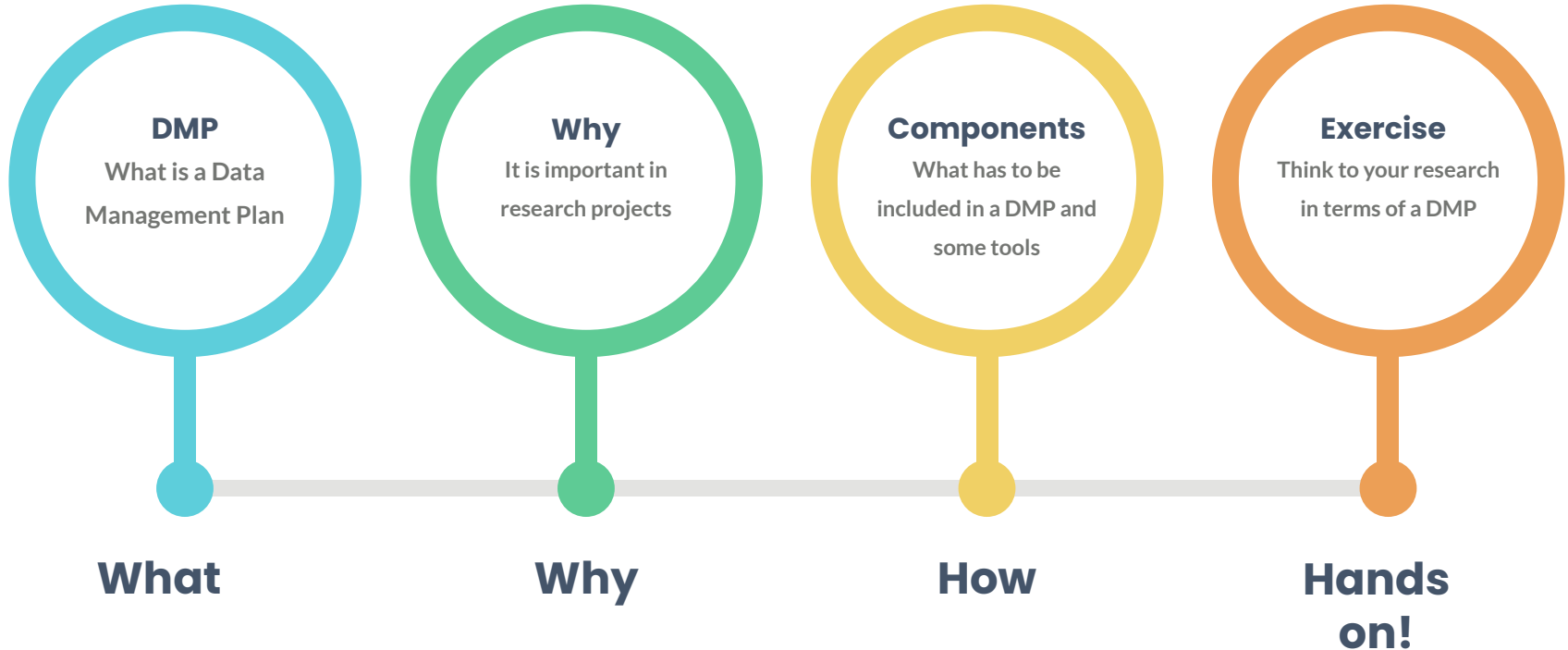
FAIR data and DMP

COURSE OUTLINE



Today's agenda

University of Pisa
PhD students



Zenodo: Licence (other)

🌟 License *

Creative Commons Attribution 4.0 International

Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your files under different licenses, please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your record and choose one of the *Other* licenses available (*Other (Open)*, *Other (Attribution)*, etc.). The supported licenses in the list are harvested from opendefinition.org and spdx.org. If you think that a license is missing from the list, please [contact us](#).

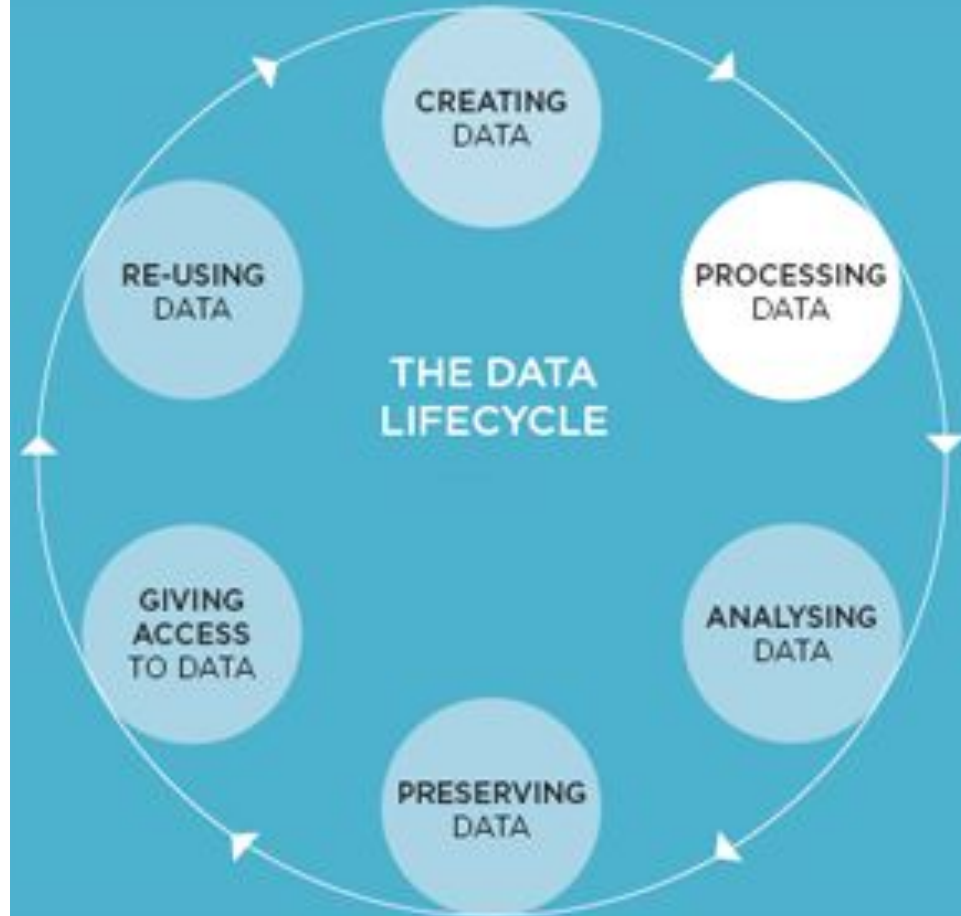
■ specpose-stopeight-e0af952	
○ ■ .github	
○ ■ workflows	
○ ■ conda-build.yml	2.6 kB
○ ■ pypi-sdist.yml	822 Bytes
○ .gitignore	17 Bytes
○ .gitmodules	195 Bytes
○ CMakeLists.txt	352 Bytes
○ LICENSE.txt	15.2 kB
○ MANIFEST.in	934 Bytes
○ README.md	186 Bytes
○ cmake.py	4.0 kB
○ doc	

<https://zenodo.org/record/4174214>

<https://github.com/specpose/stopeight/blob/0.3.15/LICENSE.txt>

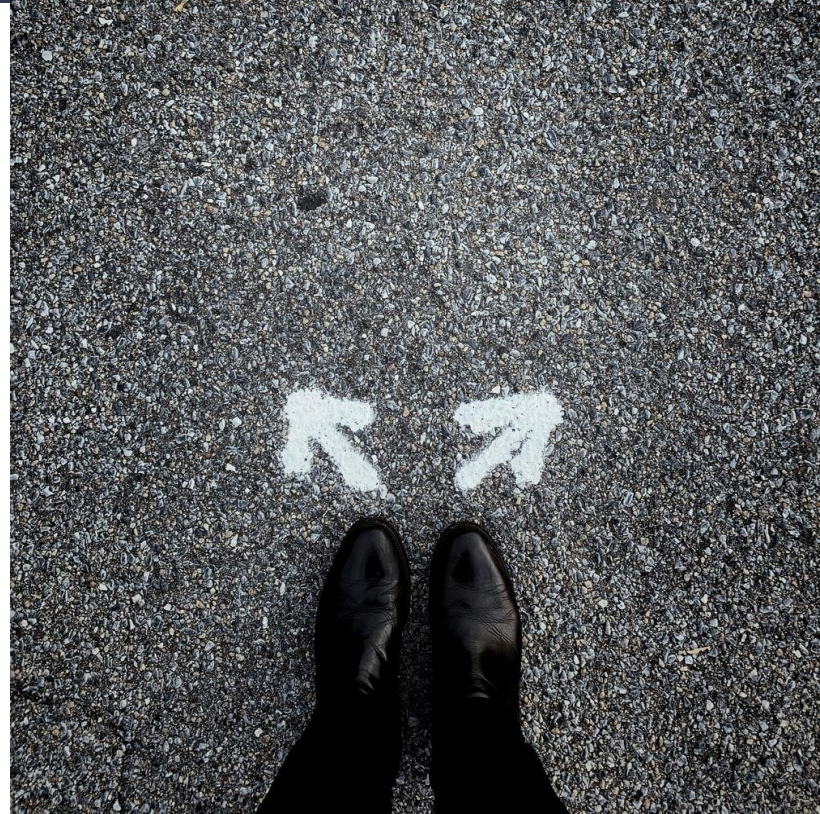
RDM practices encompass activities in every stage of the work with data

Before, during and after the research project. Choices made in one stage influence the next one.



Choices in each research phase

The different stages that your data travels through (raw, cleaned up, processed, analysed data) involve their own data management challenges.



RDM benefits

For researchers

- More visibility and citations
- Opportunity for collaboration
- Career recognition
- Decrease of non-compliance risks (legal, ethical, institutional and funders' policies)

For science

- Facilitates data finding and reuse
- Enables new research and new insights on the data
- Protection of valuable data
- Supports research integrity and reproducibility

For society

- Efficient use of public resources
- Better quality research can benefit to better decision-making
- Opportunities for citizen science
- Increased transparency and trust in science

The data Management plan



What is a DMP?

A Key tool for proper Research Data Management

A Data Management Plan is a document specifying how research data will be handled both during and after a research project.

It identifies key actions and strategies to ensure that research data are of a high-quality, secure, sustainable, and – to the extent possible – accessible and reusable.

quoted from:

<https://www.ugent.be/en/research/datamanagement/before-research/datamanagementplan.htm>

Why does DMP has to be done?

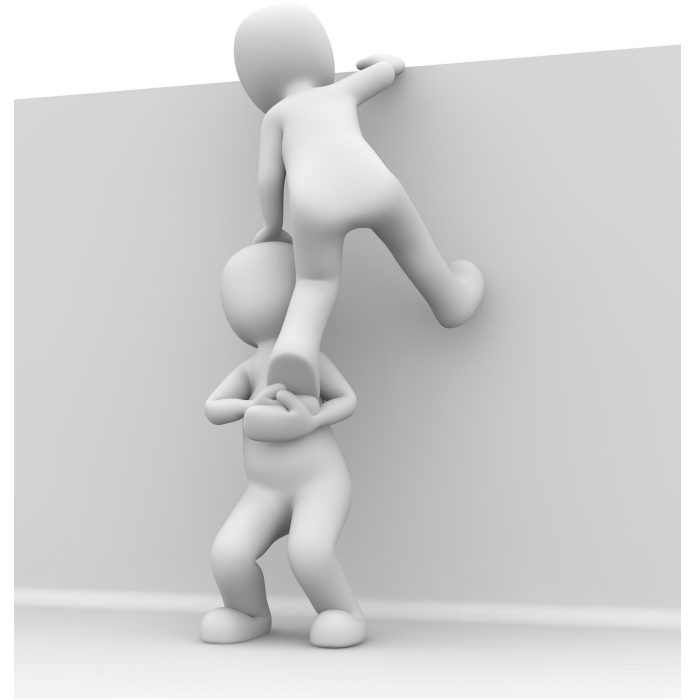
- For oneself!

The DMP is a structured approach to data management: instead of improvising when a need arises, thoughtful choices are made across the entire data lifecycle.

- Save time and try to prevent problems in the future
- Estimate costs

See also:

https://open-science.it/article?rpk=101032&prs_sel=p_researcher&tpc_sel=t_gestione_dei_dati_della_ricerca (in Italian)



Why does DMP has to be done?

For others and for mandates!

Mandates

- It is mandatory in EC and ERC funded projects
- Also other funders ask for it
- RPOs may have their own policy on RDM

Ethics

- DMPs may also be required as part of the ethical approval process

GDPR

- Even if a full DMP is not required, a record of processing activities is needed to comply with the GDPR when working with personal data.

DMP benefits



Good time investimen!

“The time invested in setting up a good data management strategy pays off when the time comes to reproduce your analysis and results.

You will be able to easily find and understand your data, increase your data's reuse potential and comply with funder mandates at the same time.”



Think ahead

- It makes you aware of possible problems at an early stage so that you can work around them. E.g. it reminds you to gain consent for future reuse and sharing from research participants.
- By thinking early about various aspects of data management, you can ensure that the material is well-managed already during the data collection period.



Allows for easy project management

- An important function of a DMP is to work as a one-stop shop to find project-related information.
- questions surrounding data management are being gathered in one place and project-related details are readily available rather than just vaguely remembered or simply forgotten.



Clarifies needed budget

- Data management is not free. Therefore, an important aspect of a DMP is its use in calculating how much money will be required for managing your research data during your research project.
 - Time and resources (money and expertise) for collecting, analysing, and publishing on data,
 - Time and resources for careful documentation as well as server space, backup solutions, and documentation software

Make data FAIR

A DMP allows you to think through beforehand how to provide a dataset to a data repository which is as FAIR as possible. A DMP:

- Makes structuring and documenting of your datasets simpler, thus making it easier for others as well as your future self to find and understand the material;
- Encourages you to think about the data format which is best suited for reuse;
- Allows you to think about the reuse license you would want to apply to your data;
- Etc.

Shows accountability

If you draw up a DMP, you are showing your affiliated institution, funders and project partners a serious approach to research data management, that includes a responsible approach towards research funds and research participants.



When the DMP has to be done?

Before or at early stage of the research activity



The DMP is a living document!



Update it where necessary in the course of the project!

You may not know all the answers at the outset, and circumstances may change.

What is normally a DMP about?



Identify the data you are working with in your project

Decide the strategy to organise your data and the standards you will use

Daily data management

- What is your plan for sharing your data?
- Will you have issues sharing your data?
- Will you need more resources/budget than expected?



Topics and aspects to address in a DMP

Data collection

Considering what data will be collected, generated, and/or reused, and how you will organize them

Research data can be gathered through observation, manual or automatic measurements in the laboratory or in the field, with remote sensing techniques, by interviews, by modelling and simulation, etc. Data can also be stored in many formats.

Types of data

Data can be described in many ways. Such as:

Primary or secondary

Source: registers, databases, types of media etc

Physical formats: numerical, textual, still image, geospatial, audio, video and software

How they are created: electronic text documents, spreadsheets, laboratory notebooks, field notebooks and diaries, questionnaires, transcripts and codebooks, audiotapes and videotapes, photographs and films, examination results, slides, algorithms...

Types of big data

Depending on their source, the [OECD](#) defines six categories of Big Data:

A: Data stemming from the transactions of government, for example, tax and social security systems.

B: Data describing official registration or licensing requirements.

C: Commercial transactions made by individuals and organisations.

D: Internet data, deriving from search and social networking activities.

E: Tracking data, monitoring the movement of individuals or physical objects subject to movement by humans.

F: Image data, particularly aerial and satellite images but including land-based video images.

D: [Social media data](#), from platforms like Facebook, Twitter, Instagram or YouTube. These data are created by the users of such platforms. Researchers can access these data in three main ways: 1) Direct cooperation with the companies/platforms, 2) Buying from data resellers, 3) Via APIs (one might add web scraping to the list but most platforms/companies discourage its use).

Data in social sciences: personal data

Notably, within the field of social sciences, you will often work with data originating from **human participants**. This can mean that you are handling (sensitive) personal data, which deserve special attention.

Personal data (GDPR) any information relating to an identified or identifiable natural person known as 'a data subject'. It is further specified that an identifiable natural person is someone who can be identified, either directly or indirectly, by reference to an **identifier such as a name, an identification number, location data, an online identifier** or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Personal data can include a variety of information, such as **names, addresses, phone numbers and IP addresses**.

The GDPR applies only to the data of living persons, but for other types of data there might still be ethical reasons for protecting this information.

Data in social sciences: sensitive data

Data that may create important risks for the fundamental rights and freedoms of the involved individual.

Within the GDPR the following categories are defined as 'special categories of personal data':

Racial or ethnic origin;

Political opinions;

Religious or philosophical beliefs;

Trade union membership;

Genetic data;

Biometric data;

Data concerning health;

Data concerning a natural person's sex life or sexual orientation.

Organize: file formats

File format:

- how information is stored within a digital file
- the format of a file is indicated by the 'extension' in the filename (e.g. .txt, .csv)

The choice of file formats to use depends on:

- Discipline-specific standards and customs
- Planned data analyses
- Software availability/cost
- Hardware used

Risks

- Formats which can only be used within specific software makes the digital data vulnerable to obsolescence of the software
- Beware to file converting!

Best practices:

- Non-proprietary (not protected by trademark, patent or copyright)
- Open, documented standard
- Common usage by research community
- Standard representation (ASCII, Unicode)

Data collection: file naming

A file name is the principal identifier of a file.

Good file names should:

- Provide useful cues to content, status and version
- Uniquely identify a file
- Help to classify and sort files

File names can be constructed using the following elements:

- Project acronym
- Content description
- Date
- Location
- Creator name/initials
- Status information (i.e. draft or final) etc

File naming: decide with your colleagues!



It can be useful if the consortium/department/group agrees on the following elements of a file name:

- **Vocabulary** – choose a standard vocabulary for file names, so that everyone uses a common language
- **Punctuation** – decide on conventions for if and when to use punctuation symbols, capitals, hyphens and spaces
- **Dates** – agree on a logical use of dates so that they display chronologically i.e. YYYY-MM-DD
- **Order** - confirm which element should go first, so that files on the same theme are listed together and can therefore be found easily
- **Numbers** – specify the amount of digits that will be used in numbering so that files are listed numerically e.g. 01, 002, etc.

Data collection: folder structure

Information on a topic is located in one place

Are there established approaches in your team or department?

Name folders appropriately - i.e. name folders after the areas of work to which they relate

Structure folders hierarchically - limited number of folders for the broader topics, and then create more specific folders within these

Consider separating ongoing and completed work

Backup – ensure that your files, whether they are on your local drive, or on a network drive, are backed up

Documentation

One of the basics of RDM. It enables you to understand/interpret data later

Study level documentation:

Contextual information (the background, aims, objectives, the hypotheses etc)

Procedural & methodological information

Data level documentation:

Information about datasets and/or individual data items

Information about variables, derived data, aggregated data etc

Legal and ethical issues

Especially when the research is handling personal or sensible data

Do you have to protect confidentiality?

Are there requirements from an ethical committee?

Have you an informed consent to distribute and get signed?

Have you an idea of the anonymization tool to use?

Storage and backup

Data should be safe!

Provide information on the type of storage is going to be used for the project data (storage on local devices, network storage - i.e. in your institution - or cloud storage)

Specify in which phase you want to use one type or another of storage

A storage question

I have terabytes of videotaped interviews from a European project, dozens of pseudonymised transcripts and informed consent forms. European partners need access to the files for data analysis. What's the best storage strategy for me?

⊖ A possible storage solution

Type of data	Storage needs	Storage solution
<i>The data which were collected are personal data.</i>	<i>High storage capacity for videos required;</i>	<i>Data are transmitted only in encrypted form. (see Security)</i>
<i>Extra security measures to protect it should be in place (see Security).</i>	<i>Remote access to videos and transcripts required;</i>	<i>Data for remote access is stored in cloud storage in Europe. (see Storage)</i>
	<i>Researchers need to work on the same files simultaneously.</i>	<i>Master copies of videos and transcripts are encrypted and backed up in the cloud and on portable hard disk and flash drives. (see Security)</i>
		<i>Backups locked away in different, secure locations. (see Backup)</i>
		<i>Consent forms and encryption keys are stored in a secure safe.</i>

Storage

Questions:

- How much storage space do I need?
- Who needs access?
- What precautions should I take to protect my data against loss?
- Which storage solutions are suitable for personal data?

Technologies:

- Portable devices: Laptops, tablets, external hard-drives, flash drives and Compact Discs
- Local storage: Desktop computers
- Cloud storage: E.g. Google Drive, OneDrive, Dropbox, a University's OwnCloud, Open Science Framework
- Networked drives: Shared drives on university servers

Back up

Backups are an important instrument to ensure that data and related files can be restored in case of loss or damage.

Among the most common causes of data loss are:

- Hardware failure;
- Software malfunction;
- Malware or hacking;
- Human error (research data accidentally gets deleted or overwritten or is lost in transport);
- Theft, natural disaster or fire;
- Degradation of storage media

Create your backup strategy

- Find out whether your institution has a backup strategy
- Determine what you want to backup
- Decide where backups will be stored
- Determine how much storage capacity will be needed
- Determine if there are tools you could use to automate backup
- Determine how long backups will be kept and how they will be destroyed
- Determine how personal data will be protected
- Devise a disaster recovery plan
- Assign responsibilities
- Determine how to check the integrity of backed-up files

Security

Why:

“To prevent unauthorised access and possible changes to your data, data security measures are in order. Such measures, on the one hand, serve to protect personal data and confidential information and on the other hand offer protection against unauthorised manipulation or erasure of files (intentional or unintentional).”

You need to arrange technical solutions and organizational measures

Possible solutions:

Passwords to lock the computer systems used to access these data files

Encryption: the process of encoding digital information in such a way that only authorised parties can view it. (there are many specific softwares)

Up-to-date virus scanners and firewalls.

Secure disposal (ie. use of software for secure erasing)

Protection

Reflect on on key legal and ethical considerations in creating shareable data.

Uphold to scientific standards

Be compliant with the law and check requirements by your institution ethical committee

Avoid social and personal harm

Tools:

Informed consent

Ethical assessments

Anonymization (software) and pseudo-anonymization

Check if data are protected by the law (confidential data, copyright and so on)

Identifiers

Direct identifiers are ones like the participant's name, address, or telephone numbers that specifically identify them;

Indirect identifiers are ones that when they are placed with other information could also reveal an individual, for example, by cross-referencing occupation, salary, age, and location.



Anonymisation and pseudonymisation

Anonymisation

irreversibly destroys any way of identifying the data subject

Anonymous data is data that cannot identify individuals in the dataset in any way. Neither directly through name or social security number, indirectly through background variables, nor through a list of names or through an encryption formula and code/scrambling key.

When anonymising, data identifiers need to be removed, generalised, aggregated or distorted.

Pseudonymisation

allows to re-identify the data subject with additional information.

The GDPR defines pseudonimisation as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information". To pseudonymise a dataset "the additional information must be kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person". Directly identifying data is held separately and securely from processed data to ensure non-attribution.

Data preservation

Keeping data available and usable in the longer term, beyond the end of your research project

Specify if data will be selected for deposit (ie. raw data, processed data...) - you can update this part later!

Detail on the Research Data repository

Detail on non-digital data and materials

Costing

Both in time and money!

Infrastructure costs:

- Digitisation
- Storage
- Licensing and Security
- Sharing and Re-use
- Archiving

Skill costs:

- Data wrangling
- Description and Documentation
- Metadata generation
- Formatting and Cleaning
- Consent and Anonymisation

✓ How much could management & deposit cost?

Some factors that affect RDM costs...



Security of potentially sensitive data



Dataset size



Length of preservation required



Remember:

Different repositories apply different charging models. Some apply a fixed-fee per data package plus an amount over a certain volume, while others only apply variable fees depending on the data volume. Some may not charge at all.

Guide on costs by Utrecht University



Search uu.nl



Nederlands

[Home](#) > [Research](#) > [Research Data Management Support](#) > [Guides](#) > [Costs of data management](#)



Research Data Management Support

[Home](#) **Guides** [Tools & Services](#) [Walk-in hours & Workshops](#) [RDM Projects & Stories](#) [FAQ](#) [Contact us](#) [About](#) [Index](#)

Guides

- > [Working safely with research data from home](#)
- > [Data management planning](#)

Costs of data management

To help you estimate the costs of data management an overview of possible costs per research phase and research activity is presented.

Data sharing

Make publicly available (outside project or research team)

Details on access: if data will be open, restricted, closed or embargoed

Details and motivation for non-open data

Information on access conditions (if restricted - ie. email, form or whatever)

Use of persistent identifiers

Information on licences

HAVE

A

BREAK



Data Management Plan Checklist

One of the million available...

To do:

- Wake up
- Make coffee
- Drink coffee
- Make more coffee



1. Administrative Data

Basic information about your project: title, acronym, ID, reference numbers

An abstract of your project highlighting the scope of data collection/creation

Details related to procedures and policies



Photo by [Wesley Tingey](#) on [Unsplash](#)

2. Data Collection



Are you using existing data?

What are the standards or methodologies that you will use to collect your data?

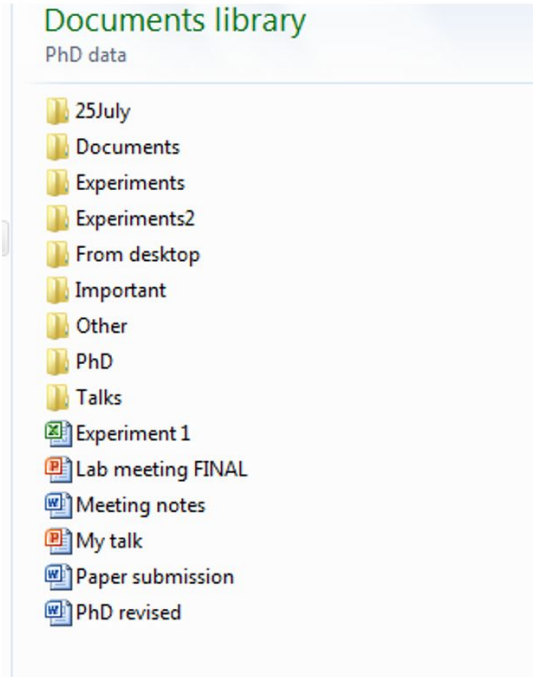
Types of data and scope

Data Formats and Software

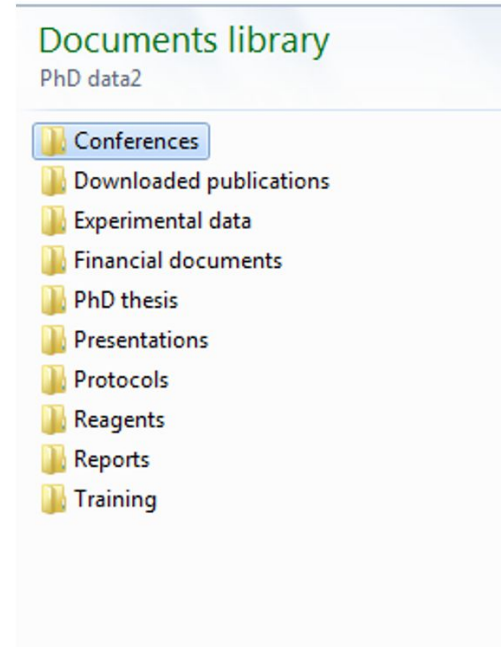
How will you structure and name your files and folders?

Folder structure: what is your strategy?

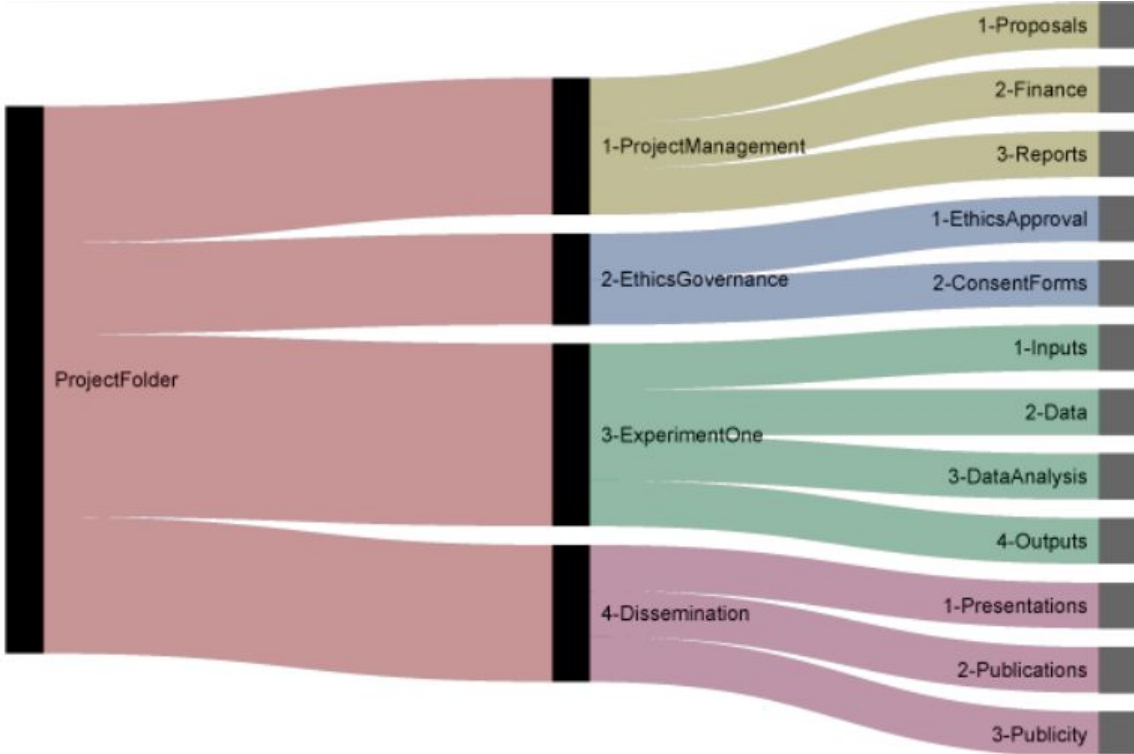
Example A



Example B



A good example



File naming



A good example

http://www.data.cam.ac.uk/files/gdl_tilSDocNaming_v1_20090612.pdf

3. Version

(upper case, max 4 chars, optional)

For documents that will continue in various versions use V followed by the version number. Use an underscore to indicate a decimal point if necessary.

Eg. PMF_PRP_ZenMonkeyProject_V2_20090607.docx

New versions should not be created for each iteration of the document, but rather at significant changes or when it has been reviewed or changed by another author.

Document naming for the TILS Division should follow this convention:

GDL_TILSDocNaming_V1_20090612.docx

A prefix shows the document type

The document title describes the content

The version number

The date in the format yyyymmdd

Prefix	Meaning
AGD	Agenda
AGR	Agreement
GDL	Guideline
MEM	Memorandum
MIN	Minutes and Notes
PRE	Presentation
PRO	Procedure
PRP	Proposal
REP	Report
TEM	Template

2. Document title/ Description

(mixed case, max 30 chars, **no spaces**)

- Describes the purpose or “business” of the document. Acronyms, capitalisations, abbreviations can be used, keep in mind that descriptions should be **meaningful** to anyone reading the file name.
- In the case of project documentation use the **project name** or its usual abbreviation
- If possible Departmental Branch and/or Section should be integrated into this field to indicate origin / ownership of document.
- Use only alpha-numeric characters, plus the hyphen and underscore.
- **Do not use spaces.**

3. Metadata and Supporting Documentation

Which documentation and metadata will support/describe your data?

How will you create the supporting documentation and metadata?

Which metadata standards will you use?

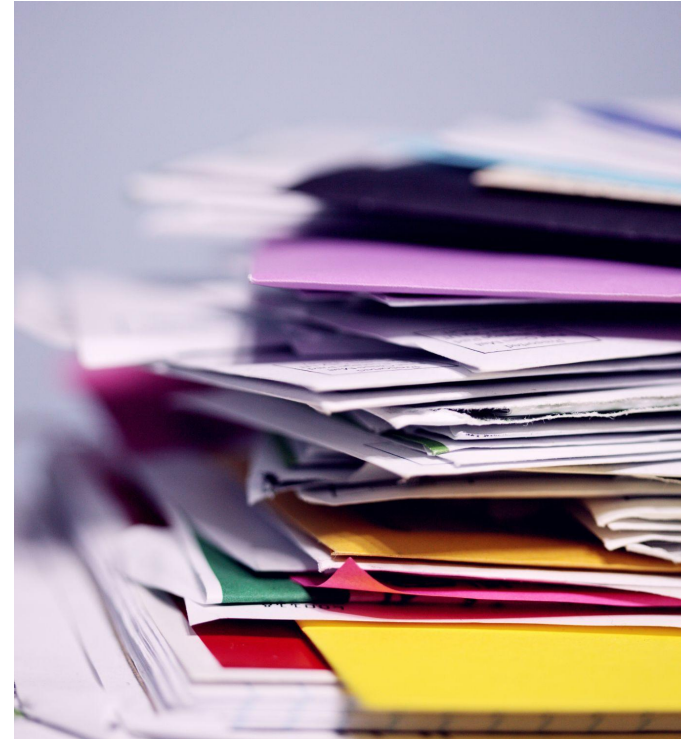


Photo by [Sharon McCutcheon](#) on [Unsplash](#)

So many ways to describe your data

How to create useful README files: <https://data.research.cornell.edu/content/readme>



```
Cornell AUTHOR_DATASET_ReadmeTemplate.txt

This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

-----
GENERAL INFORMATION
-----

1. Title of Dataset

2. Author Information

Principal Investigator Contact Information
  Name:
  Institution:
  Address:
  Email:
```

A readme file describes your data

Use a readme file for those data type that do not have a metadata standard available

README files template:

<https://cornell.app.box.com/v/ReadmeTemplate>

4. Legal and Ethical Aspects



Did you ask for an informed consent to share the data and preserve them?

How will you protect personal data?

How about data licencing?

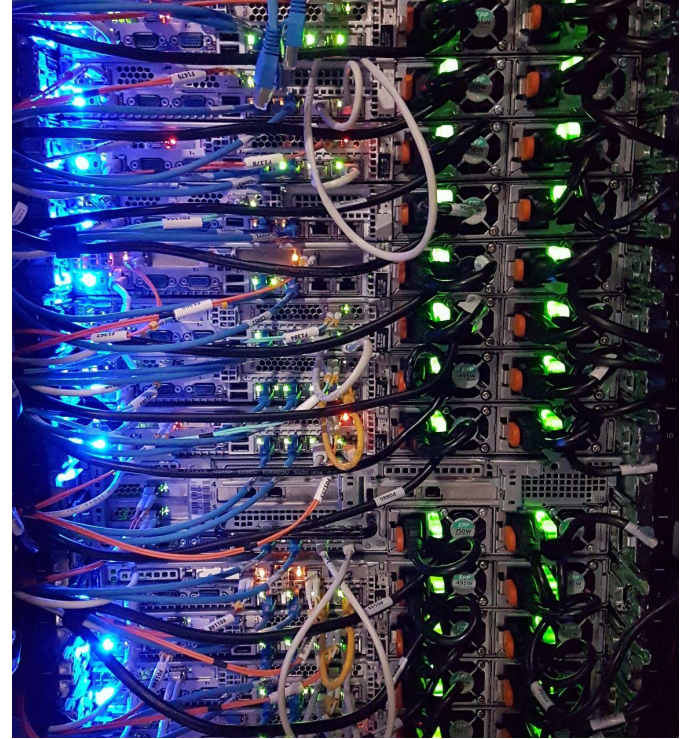
5. Data Storage and Backup

Do you have enough space to store your data or should you include costs for additional services?
Will the services be reliable/trusted?

How will you share your storage/backup with your collaborators?

Will you use cloud solutions?

Will you back your data up? How?



Do not leave it all to Google

Google services Terms of Use:

When you upload, submit, store, send or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content. The rights you grant in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones. This license continues even if you stop using our Services (for example, for a business listing you have added to

<https://policies.google.com/terms?hl=en>

You have better alternatives...



Consortium GARR

Infrastrutture Comunità Servizi Ricerca e formazione

INFRASTRUTTURA CLOUD

Infrastrutture / Infrastruttura cloud / Infrastruttura cloud

GARR, OPEN SOURCE, COSA ABBIAMO, COSA OFFRIAMO, RETE, CLOUD

INFRASTRUTTURA CLOUD

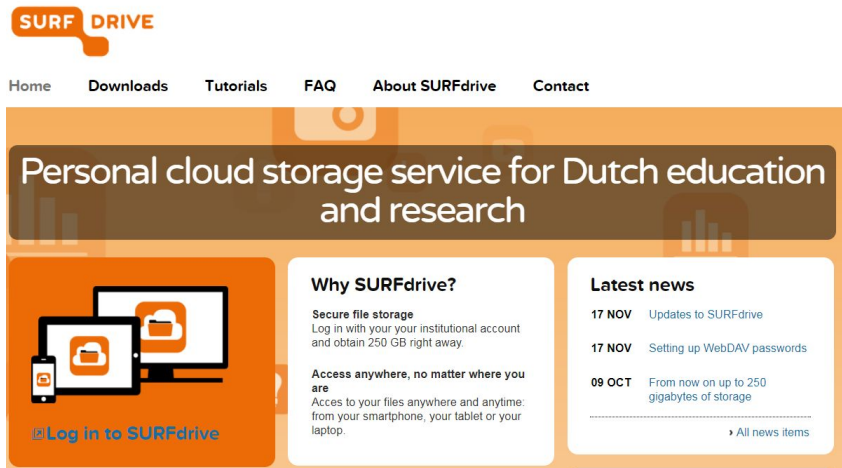
GARR affianca alla rete ad alte prestazioni un'infrastruttura per il calcolo e l'archiviazione costruita secondo il paradigma cloud. Su questa infrastruttura è stata realizzata la **piattaforma Cloud GARR**.

Qui, la comunità nazionale della ricerca e dell'istruzione può utilizzare risorse **condivise e flessibili in base alle esigenze**, riducendo i costi senza rinunciare alla **qualità dei servizi**, con garanzie di **sicurezza e confidenzialità dei dati**, ed offre una totale **indipendenza da lock-in** con fornitori di servizi di cloud commerciali. Semplicità, scalabilità ed economicità sono tra i principali benefici.

La piattaforma Cloud GARR offre attualmente tre tipologie di servizi: Macchine virtuali, Virtual Datacentre e Applicazioni Cloud in modalità PaaS.

È inoltre allo studio l'offerta di un'innovativa piattaforma di Container.

<https://www.garr.it/it/infrastrutture/infrastruttura-cloud/infrastruttura-cloud>



SURF DRIVE

Home Downloads Tutorials FAQ About SURFdrive Contact

Personal cloud storage service for Dutch education and research

Why SURFdrive?

- Secure file storage**
Log in with your institutional account and obtain 250 GB right away.
- Access anywhere, no matter where you are**
Access to your files anywhere and anytime: from your smartphone, your tablet or your laptop.

[Log in to SURFdrive](#)

Latest news

- 17 NOV** Updates to SURFdrive
- 17 NOV** Setting up WebDAV passwords
- 09 OCT** From now on up to 250 gigabytes of storage

[All news items](#)

Your institution probably provides better alternatives.
Ask your IT for support!

6. Select and Preserve



Which data shall be preserved or destroyed due to contractual, legal or administrative reason?

What are the envisaged uses of your data for research purposes?

Which data shall be preserved and potentially shared?

What is your long term preservation strategy?

Did you consider the effort and costs to prepare your data for sharing and preservation?

7. Data Sharing

Who will you share your data with? Under which conditions?

When will you share your data?

Will you need to apply any access restriction?

Which actions do you foresee to avoid or reduce access restrictions?

How will your potential users find your data?



Funder policies

Sherpa Juliet

Browse

Search

Statistics

Our APIs

Suggest

Admin

Search

Please enter a name or acronym of a funder.

Funder Name

Search

This quick search will find any items whose name or acronym (in any language) match any of the words entered.

<https://v2.sherpa.ac.uk/juliet/>

8. Responsibilities and Resources



Who is responsible to implement and revise the DMP?

How will you share responsibilities among partners in collaborative projects?

Which resources will you need to implement your DMP?

Will you need any specific external expertise or tools?

Cost assessment

The costs related to open access to research data are eligible as part of the Horizon Europe grant - refer to your Grant Agreement conditions. For further information, see [OpenAIRE guide on cost in research data management](#)



What the DMP in HE should include

From Horizon Europe Data Management Plan Template, version 1 , 5 May 2021

[https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan-template he en.docx](https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan-template_he_en.docx)

Data summary

Will you re-use any existing data and what will you re-use it for?

- You can state the reasons if re-use of any existing data has been considered but discarded.

What types and **formats** of data will the project generate or re-use?

What is the **purpose** of the data generation or re-use and its relation to the objectives of the project?

What is the **expected size** of the data that you intend to generate or re-use?

What is the **origin/provenance** of the data, either generated or re-used?

To whom might your data be useful ('data utility'), outside your project?

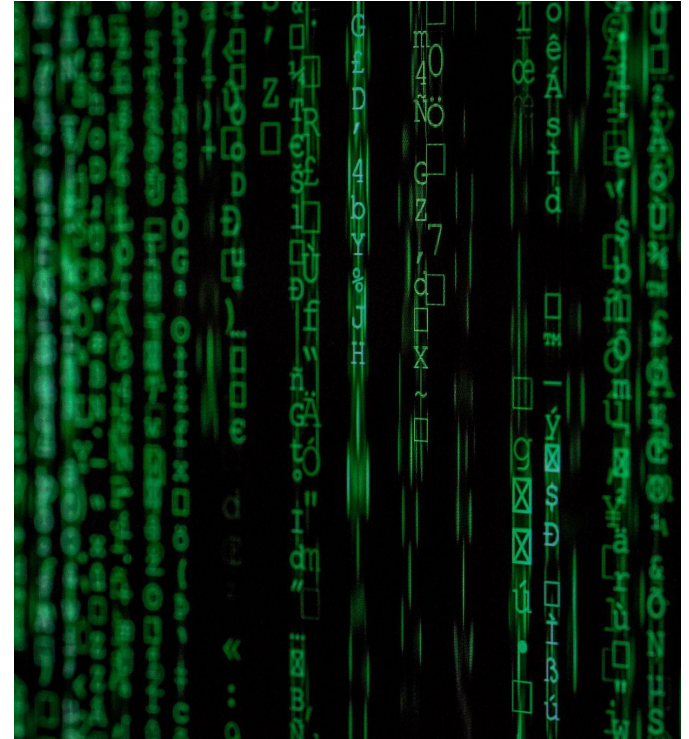


Photo by [Markus Spiske](#) on [Unsplash](#)

FAIR data – Findability

Making data findable, including provisions for metadata

Will data be identified by a **persistent identifier**?

Will rich **metadata** be provided to allow discovery?

What metadata will be created? What disciplinary or general **standards** will be followed?

- In case metadata standards do not exist in your discipline, you outline what type of metadata will be created and how.

Will search **keywords** be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be **harvested and indexed**?

FAIR data - Accessibility part 1

REPOSITORY

Will the data be deposited in a **trusted repository**?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an **identifier**? Will the repository resolve the identifier to a digital object?

METADATA

Will metadata be made openly **available** and licenced under a public domain dedication CC0, as per the Grant Agreement?

- If not, clarify why!

Will metadata contain information to enable the user to **access** the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will **documentation** or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

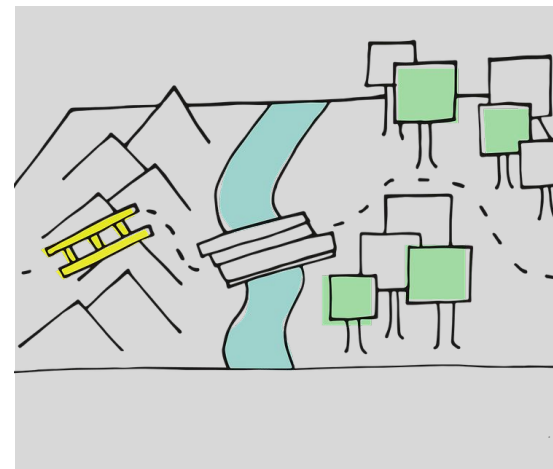


Image by [Manfred Steger](#) from [Pixabay](#)

FAIR data - Accessibility part 2

DATA

Will all data be made **openly available**?

- If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

If an **embargo** is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized **access protocol**?

If there are **restrictions** on use, how will access be provided to the data, both during and after the end of the project?

How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

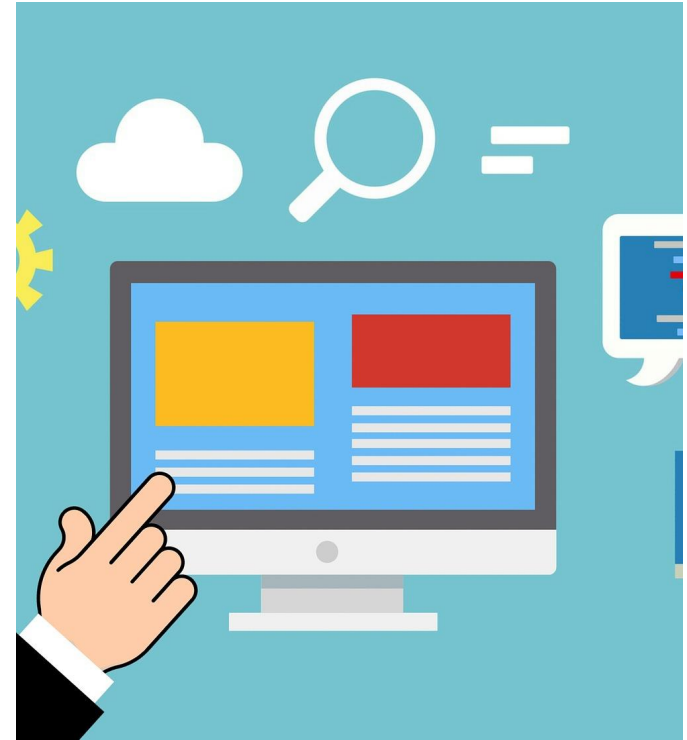


Image by [mohamed Hassan](#) from [Pixabay](#)

FAIR data – Interoperability

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data **interoperable to allow data exchange** and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide **mappings** to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified **references** to other data (e.g. other data from your project, or datasets from previous research)?

FAIR data - reusability

How will you provide **documentation** needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will your data be made freely available in the **public domain** to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Will the data produced in the project be useable by third parties, in particular after the **end of the project**?

Will the **provenance** of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes.

Further to the FAIR principles, DMPs should also address **research outputs other than data**, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.



Other research outputs



Image by [Phe Schlay](#) from [Pixabay](#)

Consider and plan for the management of other research outputs that may be generated or re-used throughout the projects. Be they either **digital** (e.g. software, workflows, protocols, models, etc.) or **physical** (e.g. new materials, antibodies, reagents, samples, etc.).

Consider which of the FAIR principles can apply to the management of other research outputs.

Strive to provide sufficient detail on how their research outputs will be managed and **shared**, or made available for **re-use**, in line with the FAIR principles.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to **storage, archiving, re-use, security**, etc.) ?

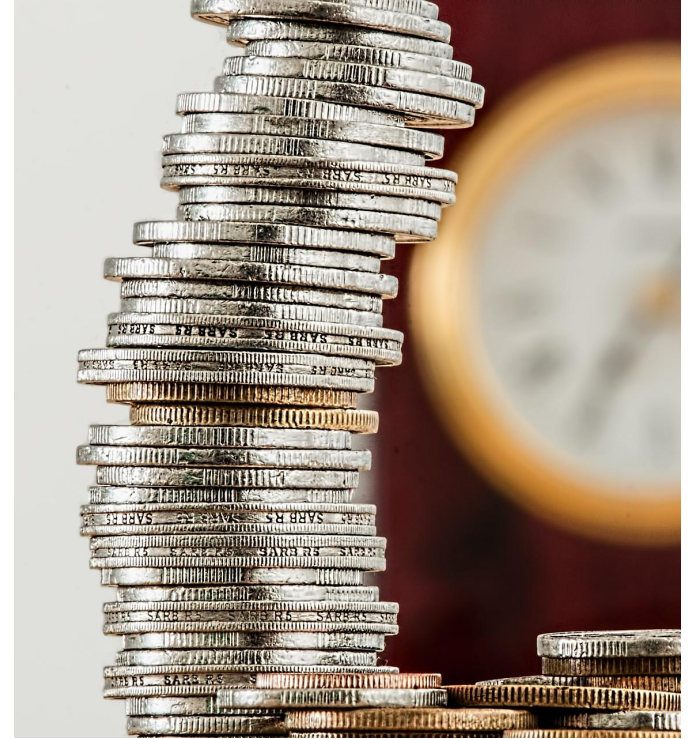
How will these be **covered**?

- Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

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

How will **long term preservation** be ensured?

- Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?



Data Security, ethics and other issues

Security

What provisions are or will be in place for data security (including data **recovery** as well as secure **storage/archiving** and transfer of **sensitive** data)?

Will the data be safely stored in **trusted repositories** for long term preservation and curation?

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Will **informed consent** for data sharing and long term preservation be included in questionnaires dealing with personal data?

Other issues

Do you, or will you, make use of **other national/funder/sectorial/departmental procedures** for data management? If yes, which ones (please list and briefly describe them)?

DMP in HE: when?

Proposal stage: concept of FAIR data management and draft of future DMP - recommended

Approved project: Beneficiaries must submit a DMP as a deliverable to the granting authority in accordance with the Grant Agreement (normally by month 6) - mandatory

see [HE Annotated model grant agreement, annex 5](#)



Photo by [Brands&People](#) on [Unsplash](#)

Mandatory updates

An updated DMP deliverable must also be produced **mid-project** (for projects longer than twelve months) and at the **end of the project** (where relevant). - see [HE Annotated model grant agreement, annex 5](#)

Image by [Gerd Altmann](#) from [Pixabay](#)



Beware!



The project will be evaluated also on the capability of properly managing data!

HE, DMP best practices

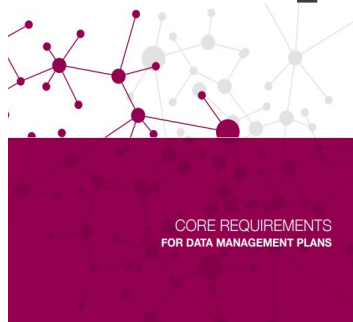
Beneficiaries should maintain the DMP as a living document and **update it over the course of the project whenever significant changes arise**. I.e.: the generation of new data, changes in data access provisions or curation policies, attainment of tasks (e.g. datasets deposited in a repository, etc.), changes in relevant practices (e.g. new innovation potential, decision to file for a patent), changes in consortium composition.

Beneficiaries are encouraged to encode their DMP deliverables as **non-restricted, public deliverables**, unless there are reasons (legitimate interests or other constraints) not to do so. In the case they are made public, it is also recommended that open access is provided under a CC BY licence to allow a broad re-use.



Resources and tools for DMP

Core Requirements



CORE REQUIREMENTS FOR DATA MANAGEMENT PLANS



When developing solid data management plans, researchers are required to deal with the following topics and answer the following questions:

- 1. Data description and collection or re-use of existing data**
 - a. How will new data be collected or produced and/or how will existing data be re-used?
 - b. What data (for example the kinds, formats, and volumes) will be collected or produced?

- 2. Documentation and data quality**
 - a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?
 - b. What data quality control measures will be used?

- 3. Storage and backup during the research process**
 - a. How will data and metadata be stored and backed up during the research process?
 - b. How will data security and protection of sensitive data be taken care of during the research?

- 4. Legal and ethical requirements, codes of conduct**
 - a. If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
 - b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
 - c. How will possible ethical issues be taken into account, and codes of conduct followed?

- 5. Data sharing and long-term preservation**
 - a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?
 - b. How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?
 - c. What methods or software tools will be needed to access and use the data?
 - d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

- 6. Data management responsibilities and resources**
 - a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?
 - b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

CESSDA DMP Expert Guide

PLAN

Overview

Title of the project

Date of this plan

Description of the project

- What is the nature of the project?
- What is the research question?
- What is the project time line?

Origin of Data

- What kind of data will be used during the project?
- If you are reusing existing data: What is the scope, volume and format? How are different data sources integrated?
- If you are collecting new data can you clarify why this is necessary?

Principal researchers

- Who are the main researchers involved?
- What are their contact details?

Collaborating researchers (if applicable)

- What are their contact details and their roles in the project?

Funder (if applicable)

- If funding is granted, what is the reference number of the funding granted?

Data producer

- Which organisation has the administrative responsibility for the data?

Project data contact

- Who can be contacted about the project after it has finished?

Data owner(s)

- Which organisation(s) own(s) the data?
- If several organisations are involved, which organisation owns what data?

Roles

- Who is responsible for updating the DMP and making sure that it's followed?
- Do project participants have any specific roles?
- What is the project time line?

Costs

- Are there costs you need to consider to buy specific software or hardware?
- Are there costs you need to consider for storage and backup?
- Are potential expenses for (preparing the data for) archiving covered?

ORGANISE & DOCUMENT

Organising and documenting your data

Data collection

- How will the data be collected?
- Is specific software or hardware or staff required?
- Who will be responsible for the data collection?
- During which period will the data be collected?
- Where will the data be collected?

Data organisation

- How will you organise your data?
- Will the data be organised in simple files or more complex databases?
- How will the data quality during the project be ensured?
- If data consists of many different file types (e.g. videos, text, photos), is it possible to structure the data in a logical way?

Data type and size

- What type(s) of data will be collected?
- What is the scope, quantity and format of the material?
- After the project: What is the total amount of data collected (in MB/GB)?

File format

- In what format will your data be?
- Does the format change from the original to the processed/final data?
- Will your (final) data be available in an open format?

Folder structure and names

- How will you structure and name your folders?

File structure and names

- How will you structure and name your files?

Documentation

- What documentation will be created during the different phases of the project?
- How will the documentation be structured?

Metadata

- What metadata will be provided with the collected/ generated/ reused data?
- How will metadata for each object be created?
- Is there any program that can be used to document the data?
- Can metadata be added directly into the files or will the metadata be produced in another program or document?

Metadata standard (if applicable)

- What metadata standard(s) will you use?

cessda

Consortium of European
Social Sciences Data Archives

Adapt your Data Management Plan

A list of Data Management Questions based on the
Expert Tour Guide on Data Management



DCC guides



because good research needs good data

[Home](#) | [Digital curation](#) | [About us](#) | [News](#) | [Events](#) | [Resources](#) | [Training](#) | [Projects](#)

[Home](#) > [Resources](#) > [How Guides](#) > [How Develop Rdm Services](#)

In this section

[How to Develop RDM Services - a guide for HEIs](#)

<https://www.dcc.ac.uk/guidance/how-guides>

<https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep>

Establishing criteria for selection decisions

You should establish criteria to guide selection decisions. The DCC's How to Select and Appraise Research Data for Curation[56] proposes seven criteria as outlined below:

1. **Relevance to mission:** the resource content fits any priorities stated in the institution's mission, or funding body policy including any legal requirement to retain the data beyond its immediate use.
2. **Scientific or historical value:** is the data scientifically, socially, or culturally significant? Assessing this involves inferring anticipated future use, from evidence of current research and educational value.
3. **Uniqueness:** the extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere.
4. **Potential for redistribution:** the reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property or human subjects issues are addressed.
5. **Non-replicability:** it would not be feasible to replicate the data/resource or doing so would not be financially viable.
6. **Economic case:** costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate.
7. **Full documentation:** the information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource's provenance and the context of its creation

DATA MANAGEMENT PLAN CHECKLIST /Griglia per il piano di gestione dei dati

Nel maggio 2017 un gruppo di lavoro informale sui dati della ricerca (costituito da Politecnico di Milano, Università di Milano, Università di Torino, Università di Trento, Università Ca' Foscari Venezia) ha redatto una checklist con una griglia in lingua italiana per l'elaborazione di un Data Management Plan.

ADMINISTRATIVE PLAN DETAILS	Informazioni generali sul progetto di ricerca
Project Name	<i>Inserire il nome del progetto</i>
Acronimo	<i>Inserire l'acronimo del progetto, se applicabile</i>
Grant Reference Number	<i>Inserire il riferimento alla call (es: call Horizon2020 ID:.....) e al numero di identificazione del progetto presentato, se disponibile</i>
Persistent Identifier	<i>handle o DOI del DMP, ricavabile dopo l'inserimento nel repository</i>
Funder	<i>Inserire il nome del finanziatore/dei finanziatori Es: European Commission (H2020)</i>
Principal Investigator/Researcher	<i>Inserire il nome del ricercatore autore del documento Es: Laura Rossi</i>
Principal Researcher ID ORCID	<i>Inserire l'identificativo ORCID del ricercatore Es: 0000-0003-4170-6345</i>

Citing data

Citing data is important in order to:

- Give the data producer appropriate credit
- Allow easier access to the data for repurposing or reuse
- Enable readers to verify your results

Citation Elements

A dataset should be cited formally in an article's reference list, not just informally in the text. Many data repositories and publishers provide explicit instructions for citing their contents. If no citation information is provided, you can still construct a citation following generally agreed-upon guidelines from sources such as the Force 11 Joint Declaration of Data Citation Principles and the current DataCite Metadata Schema.

Core elements

- There are 5 core elements usually included in a dataset citation, with additional elements added as appropriate.
 - **Creator(s)** – may be individuals or organizations
 - **Title**
 - **Publication year** when the dataset was released (may be different from the Access date)
 - **Publisher** – the data center, archive, or repository
 - **Identifier** – a unique public identifier (e.g., an ARK or DOI)
- Creator names in non-Roman scripts should be transliterated using the [ALA-LC Romanization Tables](#).

Common additional elements

- Although the core elements are sufficient in the simplest case – citation to the entirety of a static dataset – additional elements may be needed if you wish to cite a dynamic dataset or a subset of a larger dataset.
 - **Version** of the dataset analyzed in the citing paper
 - **Access date** when the data was accessed for analysis in the citing paper
 - **Subset** of the dataset analyzed (e.g., a range of dates or record numbers, a list of variables)
 - **Verifier** that the dataset or subset accessed by a reader is identical to the one analyzed by the author (e.g., a Checksum)
 - **Location** of the dataset on the internet, needed if the identifier is not "actionable" (convertable to a web address)

Example citations

- Kumar, Sujai (2012): 20 Nematode Proteomes. figshare. <https://doi.org/10.6084/m9.figshare.96035.v2> (Accessed 2016-09-06).
- Morran LT, Parrish II RC, Gelarden IA, Lively CM (2012) Data from: Temporal dynamics of outcrossing and host mortality rates in host-pathogen experimental coevolution. Dryad Digital Repository. <https://doi.org/10.5061/dryad.c3gh6>
- Donna Strahan. "08-B-1 from Jordan/Petra Great Temple/Upper Temenos/Trench 94/Locus 41". (2009) In Petra Great Temple Excavations. Martha Sharp Joukowsky (Ed.) Releases: 2009-10-26. Open Context. <https://opencontext.org/subjects/30C3F340-5D14-497A-B9D0-7A0DA2C019F1> ARK (Archive): <http://n2t.net/ark:/28722/k2125xk7p>
- OECD (2008), Social Expenditures aggregates, OECD Social Expenditure Statistics (database). <https://doi.org/10.1787/000530172303> (Accessed on 2008-12-02).
- Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. https://doi.org/10.1594/WDC/dphase_mpeps
- Manoug, J L (1882): Useful data on the rise of the Nile. Alexandria : Printing-Office V Penasson. <http://n2t.net/ark:/13960/t44q88124>

Useful links

- **Zenodo - CERN-OpenAIRE OA repository - catch all**

www.zenodo.org

- **Choose a license - Creative Commons**

<https://creativecommons.org/choose/?lang=en>

<https://chooser-beta.creativecommons.org/>

- **DMP examples by subject - LIBER**

<https://libereurope.eu/dmpcatalogue/>

- **Tools to create your DMP**

<https://www.openaire.eu/argos/>

<https://dmponline.dcc.ac.uk/>

<https://argos.openaire.eu/splash/>

- **Re3Data**

<https://www.re3data.org/>

- **Metadata standard Directory - Research Data Alliance**

<https://rd-alliance.github.io/metadata-directory/>

Let's look at some examples!

Go to the Zenodo community "**Liber DMP Catalogue**"

<https://zenodo.org/communities/liber-dmp-cat/?page=1&size=20>

An example of workflow

You need to be up to date in your research field



But you need to be well organized!



A good (and organized) workflow for researchers

← Thread




Maya Gosztyla 🧠🔬

@MayaGosztyla



Three years into my PhD program, I think I've finally figured out a paper reading system that works well for me.

There are 4 main steps in my workflow: finding papers, storing papers, reading papers, and annotating papers.

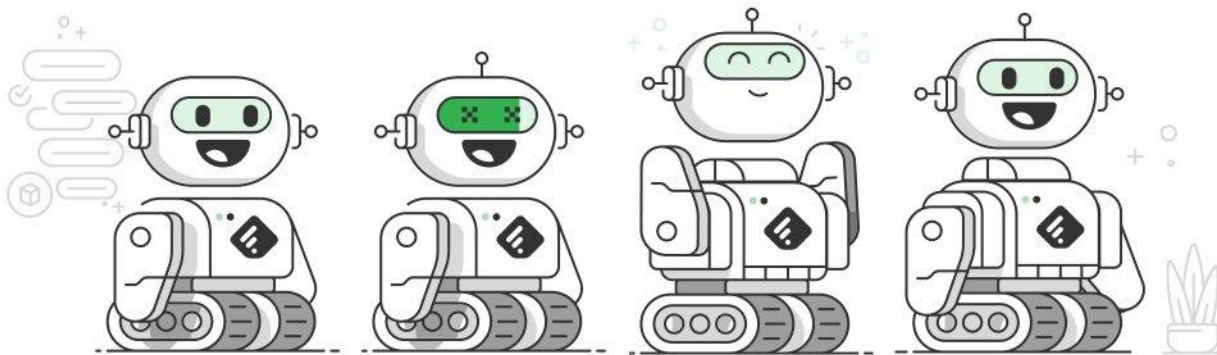
A brief  on my what I use:

9:23 PM · Apr 19, 2022 · Twitter Web App

<https://twitter.com/MayaGosztyla/status/1516497730574557186>

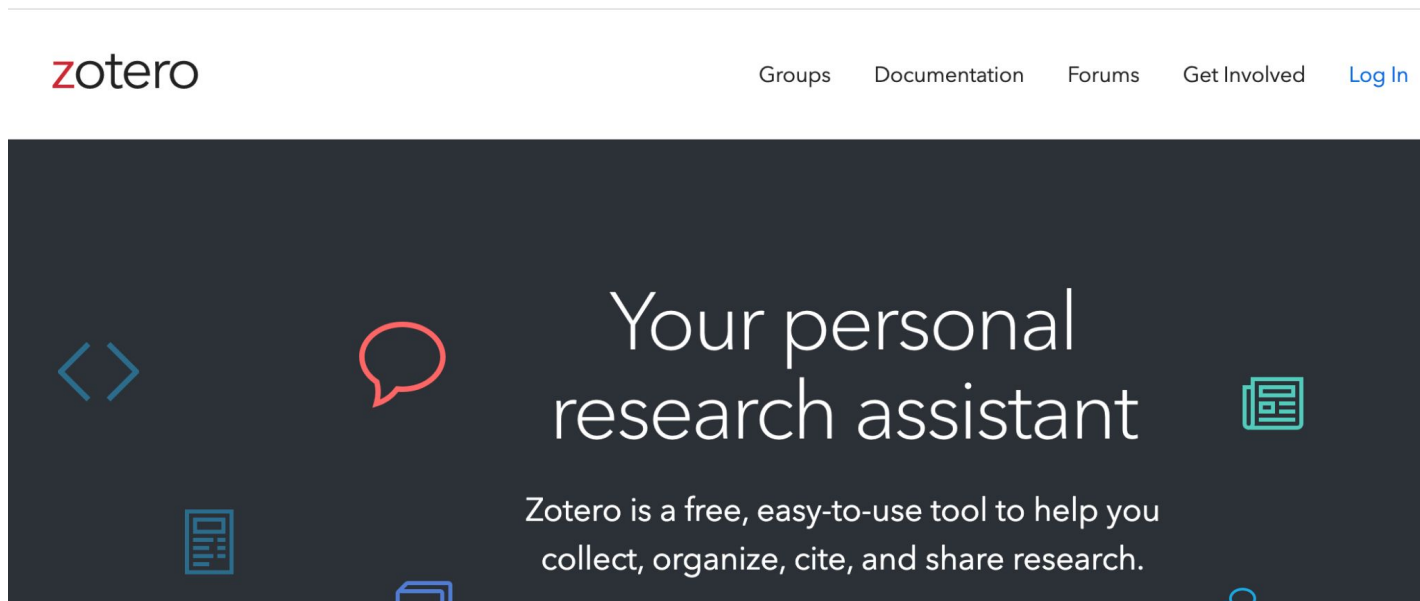
Step 1: find papers

- Rss aggregators (like [Feedly](#)) - feeds for keywords and journals of your field
- Twitter - to avoid distractions arrange a list with scientists/RPOs in your field



Step 2: store papers to read later

Zotero to store papers (with browser plugin)



Step 3: read and annotate papers

Use a system for taking notes, highlighting (Zotero - that has a built-in PDF viewer where you can make notes, highlights - or others)



Summarize key concepts

[NotionHQ](#) with the Notero plugin to organize my papers database and store detailed notes, and maybe add a screenshot of some important figures



And here's the template:



Hide description

Papers Template

Created by Maya Gosztyla (Twitter @MayaGosztyla)

 Papers to Read  Grouped by Project  Action Items  All Papers  Annotated  Skimmed

[Filter](#) [Sort](#) 

 Name	 Title	 Status	 Type	 Interesting?	 Action Items
 READ ME	How to us this template	Need to Read		★★★★	!!
 Smith et al. 2022	Example paper	Need to Read	Manuscript	★★★	!

<https://dust-bovid-04d.notion.site/ba1bc5022a4d469abd2ea559e80bc56b?v=a953e6bf2a3449e9ab3083154ce2b238>

How to use NotionHQ: <https://www.youtube.com/c/ThomasFrankExplains>

Hands on!



Go to Jamboard and follow the instructions

- Go to:
<https://jamboard.google.com/d/1Ql4-TMlfGhgFttjrbqO1nAPPAlIW42FEc6io60MOvCI/edit?usp=sharing>
- Think on your PhD research project in terms of DMP
- Defined the contents of the main sections of a DMP

Thank you!

Ask questions and interact in the VRE:

https://services.d4science.org/group/phdunipi_os21-22

gina.pavone@isti.cnr.it



Consiglio Nazionale
delle Ricerche

