
Data Management Plan

The Horizon Europe template

University of Pisa,
22 november 2022

Gina Pavone, CNR  0000-0003-0087-2151

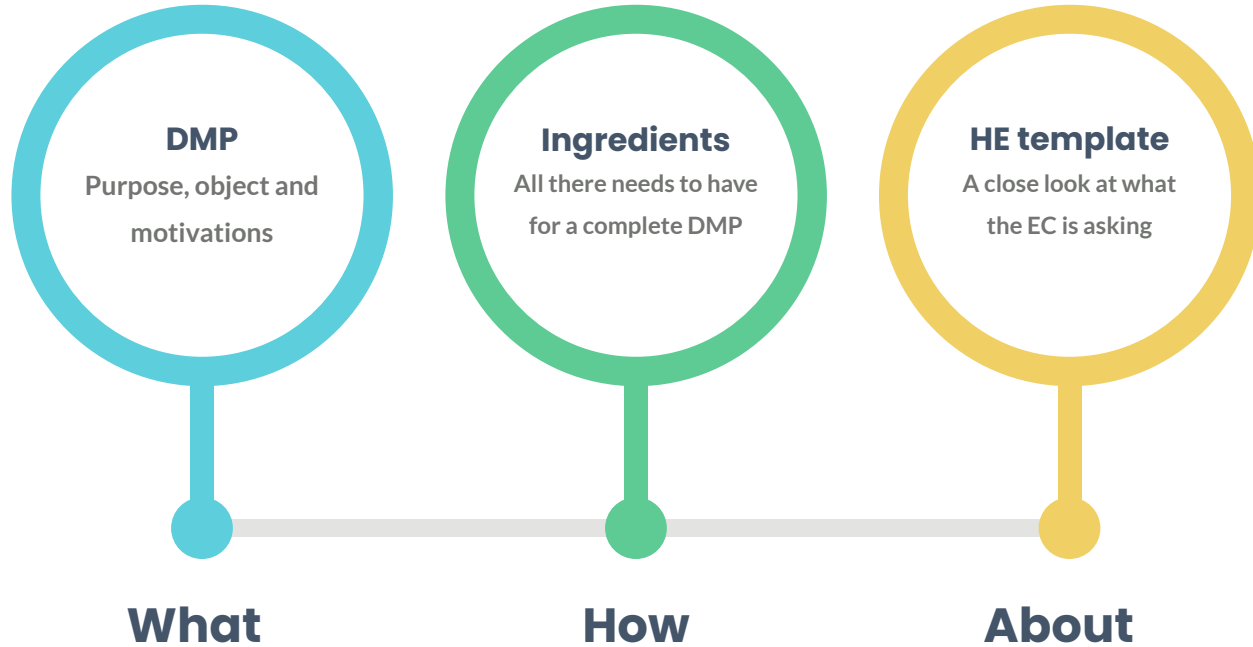
Gina Pavone

- Research fellow at the Institute of Information Science and Technologies of the Italian National Research Council in Pisa, Italy.
- Research focus: Open Science and Open Access; Research Data Management
- OpenAIRE National Open Access Desk (NOAD) for Italy
- Coordinator of the editorial board of open-science.it website
- My background: data journalism

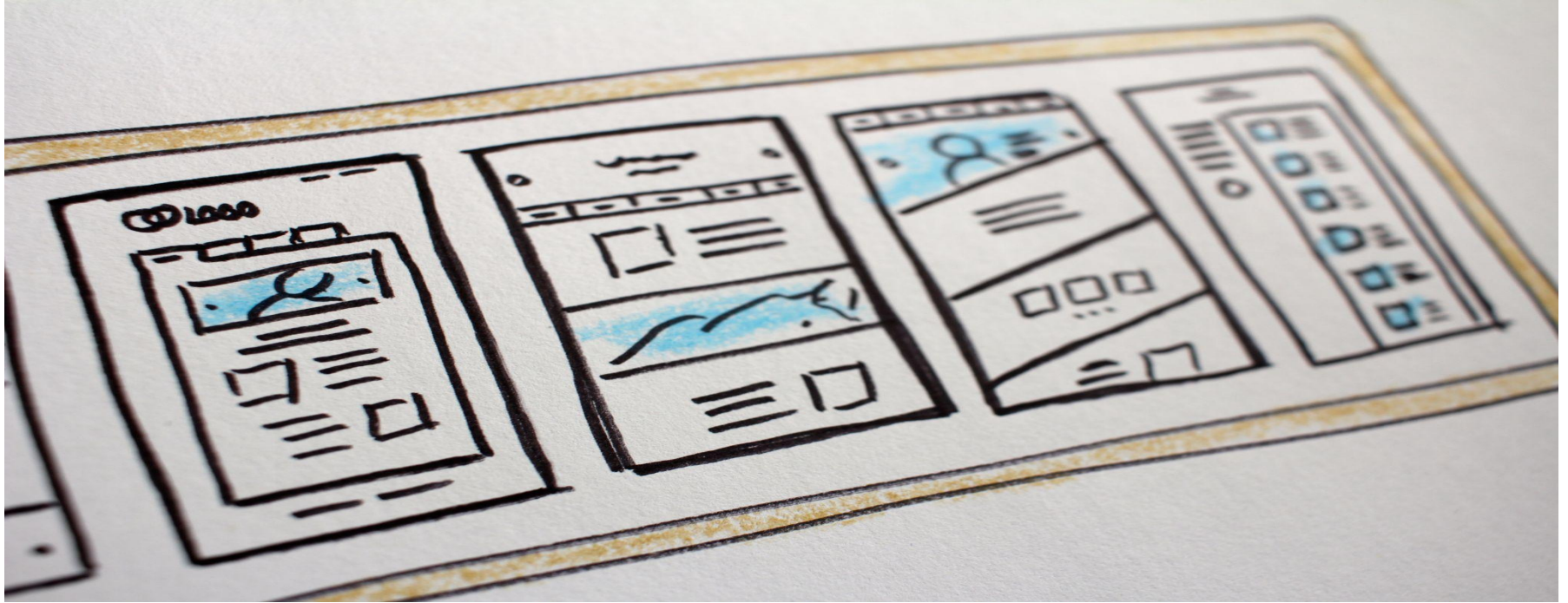


Agenda of today

University of Pisa



A planning effort



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?

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.

For oneself

- Save time and try to prevent problems in the future -
- Estimate costs -
- Get credit for your data and do not drown in irrelevant data -

For mandates

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

For others

- Produce FAIR data, easier to find, understand and reuse -
- 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 (to minimize risk)

- 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.



Photo by Belinda Fewings on Unsplash

DMP: so many benefits



Make data FAIR

- 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;
- Choose a proper repository etc.



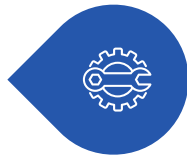
Clarifies needed budget

- calculating time and resources for careful documentation as well as server space, backup solutions, hardware and software etc.
- Calculating time and resources (money and expertise) for collecting, analysing, and publishing on data.



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.



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.

A timely approach to RDM



The DMP has to be done before or at early stage of the research activity

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

1. [Horizon EUrope Annotated model grant agreement, annex 5](#)
2. [Horizon Europe DMP template](#)



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.



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

Mandatory updates

In HE 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](#)



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.

What is normally a DMP about?

Very basic aspects



Identify

the data you are working with in your project.

- Accurately describe the types of data to be used
- Why do you need that data?
- What is the research question to be answered?



Decide

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

- Make careful choices to document all steps
- In the future it will be easy to understand and retrieve all the information?

Manage

Make decisions about 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?



Topic and aspects to
address in a DMP

What are data?

Data or it didn't happen!

Facts, observations or experiences on which an argument or theory is constructed or tested.

**Data are information!
(in a variety of forms and
formats)**

UCL Research Data Policy

<https://www.ucl.ac.uk/library/research-support/research-data-management>

Data are first-class research objects

Check
Validation
Follow-ups
New research questions
Teaching
Business applications
...



PUBLICATIONS AND DATA

Data summary

Main elements to describe

- types of data
- purpose of the data
- file formats
- organization: file naming and folder structure
- Expected size of the data
- Provenance (origin/source)
- Is that data potentially useful for others?

Types of research data

There is a huge variety of data types.

Research data can be classified in different ways, for example based on their:

Content: numerical, textual, audiovisual, multimedia...

Form: spreadsheets, databases, images, maps, audio files, (un)structured text...

Mode of data collection: experimental, observational, simulation, derived/compiled from other sources

Digital (born-digital or digitized) or **non-digital nature** (e.g. paper surveys, samples, notes...)

Primary (generated by the researcher for a particular research purpose or project) or **secondary nature** (originally created by someone else for another purpose)

Raw or processed nature

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...



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

- <https://www.ugent.be/en/research/datamanagement/why/rdm-explained.htm>
- <https://dmeg.CESSDA.eu/Data-Management-Expert-Guide/1.-Plan/Benefits-of-data-management>

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).

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)

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
- WP number
- 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.

A typical situation



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.**

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.

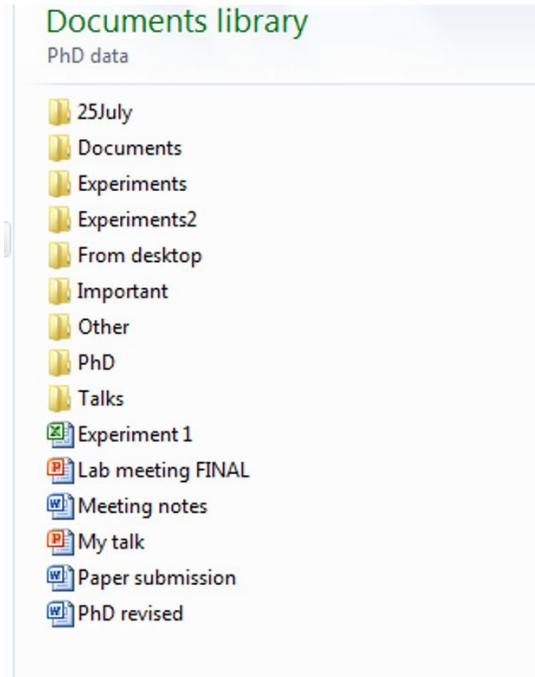
Structure folders hierarchically - limited number of folders for the broader topics, and then create more specific folders within these.

Consider separating ongoing and completed work.

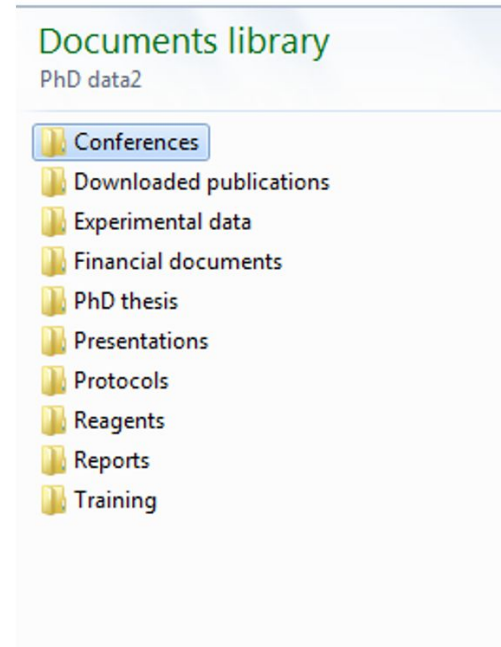
Backup – ensure that your files, whether they are on your local drive, or on a network drive, are backed up.

A typical situation

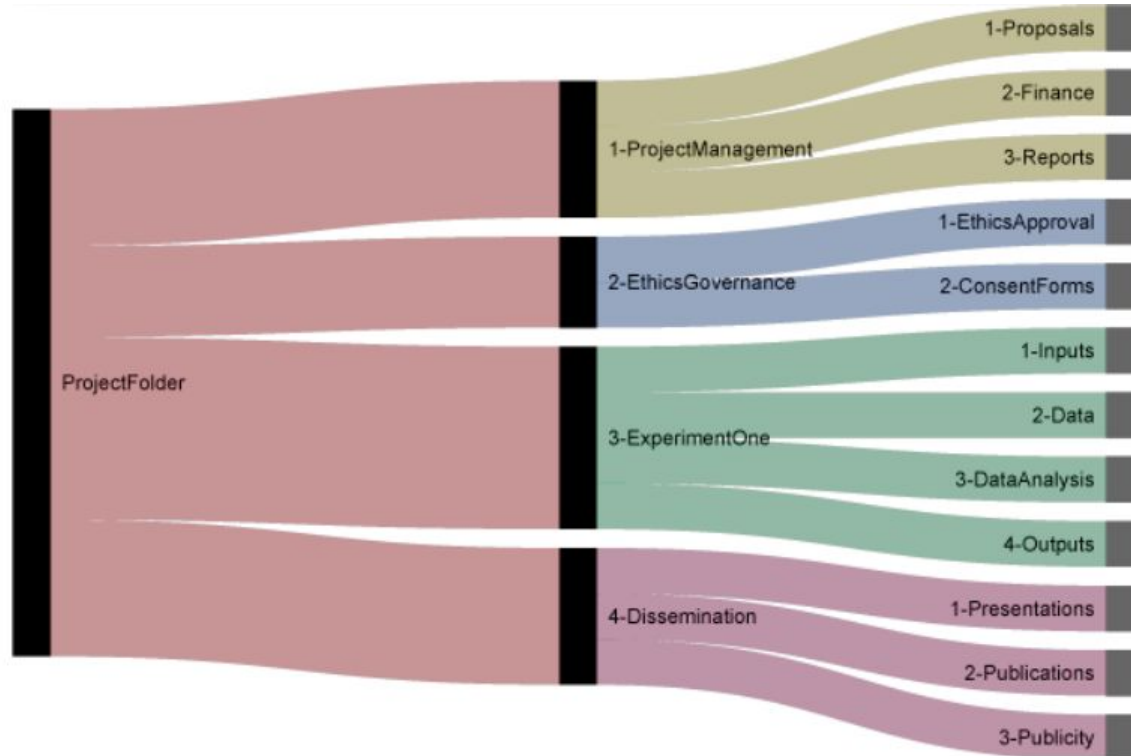
Example A



Example B



A good example



FAIR data

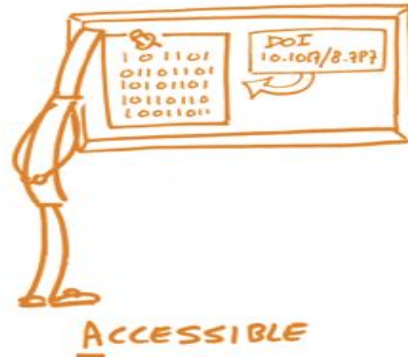
The meaning of the acronym

FAIR DATA PRINCIPLES



Findable:

Others can easily discover your data



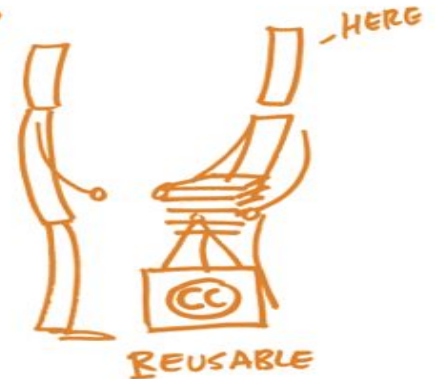
Accessible:

It is clear who, when and how can access your data (does not mean open)



Interoperable:

Your data can be integrated with other data and/or they can be easily used and read by machines



Reusable:

Your data can be reused by others in new research

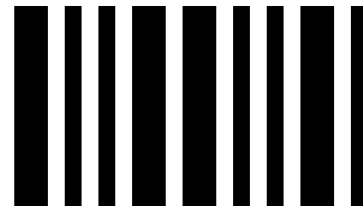
Findable

Main elements to include:

- Persistent identifiers
 - Metadata (to allow discovery)
-

Persistent Identifiers

- A **persistent identifier** (PI or PID) is a long-lasting reference to a document, file, web page, or other object.
- The term persistent identifier is usually used in the context of **digital objects** that are accessible over the Internet.
- Typically, such an identifier is not only persistent but **actionable**: you can plug it into a web browser and be taken to the identified source.
- It is like the barcode used on products...



Many types of PIDs

People - ORCID

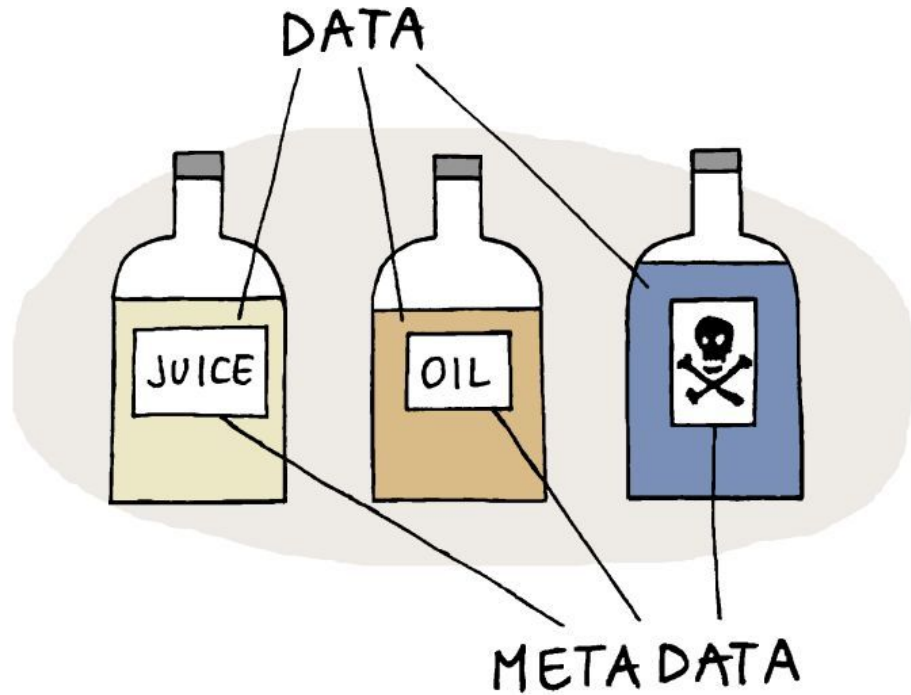
Projects - RAiD www.raid.org.au

Digital objects - DOI

Physical samples IGSN - <https://www.igsn.org/>

Example services that supply globally unique and persistent identifiers

- Identifiers.org provides resolvable identifiers in the form of URIs and CURIEs: <http://identifiers.org>
- Universally unique identifier: https://en.wikipedia.org/wiki/Universally_unique_identifier
- Persistent URLs: <http://www.purlz.org>
- Digital Object Identifier: <http://www.doi.org>
- Archival Resource Key: <https://escholarship.org/uc/item/9p9863nc>
- Research Resource Identifiers: <https://scicrunch.org/resources>
- Identifiers for funding organisations (see F3 & R1): <https://www.crossref.org/services/funder-registry/>
- Identifiers for the world's research organisations (see F3 & R1): <https://www.grid.ac>



PioloDataedo

The difference between data and metadata

Piotr Kononow through <https://twitter.com/aabella/status/1527533226574680064/photo/1>

Main types of metadata

Descriptive metadata	For finding or understanding a resource
Administrative metadata: <ul style="list-style-type: none">- Technical metadata- Preservation metadata- Rights metadata	<ul style="list-style-type: none">- For decoding and rendering files- Long-term management of files- Intellectual property rights attached to content
Technical metadata	For instance, those captured by a device/tool/machin etc.
Structural metadata	Relationships of parts of resources to one another

Types of metadata: some examples

Metadata type	Example properties	Primary uses
Descriptive metadata	<ul style="list-style-type: none">- Title- Author- Subject- Genre- Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management

Examples of metadata schema

- Dublin core - describing resources on the web
- Schema.org - commercial applications
- Crossref - research outputs
- Datacite metadata schema - describing research data
- Disciplinari metadata
 - Data Documentation Initiative
 - Darwin Core - biological sciences

Use your discipline specific standard!

You will spend less time curating and interpreting data and more time to actually make science!

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



Registries and other tools for findability

- OpenDOAR <https://v2.sherpa.ac.uk/opendoar/>
A quality-assured, global Directory of Open Access Repositories
- FAIRsharing: <https://fairsharing.org/>
A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.
- Re3data <https://www.re3data.org/>
Registry of research data repositories
- Roar <http://roar.eprints.org/>
Registry of Open Access repositories
- Google search/scholar + Unpaywall <https://unpaywall.org/>
Unpaywall is database of scholarly articles and a browser extension skip the paywall on millions of peer-reviewed journal articles: it free, and legal!

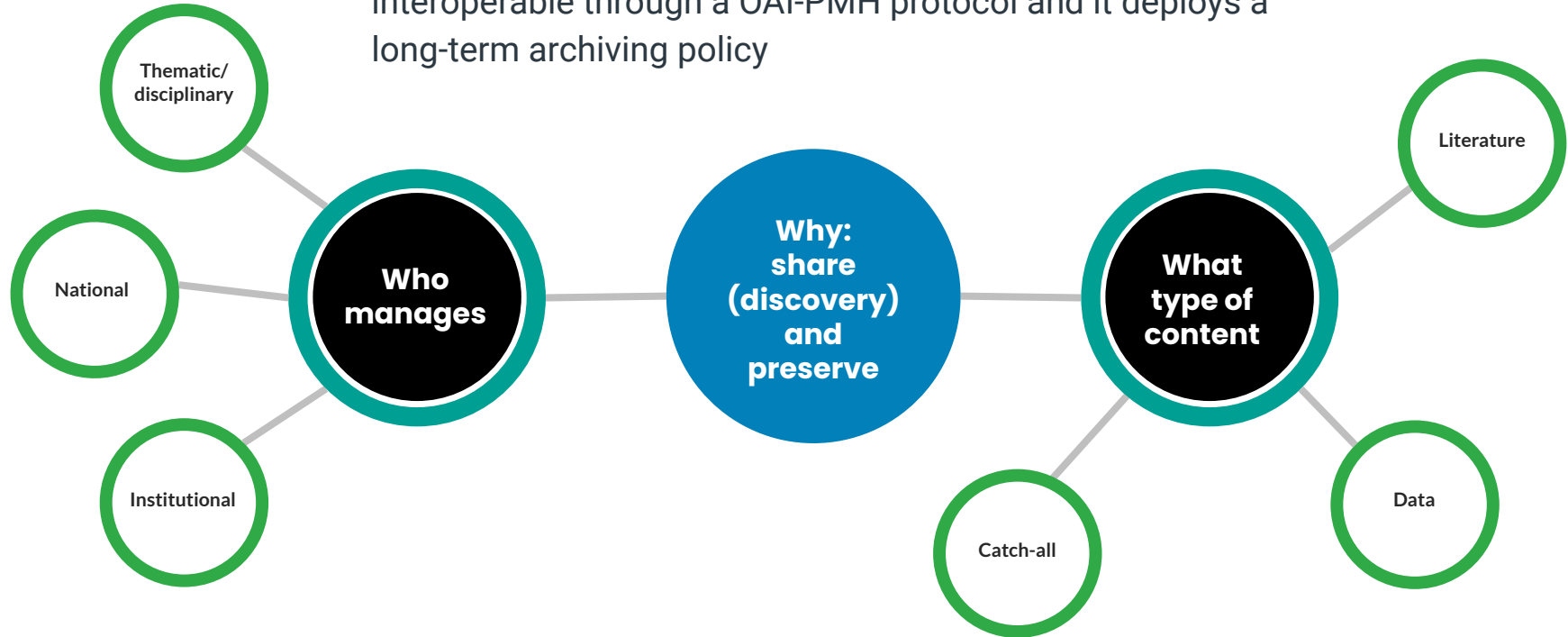
Accessible

Main elements to consider:

- Trusted repository
- If data can be publicly available
- If some data need to be protected

Open Access repositories

A repository stores Open Access digital objects and makes them available and downloadable. It's accessible and interoperable through a OAI-PMH protocol and it deploys a long-term archiving policy



Trusted repositories

Definition contained in the HE Model Grant Agreement



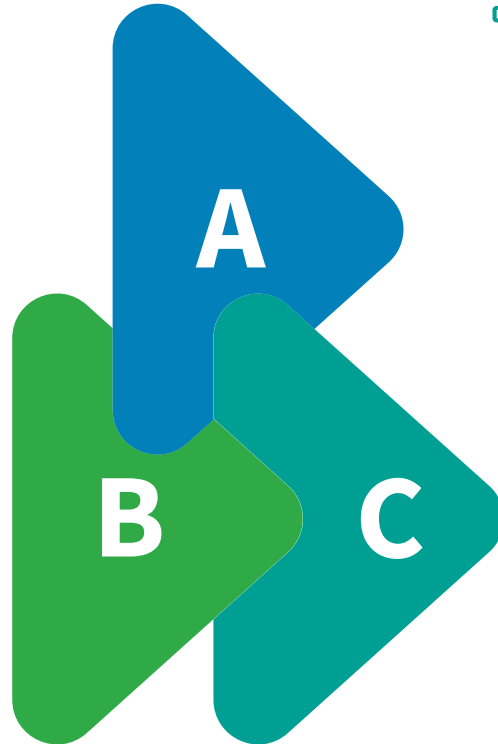
Certified repositories

E.g. CoreTrustSeal,
nestor Seal DIN31644,
ISO16363 etc.



Disciplinary or domain specific repos

Commonly used, endorsed by the research communities and internationally recognized



General purpose or institutional repositories

That present the essential characteristics of trusted repositories:





- Display services, mechanisms and/or provisions that are intended to secure the integrity and authenticity of their contents; display policy
- Provide broad, equitable and ideally open access to content free at the point of use, as appropriate, and respect applicable legal and ethical limitations. They assign PIDs. Have metadata enabling discovery
- Facilitate mid- and long-term preservation of the deposited material.

Define access rights

License

required 

Access right *

-  Open Access
-  Embargoed Access
-  Restricted Access
-  Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

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](#).

Open Access should be the default access right

Embargoed Access: it when you have a valid reason to delay access

Restricted access: use it when you have a valid reason to restrict the access

Always specify conditions under which you grant access (who, how, why can get access to your payload)

Closed access: are you really sure you need this?

consider restricted or embargoed access instead!

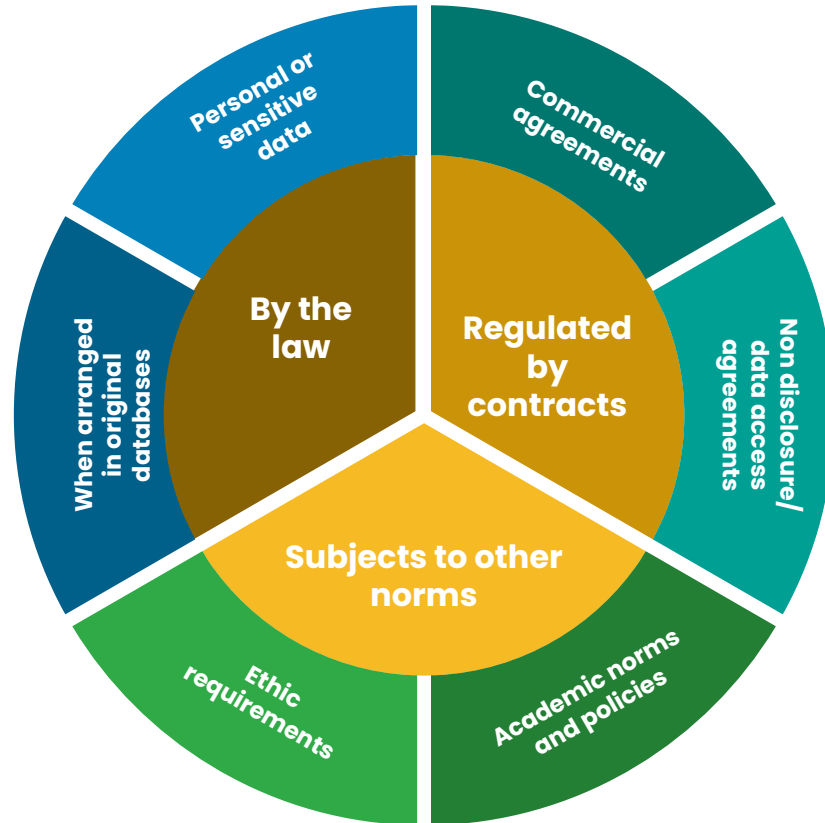
Note: metadata is always accessible to everyone



Are there any restrictions on access e.g. because of sensitive data? Conditions of access (e.g., who to contact and how) should be clearly specified

Data can be protected

Multiple types of protection might exist in research data, or there may be elements that have no legal protection



(FAIR) Open Data

Data can be freely used, shared, enriched by anyone, anywhere for any purpose.

FAIR Data





Data follow a series of good practices to allow data access, still respecting any ethical, legal and contractual restriction.

Why do we need a distinction?



Photo by [Possessed Photography](#) on [Unsplash](#)

Research data could:

-  Contain personal information (privacy e GDPR)
-  Fall under copyright (in the case of a database with creative structure)
-  Fall under the Sui Generis right (database obtained thanks to a substantial investment)
-  Be protected by patent or industrial secret

Data sharing needs to respect the specific law.
Data needs to be protected against non authorised access.

Data to be handled with great care:

- **Personal data:** any information about an identified or identifiable natural person (directly or indirectly)
- Personal **sensitive data** (i.e. revealing racial or ethnic origin, political views, religious or philosophical beliefs, membership of a trade union, genetic data, biometric data, data about health or someone's sexual behavior or sexual orientation)
- Data protected by **IPR** (Intellectual Property Rights) agreements
- **Confidential data** (i.e. commercial agreements)

This means that access to the data must be managed and restricted.

They still can be FAIR

Interoperable

Main aspects to consider:

- vocabularies
 - references to other metadata
-

Integration with other products

By humans and machines

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Standards and formats

To make your data understandable to others (humans and machines) you need to use adequate standards and formats

- Formats may refer to:
 - File format (.txt, .docx, .jpeg, etc)
 - Metadata (Dublin core, discipline specific standards)
 - Data organisation/visualisation
- Use specific ontologies and vocabularies to make your data easy to read
- Use your discipline specific standards: you will spend less time curating and interpreting data and more time to actually make science!

What is a controlled vocabulary

Controlled vocabularies are standardized and organized arrangements of words and phrases and provide a consistent way to describe data. Metadata creators assign terms from vocabularies to improve information retrieval.

<https://guides.lib.utexas.edu/metadata-basics/controlled-vocabs>

Controlled vocabulary schemes mandate the use of predefined, authorised terms that have been preselected by the designers of the schemes, in contrast to **natural language** vocabularies, which have no such restriction.

https://en.wikipedia.org/wiki/Controlled_vocabulary

DDI Controlled Vocabulary for Mode Of Collection

The procedure, technique, or mode of inquiry used to attain the data.

Value of the Code	Descriptive Term of the Code	Definition of the Code
Interview	Interview	A pre-planned communication between two (or more) people - the interviewer(s) and the interviewee(s) - in which information is obtained by the interviewer(s) from the interviewee(s). If group interaction is part of the method, use "Focus group".
Interview.FaceToFace	Face-to-face interview	Data collection method in which a live interviewer conducts a personal interview, presenting questions and entering the responses. Use this broader term if not CAPI or PAPI, or if not known whether CAPI/PAPI or not.
Interview.FaceToFace.CAPIorCAMI	Face-to-face interview: Computer-assisted (CAPI/CAMI)	Computer-assisted personal interviewing (CAPI), or computer-assisted mobile interviewing (CAMI). Data collection method in which the interviewer reads questions to the respondents from the screen of a computer, laptop, or a mobile device like tablet or smartphone, and enters the answers in the same device. The administration of the interview is managed by a specifically designed program/application.
Interview.FaceToFace.PAPI	Face-to-face interview: Paper-and-pencil (PAPI)	Paper-and-pencil interviewing (PAPI). The interviewer uses a traditional paper questionnaire to read the questions and enter the answers.
Interview.Telephone	Telephone interview	Interview administered on the telephone. Use this broader term if not CATI, or if not known whether CATI or not.
Interview.Telephone.CATI	Telephone interview: Computer-assisted (CATI)	Computer-assisted telephone interviewing (CATI). The interviewer asks questions as directed by a computer, responses are keyed directly into the computer and the administration of the interview is managed by a specifically designed program.
Interview.Email	E-mail interview	Interviews conducted via e-mail, usually consisting of several e-mail messages that allow the discussion to continue beyond the first set of questions and answers, or the first e-mail exchange.
Interview.WebBased	Web-based interview	An interview conducted via the Internet. For example, interviews conducted within online forums or using web-based audio-visual technology that enables the interviewer(s) and interviewee(s) to communicate in real time.
SelfAdministeredQuestionnaire	Self-administered questionnaire	Data collection method in which the respondent reads or listens to the questions, and enters the responses by him/herself; no live interviewer is present, or participates in the questionnaire administration. If possible, use a narrower term. Use this broader term if the method is not described by any of the narrower terms - for example, for PDF and diskette questionnaires.
SelfAdministeredQuestionnaire.Email	Self-administered questionnaire: E-mail	Self-administered survey in which questions are presented to the respondent in the text body of an e-mail or as an attachment to an e-mail, but not as a link to a web-based questionnaire. Responses are also sent back via e-mail, in the e-mail body or as an attachment.

References to other research outputs

odo.org/deposit/3778807

funding agency know:

Grants

European Commission (EU)

OpenAIRE-Advance

777541

OpenAIRE Advancing

European Commission (EU)

EOSCsecretariat.eu

831644

EOSCsecretariat.eu

Optional. OpenAIRE-supported pro
Note: a human Zenodo curator wil

+ Add another grant

Related/alternate identifiers

Specify identifiers of related publications and datasets. Supported identifiers i
arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs and URLs.

Related identifiers

10.5281/zenodo.380176

subject

10.5281/zenodo.382618

10.5281/zenodo.390163

+ Add another related identifie

Contributors

References

cites this upload

is cited by this upload

is supplemented by this upload

is a supplement to this upload

is referenced by this upload

references this upload

published this upload

is previous version of this upload

is new version of this upload

✓ continues this upload

is continued by this upload

describes this upload

is described by this upload

has this upload as part

is part of this upload

reviews this upload

is reviewed by this upload

documents this upload

is documented by this upload

is compiled/created by this upload

compiled/created this upload

is the source this upload is derived from

has this upload as its source

is required by this upload

requires this upload

replaces this upload

is replaced by this upload

object

Publication date:

April 30, 2020

DOI:

DOI 10.5281/zenodo.3778807

Keyword(s):

Open Science, Open Access, OpenAIRE, funders mandates

Grants:

European Commission:

- OpenAIRE-Advance - OpenAIRE Advancing Open Scholarship (777541)
- EOSCsecretariat.eu - EOSCsecretariat.eu (831644)

Related identifiers:

Continued by

10.5281/zenodo.3801760 (Presentation)

10.5281/zenodo.3826183 (Presentation)

10.5281/zenodo.3901639 (Presentation)

Communities:

Open Science in Italy

License (for files):

Creative Commons Attribution 4.0 International

Reusable

Main elements to include:

- documentation
 - licences
 - quality assurance
-

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

So many ways to describe your data

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



The screenshot shows a web interface for a README file template. At the top left is the Cornell University logo. To its right is a document icon and the filename 'AUTHOR_DATASET_ReadmeTemplate.txt'. Below this is a light gray box containing the template text. The text starts with a header line: 'This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]'. This is followed by a section separator consisting of two dashed lines and the text 'GENERAL INFORMATION'. Below this are two numbered list items: '1. Title of Dataset' and '2. Author Information'. At the bottom of the template is a section for 'Principal Investigator Contact Information' with fields for 'Name:', 'Institution:', 'Address:', and 'Email:'.

```

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>



Licenses

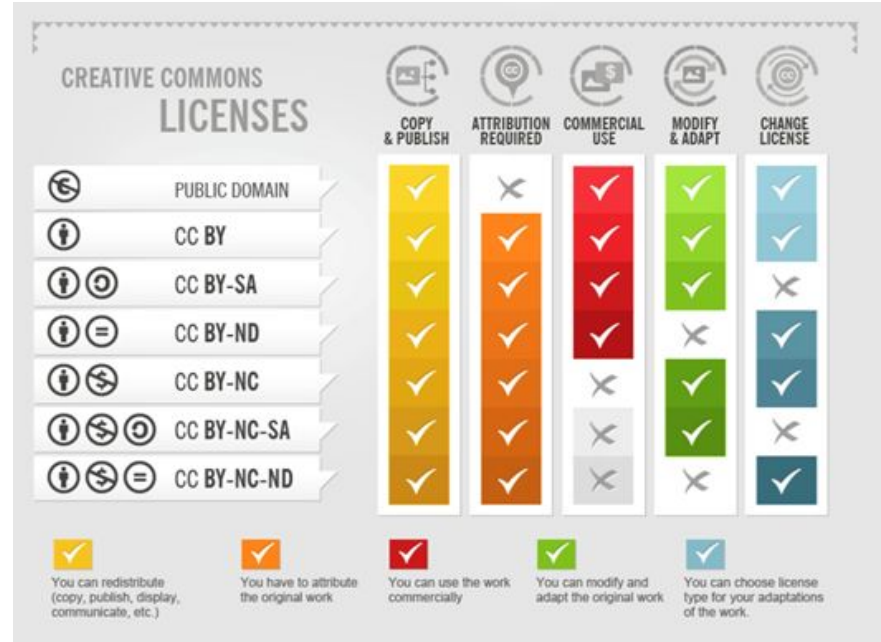
Tell other what they can
do with your data

Creative Commons

Not all of us are legal experts capable of writing proper licenses.

Creative Commons and Public Domain create legal certainty for everyone, who wants to use works, that are licensed respectively.

Usually, in OS framework and in OS mandates, the required licence is CC-BY, CC0 or equivalent.



The infographic displays the Creative Commons license matrix. It lists seven license types on the left, each with its icon. To the right, five columns represent permissions: Copy & Publish, Attribution Required, Commercial Use, Modify & Adapt, and Change License. Each cell in the matrix contains a checkmark (indicating permission) or an 'X' (indicating restriction). A legend at the bottom explains the meaning of the checkmarks in each column.

CREATIVE COMMONS LICENSES		COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
PUBLIC DOMAIN		✓	✗		✓	✓
CC BY		✓	✓	✓	✓	✓
CC BY-SA		✓	✓	✓	✓	✗
CC BY-ND		✓	✓	✓	✗	✓
CC BY-NC		✓	✓	✗	✓	✓
CC BY-NC-SA		✓	✓	✗	✓	✗
CC BY-NC-ND		✓	✓	✗	✗	✓

Legend:

- You can redistribute (copy, publish, display, communicate, etc.)
- You have to attribute the original work
- You can use the work commercially
- You can modify and adapt the original work
- You can choose license type for your adaptations of the work.

Quality assurance processes

Any measure to ensure the quality and reliability of the data.

For instance, it can be using data normalization protocols, staff and stakeholder training in order to promote data consistency, establishing data handling and/or analysis procedures and so on

Some examples:

- Checking for equipment and transcription errors
- Quality control of materials
- Data integrity checks
- Calibration procedures
- Data capture resolution and repetitions
- Other procedures related to data quality such as weighting, calibration, reasons for missing values, checks and corrections of transcripts, transformations...

Allocation of resources

Costs

Main elements to include:

- all the predictable expenses
 - who is responsible for RDM
-

Costs - 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

✓ 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 security

Main aspects to detail:

- Security
- Storage solutions
- Data access inside the project (ie. consortium)
- Data recovery strategies

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)

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

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>

Consider alternatives...



The screenshot shows the top navigation bar of the Consortium GARR website. The main header is dark blue with the GARR logo and navigation links: Infrastrutture, Comunità, Servizi, and Ricerca e formazione. Below this is a banner for 'INFRASTRUTTURA CLOUD' with a blue and green abstract background. The main content area has a white background and contains the following text:

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>



The screenshot shows the SURF DRIVE website homepage. The top navigation bar is orange and white, with the SURF DRIVE logo and navigation links: Home, Downloads, Tutorials, FAQ, About SURFdrive, and Contact. The main content area has an orange and white color scheme and contains the following text:

Personal cloud storage service for Dutch education and research



[Log in to SURFdrive](#)

Why SURFdrive?

Secure file storage
Log in with your 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.

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!

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>

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

Ethics

Ethical aspects

Main aspects to include:

- how personal and sensitive data will be protected
- management of informed consents
- reference to ethics deverbale

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

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

Tools:

Informed consent

Ethical assessments

Anonymization (software) and pseudo-anonymization

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 pseudonymisation 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.

Other issues and other research outputs

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...).

Detail on the Research Data repository.

Detail on non-digital data and materials.

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.

Resources and materials

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

- How will new data be collected or produced and/or how will existing data be re-used?
- What data (for example the kinds, formats, and volumes) will be collected or produced?

2. Documentation and data quality

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

3. Storage and backup during the research process

- How will data and metadata be stored and backed up during the research process?
- How will data security and protection of sensitive data be taken care of during the research?

4. Legal and ethical requirements, codes of conduct

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

5. Data sharing and long-term preservation

- How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?
- How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?
- What methods or software tools will be needed to access and use the data?
- 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

- Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?
- 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



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>

<http://wikimedia.sp.unipi.it/images/Grigliapianodigestionedatiricerca.pdf>

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

- **Horizon Eurppe DMP template**

<https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf>

- **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/>

- **Research Data Management decision tree**

<https://zenodo.org/record/7190005#.Y3s5guzMI-Q>

Thank you!

gina.pavone@isti.cnr.it



Consiglio Nazionale
delle Ricerche



UNIVERSITÀ DI PISA