

Open Science and Data Management Plan

Module 3.1 - Horizon Europe and EOSC

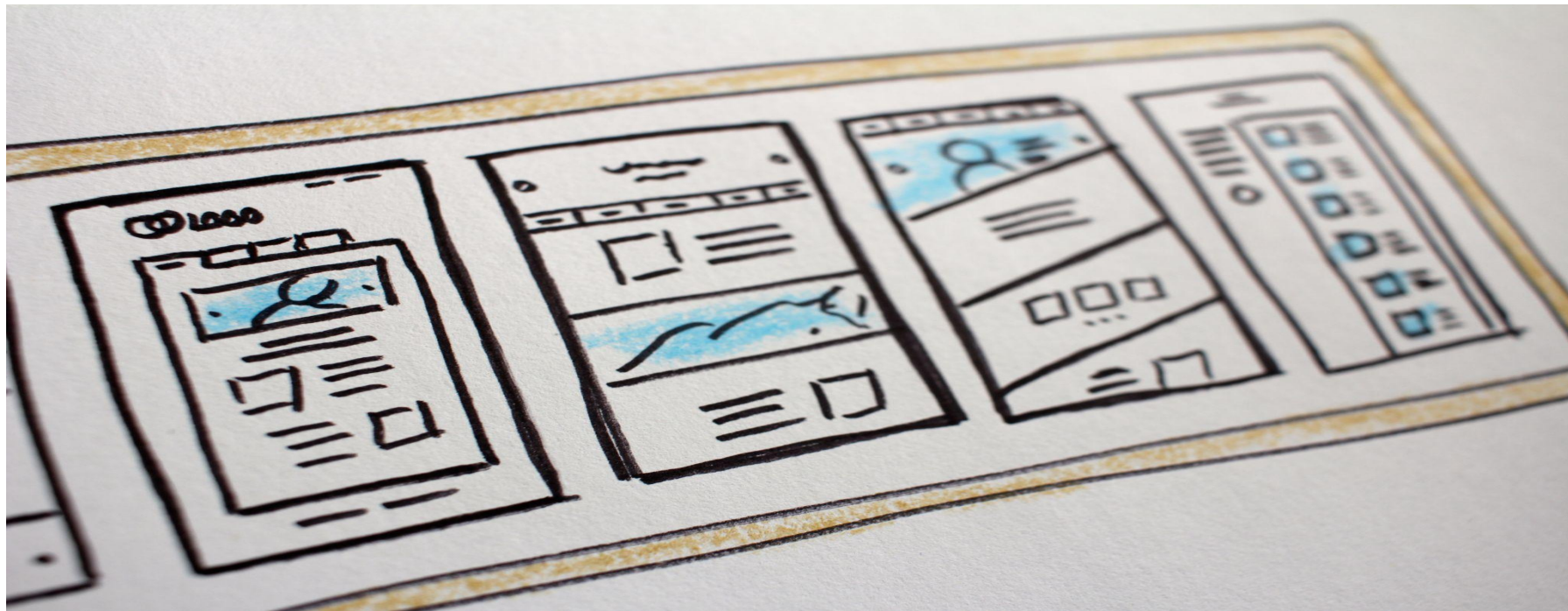
15/05/2023, Area Science Park

Gina Pavone, CNR-ISTI

ORCID 0000-0003-0087-2151

Data Management Plan

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

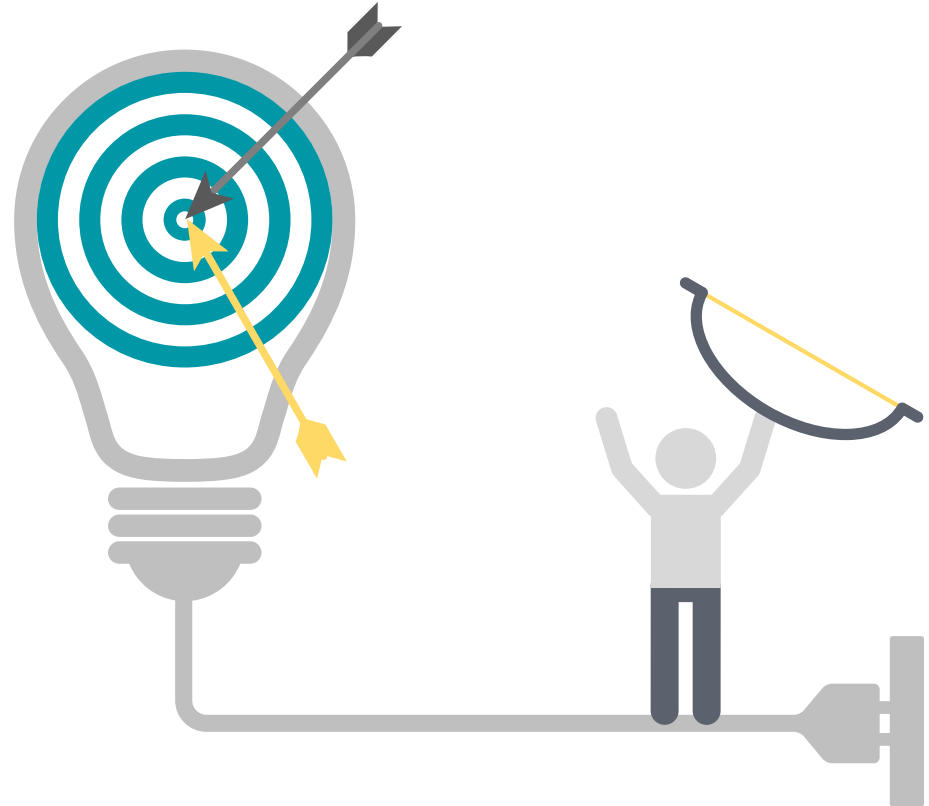
- It is mandatory in EC and ERC funded projects -
- Also other funders ask for it -
- RPOs may have their own policy on RDM -

For others

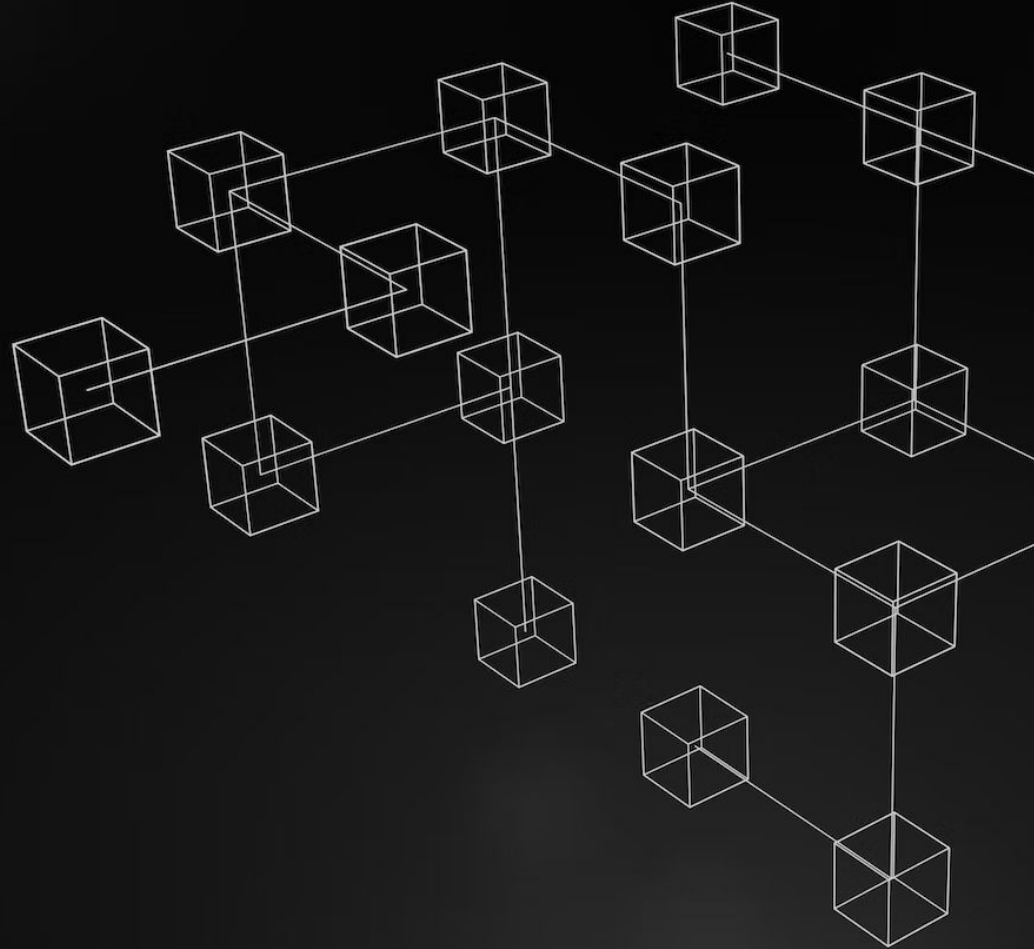
- Produce FAIR data, easier to find, understand and reuse -
- DMPs may also be required as part of the ethical approval process -

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



DMPs are useful
whenever
researchers are
creating and
managing data for
research

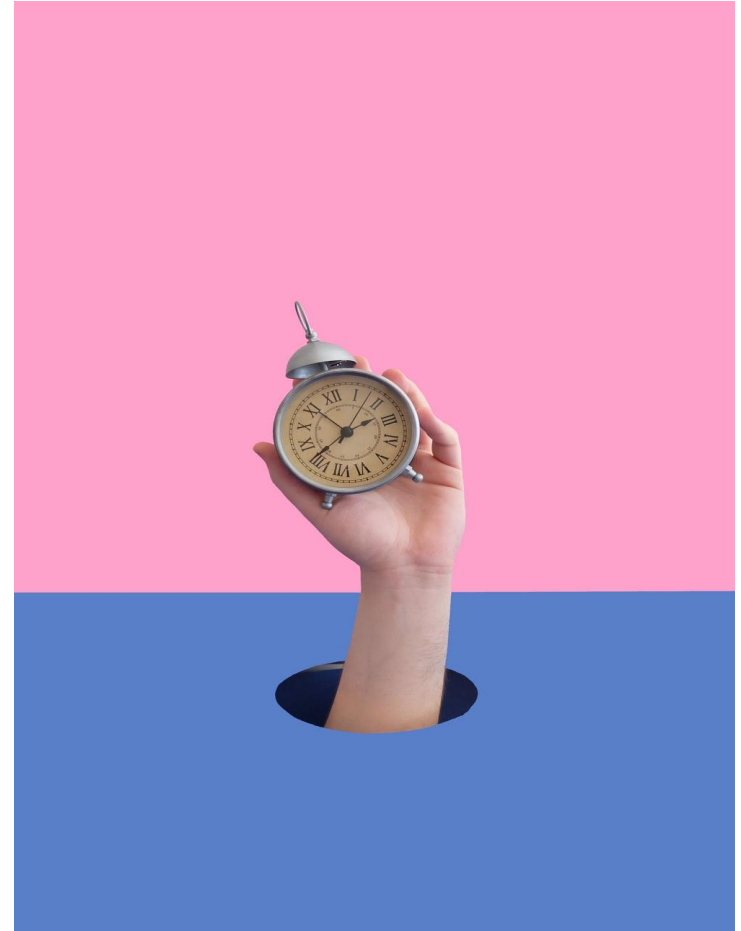


DMP benefits

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

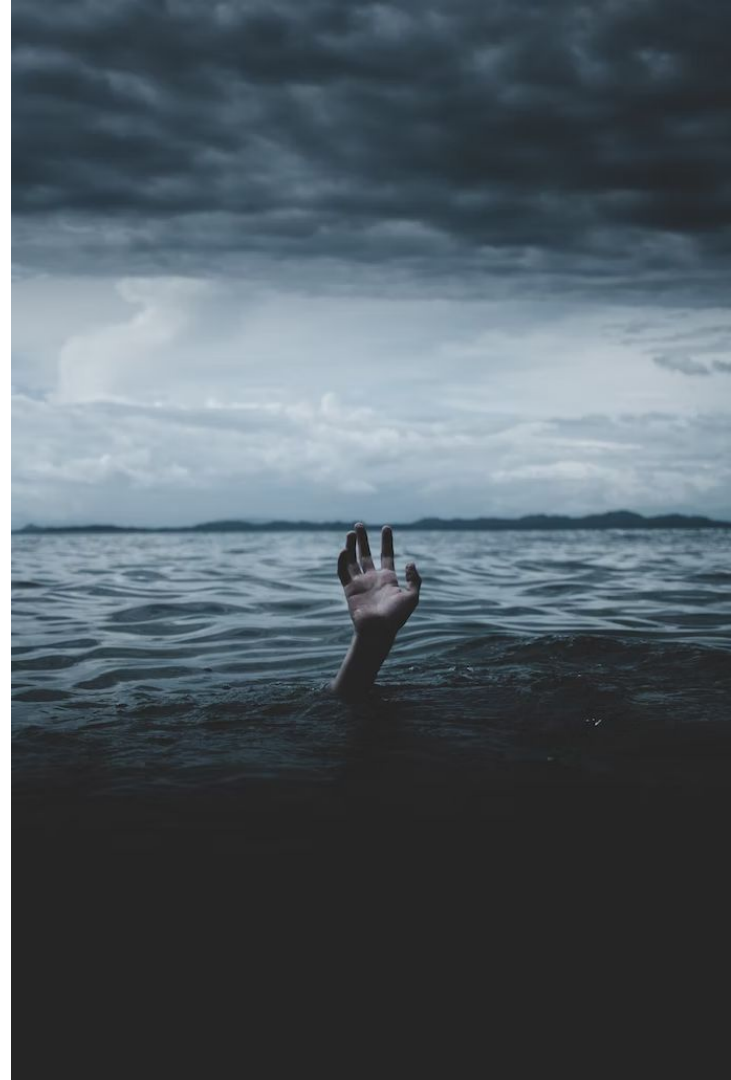
- An important function of a DMP is to work as a one-stop shop to find project-related information.
- Questions surrounding data management are being gathered in one place and project-related details are readily available rather than just vaguely remembered or simply forgotten.



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.

Stop drowning in irrelevant stuff



Time is the enemy!

- Accumulating large quantities of disorganized files
- Lack of information describing content in digital files
- Changes to hardware, software, and file formats in common use
- File corruption
- Failure of storage media
- Data leaving with collaborators

And - last but not least - tons of daily activities!

DMP timing

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](#)



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

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.

Topics and aspects to
address in a DMP

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?



Sections and elements to include

- **Administrative data**
- **Data summary.** A clear and exhaustive description of the data:
 - Types of data
 - Provenance (origin/source) or mode of collection
 - Purpose of the data
 - File formats
 - Expected size of the data
- **FAIR principles**
- Organization and data management within the consortium:
 - file naming and folder structure
 - storage
- **Security**
 - Backup strategy
 - Disaster recovery strategy
- **Ethical issues**
- **Allocation of resources.** Costs
- Data utility and data sharing (outside the consortium)
 - repository
 - access
 - licences

Be concise, but precise!

We do not need many words

But we need to be detailed!



Interaction

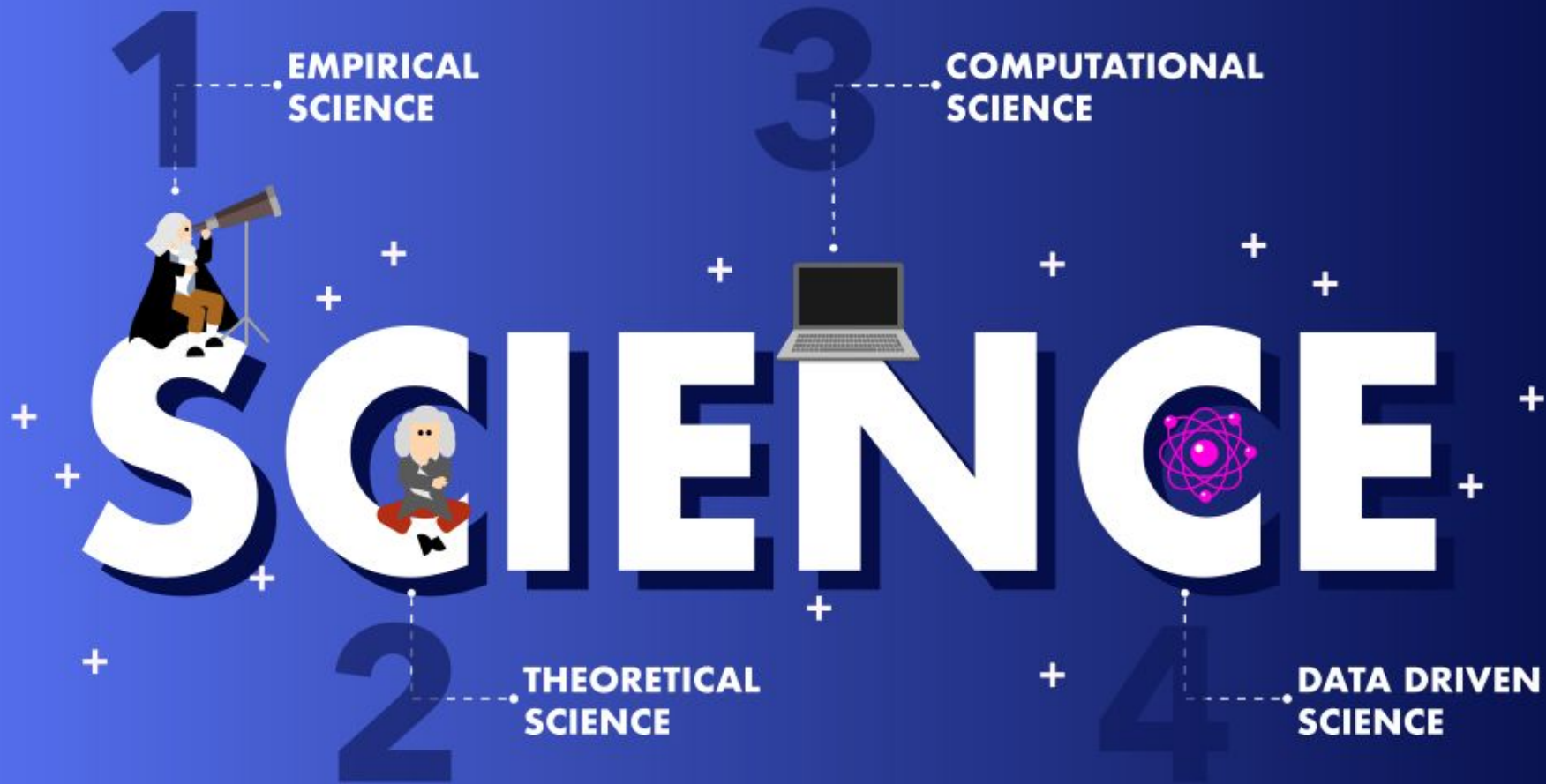
Go to:

<https://www.menti.com>

Voting code: 6844 5122







Data are first-class research objects

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



PUBLICATIONS AND DATA

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>

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...

Format: 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, 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

<https://www.ugent.be/en/research/datamanagement/why/rdm-explained.htm>



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

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

While planning, carefully consider what data will be produced in the course of their project.

E.g. Spatial, temporal, instrument-generated, models, simulations, images, video etc.

Research data means data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media.

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.

Example of data types

Text: Field or laboratory notes, survey responses

Numeric: Tables, counts, measurements

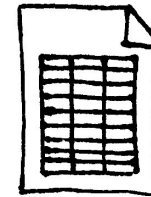
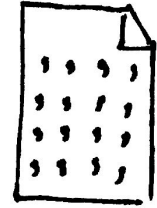
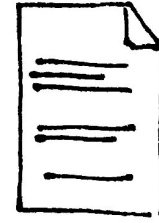
Audiovisual: Images, sound recordings, video

Code and models: Python, MATLAB, R, etc.

Discipline-specific: FITS in astronomy, CIF in chemistry

Instrument-specific: Equipment outputs

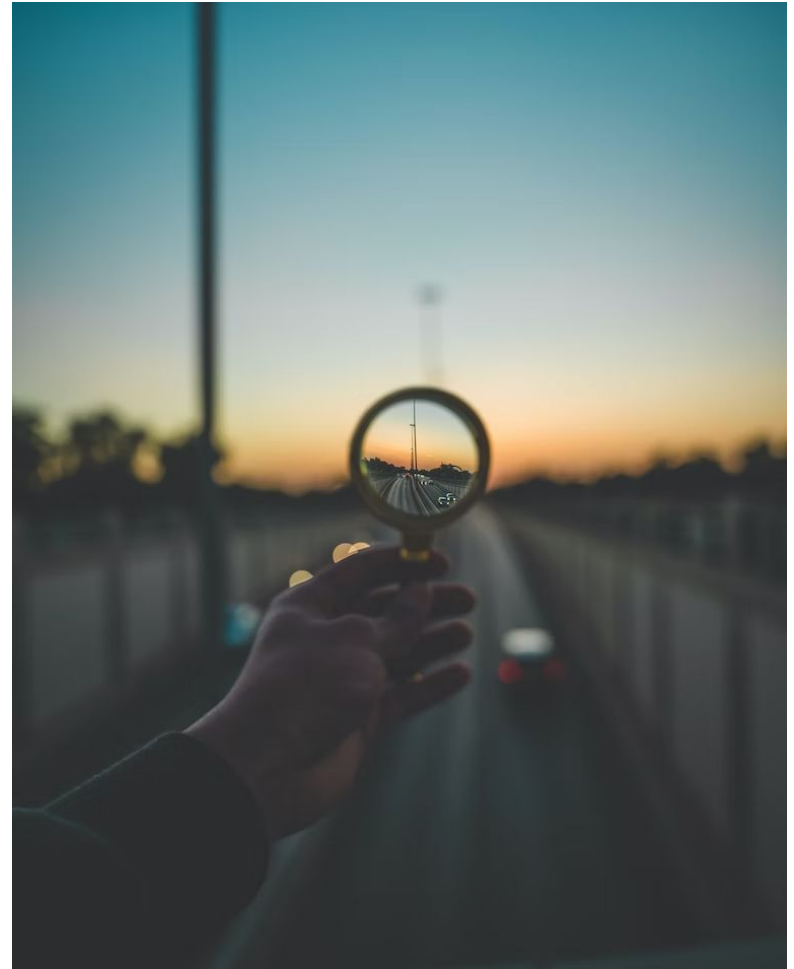
Spatial, temporal, instrument-generated, models, simulations, images, video



Purpose of the data

Describe what each data type will be used for in relation to the objectives of the research project

<https://unsplash.com/photos/ZAvhxLTcSok>



Data formats

Recommendation: try to capture your data in (or convert it to) community-accepted data formats.

Using standard or widely-adopted formats will make your data interoperable. Open or non-proprietary formats are preferable, as you and others will have less trouble processing these later.

All digital information is – by nature – software dependent.

Problem: All digital data may thus be endangered by the obsolescence of the hardware and software environment on which access to data depends.

The safest option to guarantee long-term data access is to convert data to standard formats that most software are capable of interpreting, and that are suitable for data interchange and transformation.

Recommended file formats by extension

Content type	File formats
Text	PDF/A, HTML, XML, TXT
Tabular data (spreadsheets and databases)	XML, CSV
Numbers and statistics	TXT, DTA, POR, SAS, SAV
Geospatial	SHP, DBF, GeoTIFF, NetCDF
Audio	WAVE, AIFF, MP3, MXF
Images	TIFF, JPG, JP2, PNG, GIF, BMP
Moving Images	MOV, MPEG-4, AVI, MXF
Web Archive	WARC
Containers	TAR, GZIP, ZIP

If you deposit in an archive some formats may be preferable

FILE FORMATS CURRENTLY RECOMMENDED BY THE UK DATA ARCHIVE FOR LONG-TERM PRESERVATION OF RESEARCH DATA

TYPE OF DATA	RECOMMENDED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION
Quantitative tabular data with extensive metadata a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	SPSS portable format (.por) delimited text and command ("setup") file (SPSS, Stata, SAS, etc.) containing metadata information some structured text or mark-up file containing metadata information, e.g. DDI XML file
Quantitative tabular data with minimal metadata a matrix of data with or without column headings or variable names, but no other metadata or labelling	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate
Geospatial data vector and raster data	ESRI Shapefile (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) geo-referenced TIFF (.tif, .tiff) CAD data (.dwg) tabular GIS attribute data
Qualitative data textual	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)
Digital image data	TIFF version 6 uncompressed (.tif)
Digital audio data	Free Lossless Audio Codec (FLAC) (.flac)
Digital video data	MPEG-4 (.mp4) motion JPEG 2000 (.jp2)
Documentation	Rich Text Format (.rtf) PDF/A or PDF (.pdf) OpenDocument Text (.odt)

Note that other data centres or digital archives may recommend different formats.

File naming and folder structure

Well-organised file names and folder structures make it easier to find and keep track of data files.

Develop a system that works for your project and use it consistently!

<https://unsplash.com/photos/pONH9yZ-wXg>



File naming

About:

- cues to the content
- status of a file
- uniquely identify a file
- help in classifying files

What can contain:

- project acronyms
- researchers' initials
- file type information
- a version number
- file status information
- date

Best practices:

- create meaningful but brief names
- Provide useful cues to content, status and version
- Make it unique, descriptive and informative about the content.
- Make it independent of the location of the file
- avoid using spaces and special characters
- Keep it short
- keep the file extension at the end

Adapted from: <https://dam.ukdataservice.ac.uk/media/622417/managingsharing.pdf> and https://rdm.elixir-belgium.org/file_naming.html

File naming multiple possibilities



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.

File naming



A good example

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

3. Version

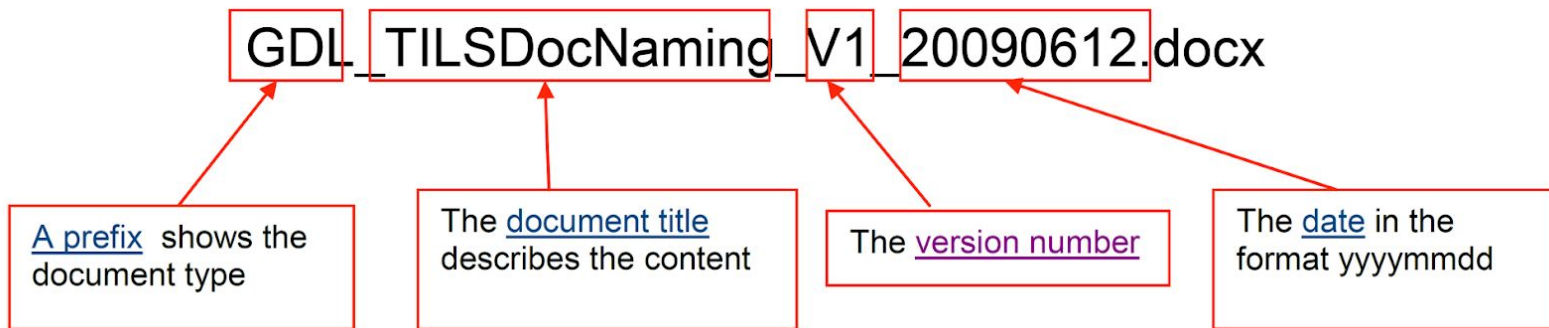
(upper case, max 4 chars, optional)

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:



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.**

File naming: what does the following names tell you?

1. Codecs_2022survey_Italy_20230126_GP.xlsx

Within the Codecs project (full name: Maximising the CO-benefits of agricultural Digitalisation through conducive digital ECoSystems) this file contains the results of a (yearly?) survey held in 2022 in Italy. It was last edited on January 26, 2023 by GP (Gina Pavone?). Based on the file name structure, you can expect other files to contain results to the same survey in different locations.

2. LabMeeting_20180712_RDM.docx

These are most probably notes taken in a lab meeting on July 12, 2018. The main subject was apparently research data management (RDM). The generic aspect (lab meeting) is placed first, then the date since it is a regular occurrence, and the specific subject comes last.

3. CONS_INT1_12-03-2019.rtf

Result from Interview 1 of the Consumers research on 12/13/2019 (provided that you know the abbreviations used)

4. GC-MS1_20180912_POLY03.ms

Polymer 3 measured on GC-MS machine 1 on the September 1 2018 (GC-MS is an analytical method and stands for gas chromatography–mass spectrometry)

5. FR3S_140623_129C_2653_W.JPG

This illegible file name can only make sense if it is accompanied with a codebook. This documentation will let you understand the detailed information displayed within the file name. In specific cases – such as massive generic file collections –, this approach can make a lot of sense, as long as the naming convention is well-documented

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.

Folder structure

Do's

- List all the type of files required for your work and try to group them according to logic criteria.
- Consider the best hierarchy for files
- Go from a general, high-level folder (e.g., a single folder for the project, using its name or acronym) to more specific lower-level folders
- Make it intuitive, clear and understandable for everyone.
- Apply file naming guidelines for naming folders.
- Include documentation to explain a complex folder structure (such as a README.txt) at the root of your folders.
- Separate raw and processed data

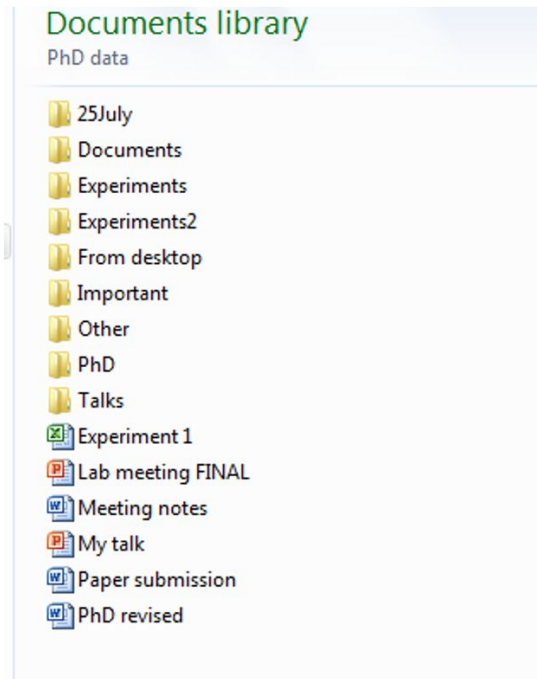
Dont's

- Make your structure too deep nor too shallow (the number of levels depends on the project).
- Use a generic “current stuff” or “my stuff” folder.
- Create researcher-specific folders (“Name_Surname” folder) within a project: folders are about the contents, not the authors.
- Create similar folders in different places (overlapping categories or folder redundancy).
- Create copies of files in different folders.

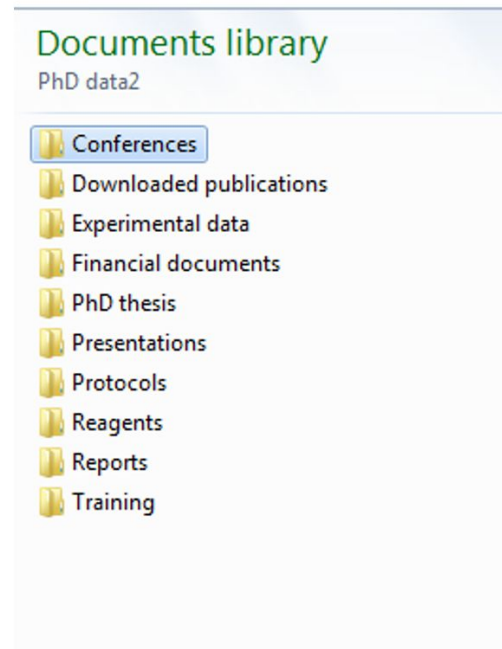
For example: Projects (README.txt); Administration (Budget, Approval, Travels); Planning (DMP, Ethics,); Literature; Experiments (ExperimentA, ExperimentB...); Dissemination (Posters, Presentation, Publications).

Folder structure: what is your strategy?

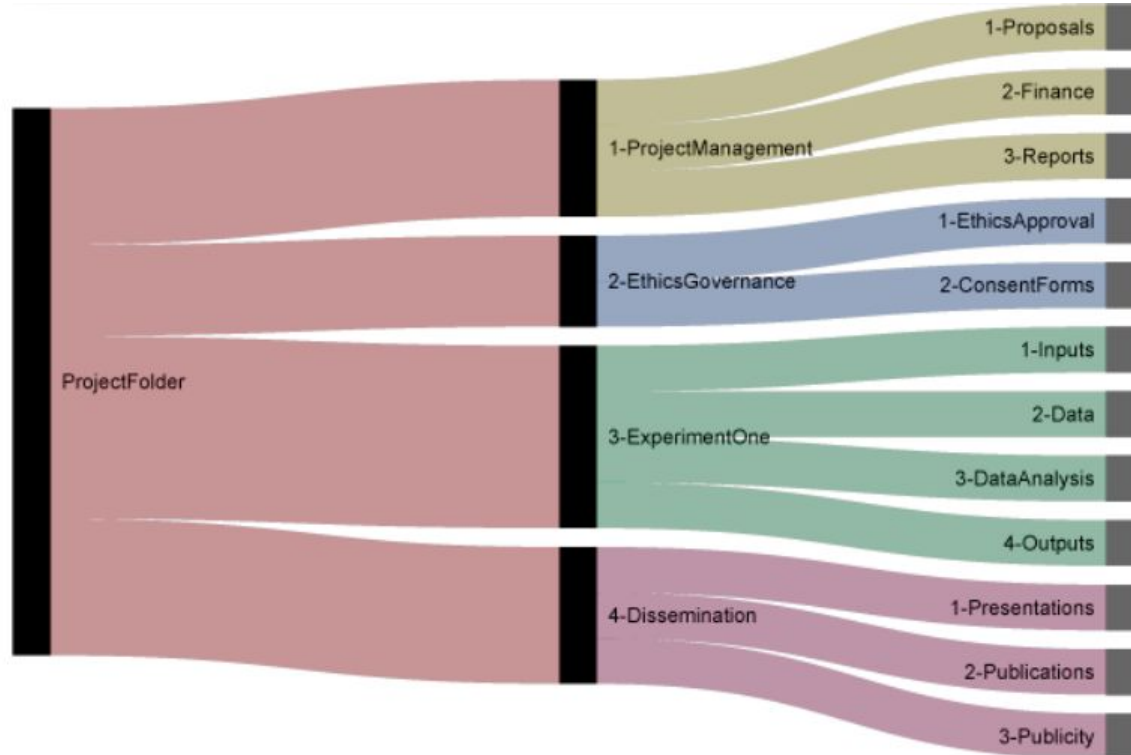
Example A



Example B



A good example



Exercise: folder and filename assessment - 5 minutes

1. Open the master folder for one of your current research projects and candidly assess how clearly and consistently you are naming your folders and files. Consider, for instance, the following:
 - How many levels of folders do you have – levels within levels within levels? Are there enough levels, too many, or too few?
 - Can you tell, without opening a folder, what is in it?
 - Is there some logic behind the organization of your folders, or does it seem you just created a new one whenever you needed one?
 - How consistent are the names of your files?
 - If you have multiple versions of particular files, can you tell which is the most current?
2. Create a new document (clearly named!) that will be a draft of the memo to yourself about the organization and naming of your folders and files. Create the first few bullets under “folder organization and naming” and under “file organization and naming” that could serve as instructions to yourself for (re)organizing and (re)naming your folders and files.

The goal is simply for you to begin thinking critically about how you are organizing and naming your folders and files, and to take the first steps toward more optimal organization and naming. If you find that there are some real improvements to make, think about when you will make them, cognizant that changes are much easier to make the earlier you are in your project.

FAIR data in the DMP

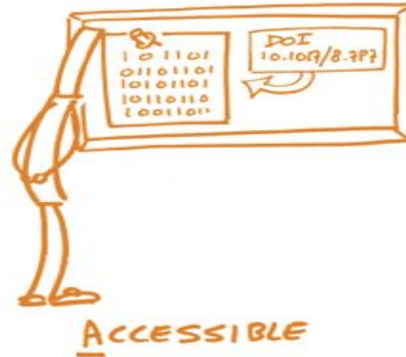
The FAIR principles

FAIR DATA PRINCIPLES



Findable:

Others can easily discover your data



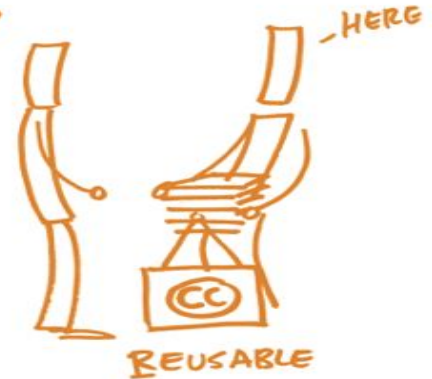
Accessible:

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:

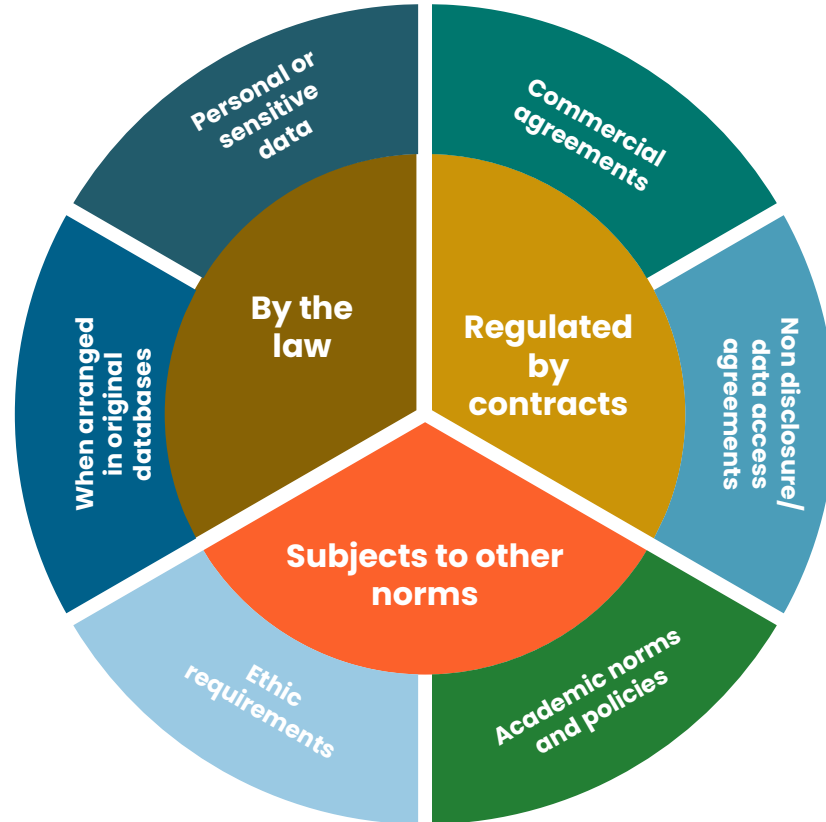
Your data can be reused by others in new research

Accessible
doesn't mean
Open



Data can be protected

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



FAIR is an evolution of the open data movement

More nuanced



More machines





- FAIR indicate a list of principles that can help you in making your data ready for Open Science
- They are **principles**, not standards!
- They were designed to enable optimal use of research data and methods
- A group of different experts designed the FAIR principles between 2014 and 2016
- They identified a set of 15 principles

Why FAIR

Making your research data Findable, Accessible, Interoperable and Reusable could:

- Help **peers** and your **future-self** understand the research project and data
- Facilitate data sharing and **collaborations**
- Increase the visibility of research and can lead to more **citations**
- Improve the **transparency, reliability and reproducibility** of research
- Prevent data loss

And thereby:

- Maximise potential from data assets
- Maximise research **impact**

<https://www.howtofair.dk/why-fair/>

Importantly, it is our intent that the principles apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data.

All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

Wilkinson et al.

FAIRification basics

- **Documentation**

Gives the context to make your data understandable by others

- **Metadata**

Make your data easy to find

- **Data formats**

Make your data simple to combine to other data and machine readable.

- **Access to data**

It means to decide who will have access to your data and how

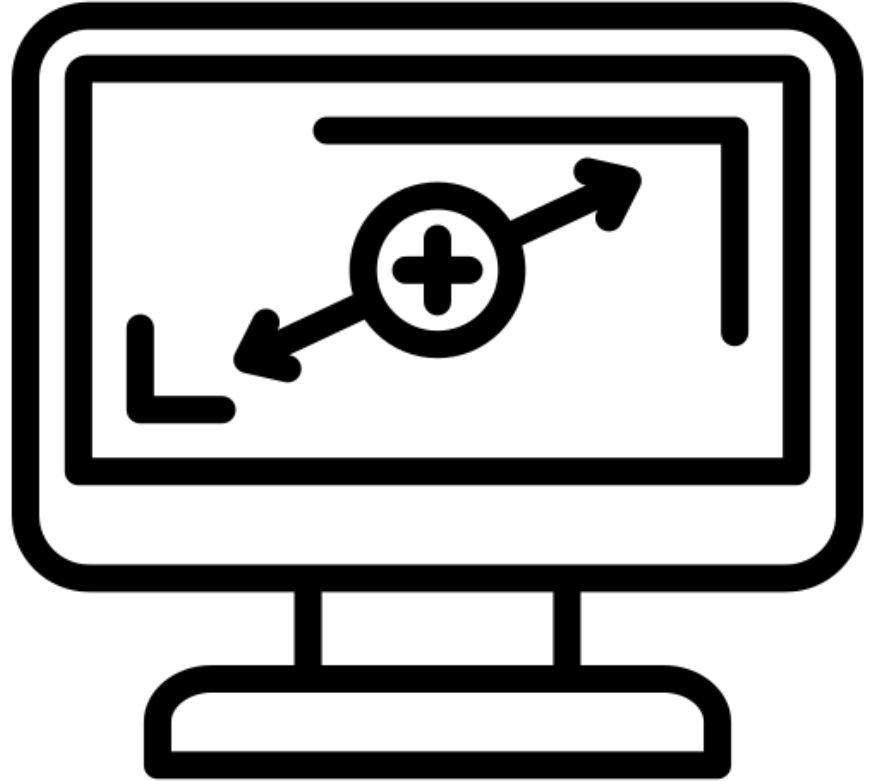
- **Persistent identifiers**

Persistent links to data that allows other to find and cite (give credit to) your data.

- **Licenses**

Are used to tell others how they can reuse your data.

No one size
fits all!

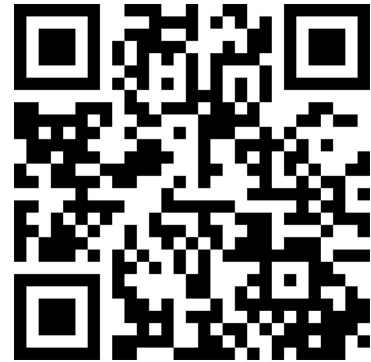


Interaction

Go to:

<https://www.menti.com>

Voting code: 6844 5122





Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource



Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
 - A1.1 The protocol is open, free, and universally implementable
 - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available



Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data



Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
 - R1.1. (Meta)data are released with a clear and accessible data usage license
 - R1.2. (Meta)data are associated with detailed provenance
 - R1.3. (Meta)data meet domain-relevant community standards

All the 15 principles

They are principles,
not standards!

Findable

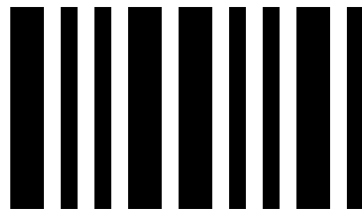
- The first step in (re)using data is to find them.
- Metadata and data should be easy to find for both humans and computers.
- Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the [FAIRification process](#).
- Use of persistent identifiers to uniquely identify some type of research product

Some "F" aspects to consider

- Use metadata and specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how.
- Use search keywords
- Persistent and unique identifiers such as DOI
- File and folder naming conventions
- Versioning of the datasets and clear version numbers

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...
- There are many types of PIDs!



Many types of PIDs

People - ORCID

Projects - RAiD www.raid.org.au

Digital objects - DOI

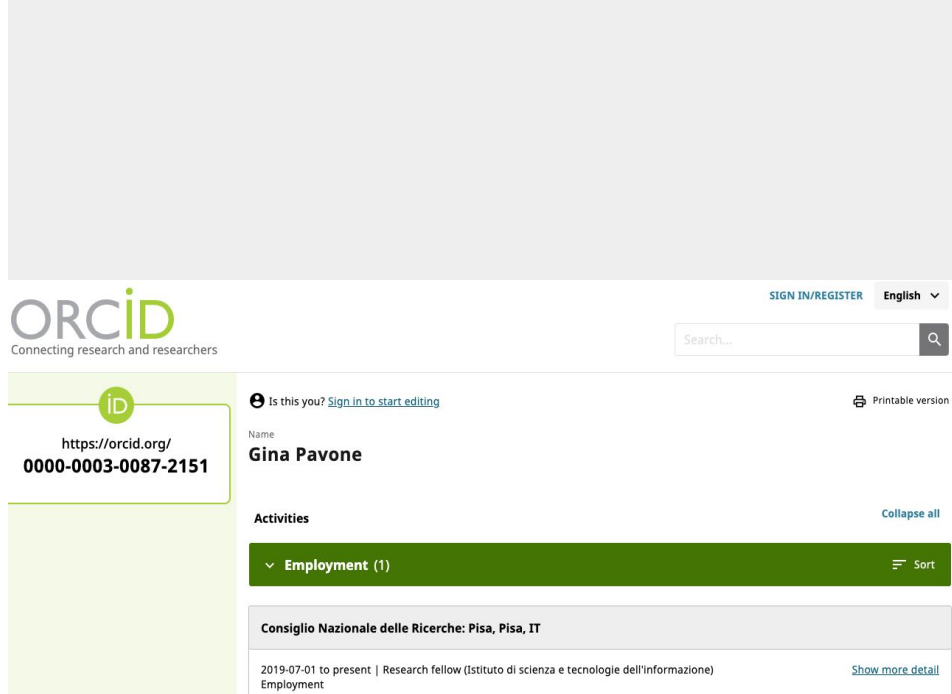
Physical samples IGSN - <https://www.igsn.org/>

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/>

ORCID: do you have one? you should...

Open Researcher and Contributor ID
is a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors and contributors



The screenshot shows the ORCID iD profile page for Gina Pavone. The header includes the ORCID logo with the tagline "Connecting research and researchers", a "SIGN IN/REGISTER" link, and a language dropdown set to "English". A search bar is located on the right. The profile section on the left displays the ORCID iD "https://orcid.org/0000-0003-0087-2151". The main content area shows the user's name "Gina Pavone" and a link to "Is this you? Sign in to start editing". Under the "Activities" section, there is a green bar for "Employment (1)" with a "Sort" button. Below this, a table lists the user's employment history.

Consiglio Nazionale delle Ricerche: Pisa, Pisa, IT	
2019-07-01 to present Research fellow (Istituto di scienza e tecnologie dell'informazione)	Show more detail

DOI - Digital Object Identifier

- In computing, a **digital object identifier** (DOI) is a [persistent identifier](#) or [handle](#) used to identify objects uniquely, standardized by the [International Organization for Standardization](#) (ISO).
- A DOI aims to be **resolvable**, usually to some form of access to the information object to which the DOI refers.
- This is achieved by **binding the DOI to [metadata](#)** about the object, such as a [URL](#), **indicating where** the object can be found
- a DOI differs from identifiers such as [ISBNs](#) and [ISRCs](#) which aim only to identify their referents uniquely

What are metadata?

Literally “data about data.”

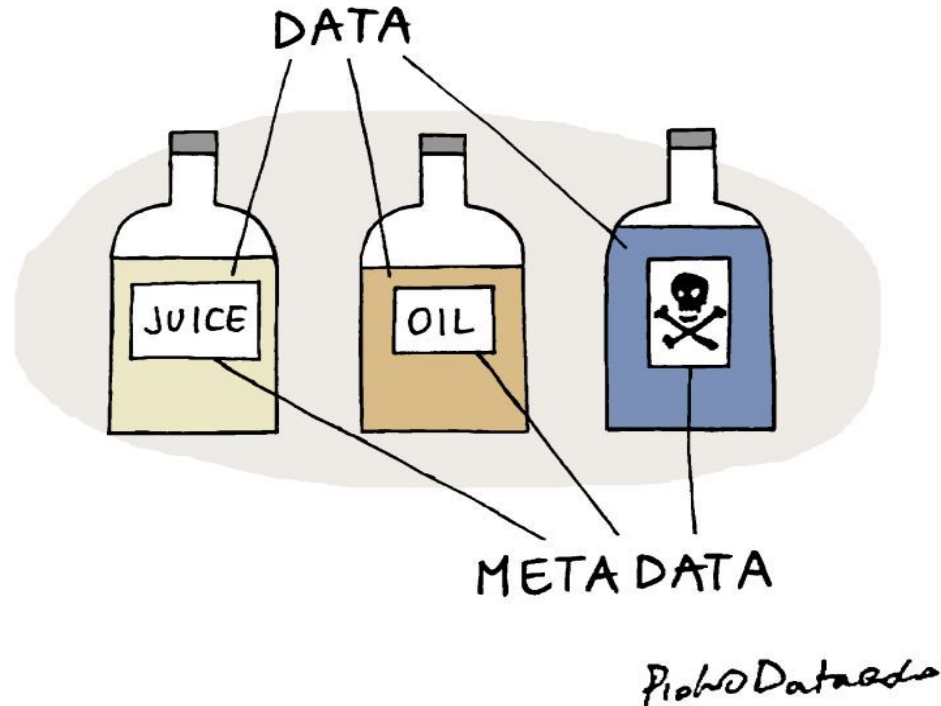
The information we create, store, and share to describe things

ie. information about the item’s creation, name, topic, features, and the like

They allow us to interact with these things to obtain the knowledge we need

Metadata is key to the functionality of the systems holding the content, enabling users to find items of interest, record essential information about them, and share that information with others.

The difference between data and metadata



‘Rich’ information

‘Intrinsic’ metadata

e.g., the data captured automatically by machines

‘Contextual’ metadata

e.g., the protocol used, with both keywords and links to a formal protocol document

The measurement devices used (with both keywords and links to manufacturers)

The units of the captured data, the species involved

The ‘object’ that is the focus of the study
(genes/proteins/whatever....)

Any other details about the experiment.



Main types of metadata

Descriptive metadata	For finding or understanding a resource
Administrative metadata: <ul style="list-style-type: none">- Preservation metadata- Rights metadata	<ul style="list-style-type: none">- Long-term management of files- Intellectual property rights attached to content
Technical metadata	For instance, those captured by a device/tool/machine 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- Genre- Author- Publication date- Subject	<ul style="list-style-type: none">DiscoveryDisplayInteroperability
Technical metadata	<ul style="list-style-type: none">File typeFile sizeCreation date/timeCompression scheme	<ul style="list-style-type: none">InteroperabilityDigital object managementPreservation
Rights metadata	<ul style="list-style-type: none">Copyright statusLicense termsRights holder	<ul style="list-style-type: none">InteroperabilityDigital 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
- Disciplinary metadata
 - Data Documentation Initiative
 - Darwin Core – biological sciences

Dublin Core metadata:

Dublin Core Metadata Element Set [\[edit \]](#)

The original DCMES Version 1.1 consists of 15 metadata elements, defined this way in the original specification:^{[6][14]}

1. Contributor – "An entity responsible for making contributions to the resource".
2. Coverage – "The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant".
3. Creator – "An entity primarily responsible for making the resource".
4. Date – "A point or period of time associated with an event in the lifecycle of the resource".
5. Description – "An account of the resource".
6. Format – "The file format, physical medium, or dimensions of the resource".
7. Identifier – "An unambiguous reference to the resource within a given context".
8. Language – "A language of the resource".
9. Publisher – "An entity responsible for making the resource available".
10. Relation – "A related resource".
11. Rights – "Information about rights held in and over the resource".
12. Source – "A related resource from which the described resource is derived".
13. Subject – "The topic of the resource".
14. Title – "A name given to the resource".
15. Type – "The nature or genre of the resource".

Each Dublin Core element is optional and may be repeated. The DCMI has established standard ways to refine elements and encourage the use of encoding and vocabulary schemes. There is no prescribed order in Dublin Core for presenting or using the elements. The Dublin Core became a NISO standards, Z39.85, and IETF RFC 5013 in 2007, ISO 15836 standard in 2009 and is used as a base-level data element set for the description of learning resources in the [ISO/IEC 19788-2 Metadata for learning resources \(MLR\) – Part 2: Dublin Core elements](#), prepared by the [ISO/IEC JTC 1/SC 36](#).

Full information on element definitions and term relationships can be found in the Dublin Core Metadata Registry.^[15]

Encoding examples [\[edit \]](#)

```
<meta name="DC.Format" content="video/mpeg; 10 minutes" />
<meta name="DC.Language" content="en" />
<meta name="DC.Publisher" content="publisher-name" />
<meta name="DC.Title" content="HYP" />
```

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-director>
y/



Registries and other tools for findability

- OpenDOAR <https://v2.sherpa.ac.uk/opensoar/>

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!



Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource



Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
 - A1.1 The protocol is open, free, and universally implementable
 - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available



Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data



Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
 - R1.1. (Meta)data are released with a clear and accessible data usage license
 - R1.2. (Meta)data are associated with detailed provenance
 - R1.3. (Meta)data meet domain-relevant community standards

All the 15 principles

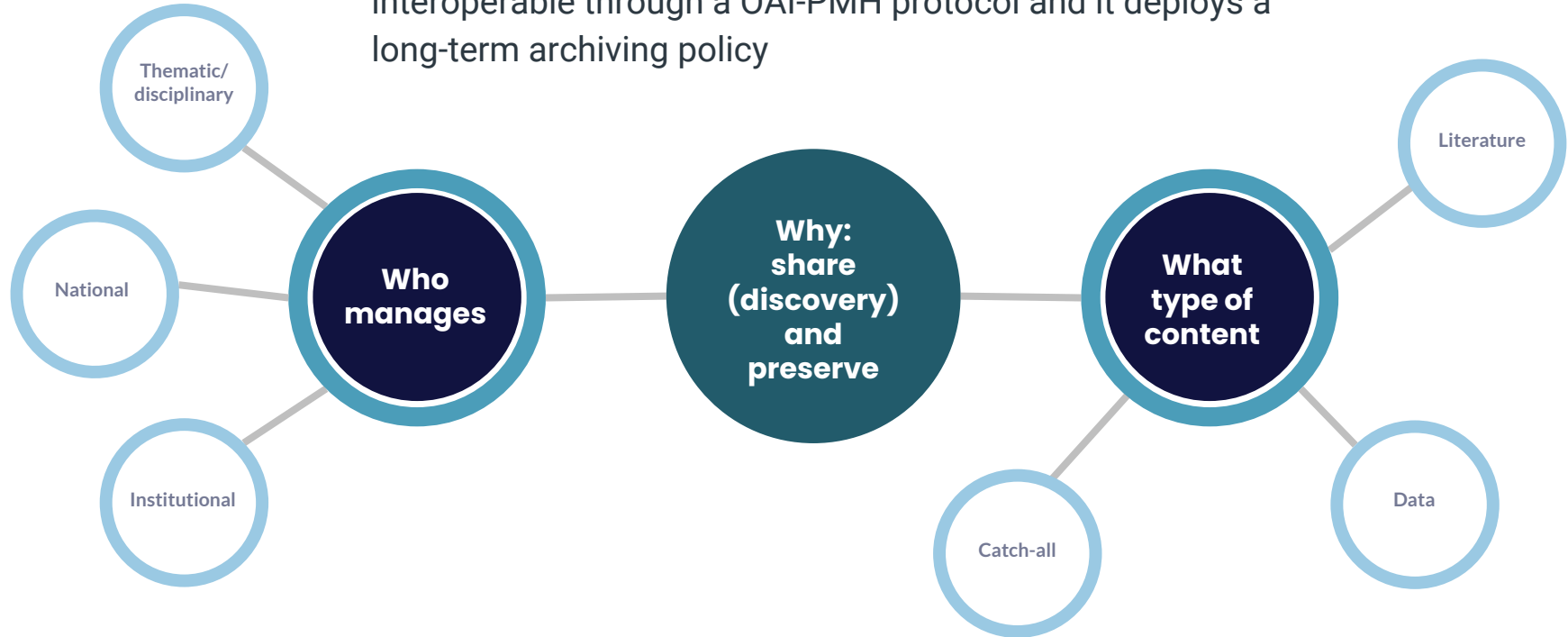
They are principles,
not standards!

Some "A" aspects to consider

- Explain which data can't be shared openly, if any
- Specify how access will be provided in case of restrictions, e.g. through a data committee, a license, or arranged with the repository
- Will methods or software tools needed to access the data (if any) be included or documented?
- Deposit the data and associated metadata, documentation and code preferably in certified repositories which support Open Access

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







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 spxd.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



Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource



Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
 - A1.1 The protocol is open, free, and universally implementable
 - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available



Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data



Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
 - R1.1. (Meta)data are released with a clear and accessible data usage license
 - R1.2. (Meta)data are associated with detailed provenance
 - R1.3. (Meta)data meet domain-relevant community standards

All the 15 principles

They are principles,
not standards!

Interoperable

- The data usually need to be integrated with other data, also with automated processes.
- The data need to interoperate with applications or workflows for analysis, storage, and processing (thus facilitating reuse).
- Mapping data to common standards makes it easier to share across disciplines

Data need to be (easily) integrated

- Ready to be combined with other datasets by humans as well as computer systems
- Use formal, broadly applicable languages, use standard vocabularies, qualified references...



Interoperable

Main aspects to consider:

- vocabularies
- references to other metadata

An example



Making assumptions

Date	Temp
28/05/2022	200
29/05/2022	195
30/05/2022	197

Determining

Date (DD/MM/YYYY)	Temp (K)
28/05/2022	200
29/05/2022	195
30/05/2022	197

Machines do not make assumptions! Using standardised formats, a computer can interpret the data even if it hasn't encountered it before

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.

	A	B	C	D
1	Attribute	Attribute Description	Example	
2	Study Title	Title of your experiment	Genes in response to drug	
3	Study Description	What your experiment is about (hypothesis, methods etc...)	Genes expression in liver of mice treated with DrugY	
4	Person Name	Name Surname	My Name	Your Name
5	Person Role	Performer, Supervisor etc...	Performer	Supervisor
6	Experimental factor(s)	Independent variable manipulated by the experimentalist	Drug	
7			PBS	
8	Experimental factor levels	Groups you want to compare	DrugY	
9	Replication type(s)	Experimental Unit, Observational Unit, Independent replicates, pseudoreplicates etc...	Experimental Unit	Observational Unit
10	Replication description	What your replicates are	Mouse	Liver
11	Organism	Species according to one taxonomy	Mm	
12	Genotype	Genetically modified or not	geneX_WT	
13	Date	Format as YYYYMMDD	YYYYMMDD	
14	Extract_Material	Extracted material type	RNA	
15	LibraryLayout	Single end or paired end (SE, PE)	SE	
16	MillionReads_Asked	How many reads are needed		10
17	ReadsLength_bp	Reads length in bp		50
18				

Controlled Vocabulary for Drug

Figure 1. Metadata fields description and controlled vocabulary.

Zenodo is integrated into reporting lines for research funded by the European Commission via [OpenAIRE](#). Specify grants which have funded your research, and we will let your funding agency know!

Grants

European Commission (EU)

OpenAIRE-Advance 777541 OpenAIRE Advancing Open Scholarship

European Commission (EU)

EOSCsecretariat.eu 831644 EOSCsecretariat.eu

Optional. OpenAIRE-supported projects only. For other funding acknowledgements, please use the **Additional Notes** field.
Note: a human Zenodo curator will need to validate your upload - you may experience a delay before it is available in OpenAIRE.

[+ Add another grant](#)

Related/alternate identifiers

recommended

Specify identifiers of related publications and datasets. Supported identifiers include: DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs and URLs.

Related identifiers

10.5281/zenodo.380176

continues this upload

Presentation

Optional. Resource type of the related identifier.

10.5281/zenodo.382618

continues this upload

Presentation

Optional. Resource type of the related identifier.

10.5281/zenodo.390163

continues this upload

Presentation

Optional. Resource type of the related identifier.

[+ Add another related identifier](#)

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)

PIDs

Communities:

Open Science in Italy

License (for files):

[Creative Commons Attribution 4.0 International](#)



Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource



Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
 - A1.1 The protocol is open, free, and universally implementable
 - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available



Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data



Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
 - R1.1. (Meta)data are released with a clear and accessible data usage license
 - R1.2. (Meta)data are associated with detailed provenance
 - R1.3. (Meta)data meet domain-relevant community standards

All the 15 principles

They are principles,
not standards!

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Some hints for reusability

- Describe the **scope of your data**: for what purpose was it generated/collected?
- Mention any **particularities or limitations** about the data that other users should be aware of.
- Specify the date of **generation/collection** of the data, the lab conditions, who prepared the data, the **parameter settings, the name and version of the software** used.
- Is it **raw or processed** data?
- Ensure that all **variable names** are explained or self-explanatory (i.e., defined in the research field's controlled **vocabulary**).
- Clearly specify and document the **version** of the archived and/or reused data.

Reusable

Main elements to include:

- documentation
- licences
- quality assurance

Some "R" aspects to consider

- License the data to permit the widest reuse possible
- Specify a data embargo, if this is needed

Beware: embargo is no longer an option in HE!

- How long will the data remain reusable?

Check the repository policies

- Describe data quality assurance processes

Documentation

Consider how you will capture this information and where it will be recorded:

- Data collection methodology;
- Analytical and procedural information;
- Definitions of variables and units of measurement;
- Assumptions made and quality indicators;
- Software used to collect and/or process the data.

In short: document and preserve everything that is needed to reproduce the study – ideally following the standard in your discipline

Distinguishing metadata and documentation

Documentation

- Can be informal.
- Often created while working on a project.
- May cover many levels (project, datasets, data files, variables and values).
- May provide general context.

Metadata

- Formally describes a particular object (can be a data file or dataset).
- Often created upon “publication” and linked to an object.
- Formatted into a record and structured according to standards.
- May derive from the documentation.

So many ways to describe your data

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



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>

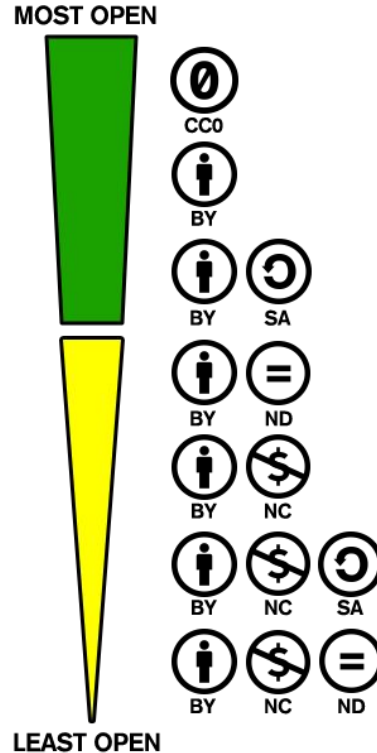
Licences

The conditions under which the data can be used should be clear to machines and humans. This has to be specified in the metadata describing a data set.

Include information about the license in the metadata. If a particular license is needed, you have to provide it along with the data set. Where possible it is suggested to use common licenses, such as CC 0, CC BY, etc., which can be referred to by URL.



Open Science aims at CC0 or CC-by



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...

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>

Sharing data: what is ment?

With collaborators while the research is active



Data are mutable

(Open) data sharing



Data are stable, searchable, citable,
clearly licensed

Storing versus archiving

Storing and backing up files while research is active



Likely to be on a networked filestore or hard drive
Easy to change or delete

Archiving or preserving data in the long-term



Data are stable, searchable, citable, clearly licensed

Likely to be deposited in a digital repository
Safeguarded and preserved

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

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

Ethics

Legal and ethical issues

Especially when the research is handling personal or sensible data

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?

Do you plan to use data subject to any kind of protection (e.g. commercial agreements or Non-Disclosure Agreements)?

Could some of the data be used for patents?

Ethical aspects

Main aspects to include in the DMP:

- How personal and special categories of data will be protected
- Management of informed consent: free, specific, informative and unambiguous
- Data flows: principle of equivalence, Intra EU/EEA transfers/Outside EU/EEA

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 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"(art. 2 n.5 GDPR). Directly identifying data is held separately and securely from processed data to ensure non-attribution.

Other research outputs

Other research outputs

In the research work a variety of outputs can be produced (not only papers):

of course data, but also for example **software, models, workflows, new materials, antibodies, reagents, samples, project deliverables, presentations, training materials and all grey literature, lab notes...**

All these can be FAIR – with PIDs, documentation, accessibility in a repository (at least the metadata), choosing proper file formats and licences and so on.

Any research output or instrument needed to validate the conclusion of a publication should be FAIR (this is a requirement in HE!)

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.

Software is a key ingredient of OS

The paper is not enough for reproducibility, but sometimes even for understanding



In all disciplines

From physics to biology, from mathematics to linguistics, from law to social sciences, computer programs are used everywhere. In order to understand, replicate, verify and reuse a research result, it is necessary to have access to the article that describes it, as well as to the data and the computer programs used to obtain it.

Software sharing



+



GitHub repositories can be deposited in Zenodo. This makes the repositories easier to reference in academic literature, creating persistent identifiers (DOIs).

<https://guides.github.com/activities/citable-code/>



Software Heritage

Long term, non for profit, shared infrastructure that collects, preserves and makes easily available the source code of all software that is publicly available, we are actually contributing to build the long overdue software pillar of Open Science.

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

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
- 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


**Utrecht
University**

Search uu.nl


q

Nederlands

Home > Research > Research Data Management Support > Guides > Costs of data management



Research Data Management Support

 **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.

Roles and responsibilities

RDM encourages heterogeneous stakeholder groups to work together for a shared societal goal. It's worth bearing in mind that RDM and data management planning similarly involve multiple stakeholder types:

- **The principal investigator** – ultimately responsible for the data and for data management
- **Researchers, research assistants and/or data managers** – involved in day-to-day data management
- **The institution's management** – draft and enforce data policies; raise data awareness
- **The institution's research office consisting of library, IT and legal services** – provide external data, tools, secure storage and access; expertise on rights management and ethics, data citation, metadata, access and licenses, funder requirements; raise data awareness
- **Research funders** – encourage good data practices; invest in data infrastructure; raise data awareness
- **Project partners** in academic and other research institutions as well as commercial partners
- **Academic publishers** – impose requirements on the availability of data underlying submitted and/or published papers; provide identifiers to cite papers and link to related data
- **Research data repositories** – preserve data long term; provide persistent identifiers and data discovery service

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?

Keep everything for always?

Select what data you'll need and want to retain.

- When regenerating data would be cheaper than archiving, don't archive.
- If you used secondary data and they are already available somewhere (if possible with a PID), there is no need to archive them again.
- 10 years is often stated in data policies and academic codes, but data can be valuable for ages, in climatology, sociology, health sciences, astronomy, linguistics, ...Look beyond minimal retention periods where relevant.
- Explain your selection criteria in the DMP.

Justify your decisions!

Of all the choices done – softwares, formats, standards, methodologies etc. – you need to demonstrate that the selections made are the most appropriate for your context, that of your discipline and future users.

Similarly, you need to present a convincing case for any restrictions on data sharing.

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

DCC Checklist

https://dmponline.dcc.ac.uk/files/DMP_Checklist_2013.pdf



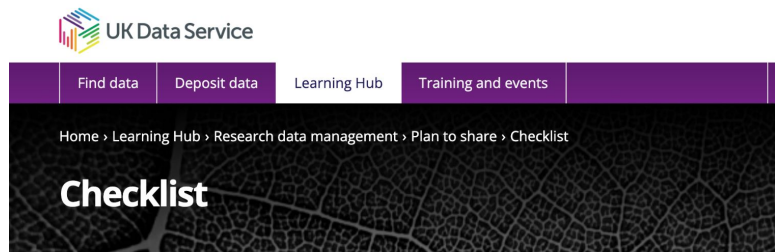
Checklist for a Data Management Plan, v4.0

Please cite as: DCC. (2013). *Checklist for a Data Management Plan*. v.4.0. Edinburgh: Digital Curation Centre. Available online: <http://www.dcc.ac.uk/resources/data-management-plans>

DCC Checklist	DCC Guidance and questions to consider
Administrative Data	
ID	A pertinent ID as determined by the funder and/or institution.
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable [POST-AWARD DMPs ONLY]
Project Name	If applying for funding, state the name exactly as in the grant proposal.
Project Description	<p>Questions to consider:</p> <ul style="list-style-type: none"> - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the data being collected or created? <p>Guidance:</p> <p>Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.</p>
PI / Researcher	Name of Principal Investigator(s) or main researcher(s) on the project.
PI / Researcher ID	E.g ORCID http://orcid.org/
Project Data Contact	Name (if different to above), telephone and email contact details
Date of First Version	Date the first version of the DMP was completed
Date of Last Update	Date the DMP was last changed
Related Policies	<p>Questions to consider:</p> <ul style="list-style-type: none"> - Are there any existing procedures that you will base your approach on? - Does your department/group have data management guidelines? - Does your institution have a data protection or security policy that you will follow? - Does your institution have a Research Data Management (RDM) policy? - Does your funder have a Research Data Management policy? - Are there any formal standards that you will adopt? <p>Guidance:</p> <p>List any other relevant funder, institutional, departmental or group policies on data management, data sharing and data security. Some of the information you give in the remainder of the DMP will be determined by the content of other policies. If so, point/link to them here.</p>
Data Collection	
What data will you collect or create?	<p>Questions to consider:</p> <ul style="list-style-type: none"> - What type, format and volume of data? - Do your chosen formats and software enable sharing and long-term access to the data? - Are there any existing data that you can reuse? <p>Guidance:</p> <p>Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.</p>
How will the data be collected or created?	<p>Questions to Consider:</p> <ul style="list-style-type: none"> - What standards or methodologies will you use? - How will you structure and name your folders and files?

UK data service checklist

<https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/checklist/>



Using a data management checklist

A checklist, such as the one below, can help with writing a data management plan, as it helps you identify what actions to take to optimise data sharing.

Planning



Documenting



- Will others be able to understand your data and use them properly?
- Are your structured data self-explanatory in terms of variable names, codes and abbreviations used?
- Which descriptions and contextual documentation explain what your data mean, how they were collected and the methods used to create them?
- How will you label and organise data, records and files?
- Will you be consistent in how data are catalogued?

Formatting



Storing



Confidentiality, ethics and consent



Copyright



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](#).

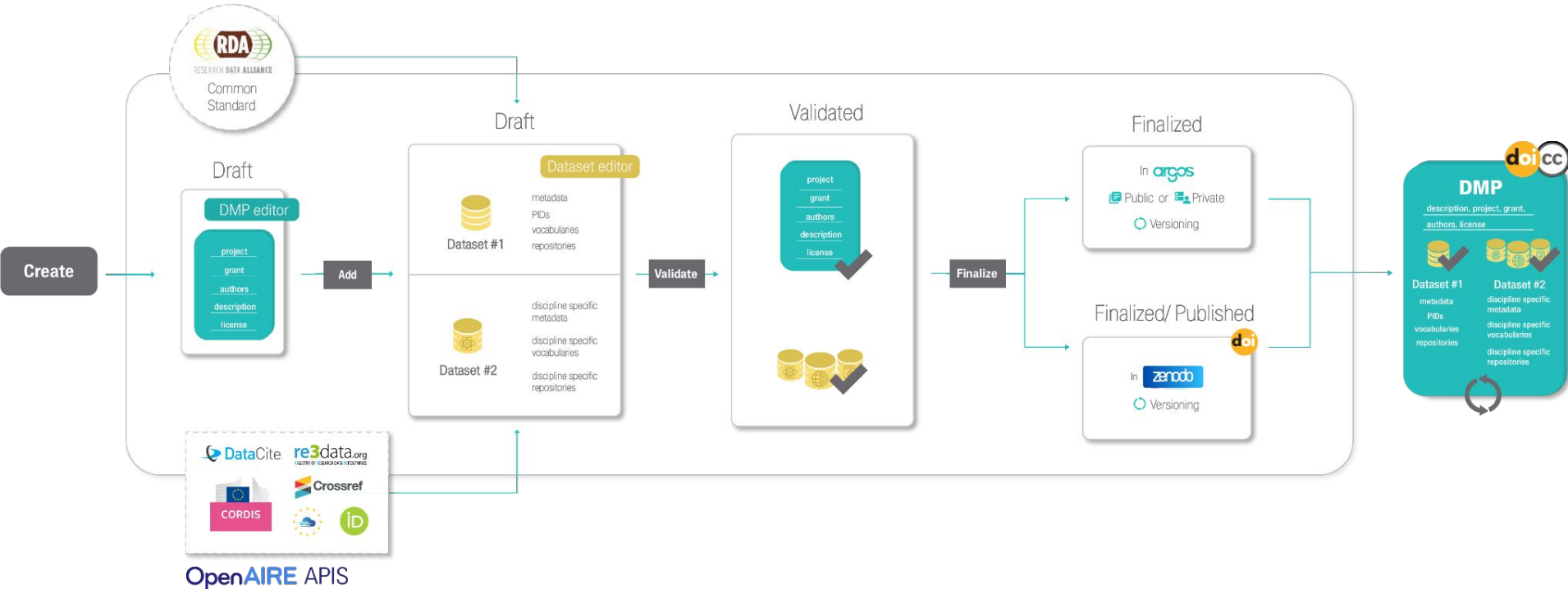
Common additional elements

- 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>

DMP tools – Argos



DMP tools – DMP online

[Home](#)[Public DMPs](#)[Funder requirements](#)[Help](#)

Plan to make data work for you

Data Management Plans that meet institutional funder requirements.



DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

DMP tools – DMP Wizard and DMPTool



Product ▾

Solutions ▾

Data Management Plans

DMPs as formal documents that outline how data were managed during and post-project have become an important and often indispensable part of grant applications, as well as a good research practice. DSW brings a complex solution for creating high-quality DMPs in any discipline.

<https://ds-wizard.org/data-management-plans>



DMPTool

Build your Data Management Plan

[Funder Requirements](#)



<https://dmptool.org/>

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/>

- **Public DMPs on DMP tool**

https://dmptool.org/public_plans

- **Tools to create your DMP**

argos.openaire.eu

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

- **Re3Data**

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

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

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

- **Research Data Management decision tree**

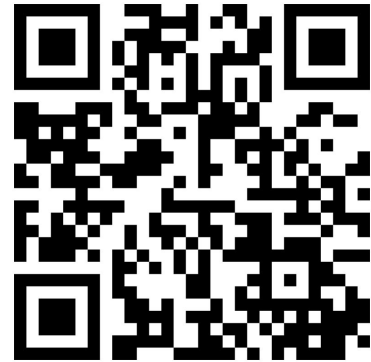
<https://zenodo.org/record/7190005#.Y3s5quzMI-Q>

Interaction

Go to:

<https://www.menti.com>

Voting code: 6844 5122



THANK YOU

GET IN TOUCH

gina.pavone@isti.cnr.it