ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

University Guidelines for Research Data Management

Guidelines prepared by:

Alma Mater Studiorum – Università di Bologna

ARIC – Research Division, Research Services and Division Projects Coordination Unit, Data Stewards

With the support of:

ARIN – Innovation Division, Knowledge Transfer Office Unit

ARPAC – Cultural Heritage Division, University Digital Library Management and Development Unit – AlmaDL; Electronic Resource Library Management and Development Unit – AlmaRE

GLOS – Open Science Working Group

Departmental Open Science Representatives

SSRD – Executive Support Services, “Personal Data Protection” Unit

APPC – Planning and Communication Division – Communication Unit – Graphic Design for Communication Office

These Guidelines accompany *The University Policy on Research Data Management*, which lays down the criteria and principles to follow in order to manage data properly and consciously, in line with international standards and domain-specific peculiarities.

This document is intended for **all researchers of the University of Bologna**, regardless of their career level or disciplinary domain.

To manage means **to take care of research data and organise them carefully throughout the research cycle**, with a view to:

- making the research process as efficient as possible;
- making data interpretable, understandable and findable over time;
- fostering research integrity;
- encouraging cooperation with other researchers.

Data management needs to be **carefully planned** when starting research, accompanies all active phases of **data production**, **collection and analysis**, up to the **preservation** (i.e. long-term archiving) and, ideally, the **sharing** of research data.

These Guidelines present, one by one, the most significant aspects of the data management process, providing procedural instructions and useful tools for every step.

More details about the topics covered by these Guidelines are included in separate fact sheets. Reference to the fact sheets is made in the text as follows: **Fact sheet title**.

Useful links

The University Policy on Research Data Management:
https://www.unibo.it/en/university/who-we-are/open-access-and-open-science

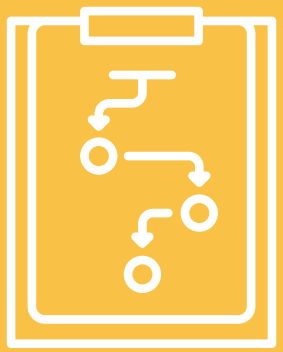
University of Bologna's playlists (in Italian):

- "Dati: conoscerli e gestirli per valorizzare la ricerca"
  https://www.youtube.com/playlist?list=PLaUmBQ7P5K-AyDDnv1f8upAyEOtAF2gj3
- "Open Access e Open Science"
  https://www.youtube.com/playlist?list=PLaUmBQ7P5K-A83TlY96DyUl6t3rCryRK_

Research Data Management Decision Tree:
https://doi.org/10.5281/zenodo.7190005

TU Delft- Research Data Management 101 (RDM101) playlist:
https://www.youtube.com/playlist?list=PLdHnT1NHNDE-A9wLpAfho_Wum9Xw6Q0_E

PLANNING DATA FLOWS



Main steps in this phase

- **Identify data** types,
  - **deciding whether to generate new data** and/or **reuse** available data from existing sources;
  - being aware of the ethical principles and privacy and intellectual property regulations to be complied with.
- **Identify the most important metadata** to describe the data you generate; when reusing existing data, pay close attention to the associated metadata.
- **Plan** how you will organise your data into **datasets.**
- **Prepare a Data Management Plan** to keep track of your choices.

Identifying data types

Research data is any **information**, in any format, which is used within a specific **research activity** and is necessary for the **validation of research results**. Data can come in a variety of types and is often identified and classified depending on the disciplinary domain **Research data: types, formats, methods**.

Right at the start of your research, you should:

- **Identify the data types** you will work with: to do this, consider every phase of your research to pinpoint **what type of information you will need to collect and/or use** to answer your research questions.
- **Always check for any existing data**, published by other researchers, that could be useful to answer your research questions.

If you produce new data:

- Always record the metadata related to the data and to their creation process (see Identifying essential metadata).
- If your research involves vulnerable people, animals or certain technologies (artificial intelligence, dual-use technology, etc.), make sure to manage your data ethically and that you are duly authorised to collect them. Get in touch with the competent ethical committees, where appropriate.
- If your data contain personal data, make sure that you take the necessary precautions **Respecting privacy**.
- Consider if there are other reasons to keep your data confidential, e.g. commercial exploitation of related results or signed agreements with third parties.

If you reuse existing data:

- Carefully check the associated metadata, which contains essential information for proper reuse (see Identifying essential metadata).
- Through the metadata, find out whether you can reuse the data for your own purposes or there are restrictions imposed by the authors **Copyright**.
- Always make sure to reuse the data in compliance with ethical principles and intellectual property and privacy regulations.

Useful links

"Call them Data Stewards: specialists in research data management at the University of Bologna": https://magazine.unibo.it/archivio/2023/06/16/call-them-data-stewards-specialists-in-research-data-management-at-the-university-of-bologna

Video "Data steward all'Università di Bologna"(in Italian) https://youtu.be/6Ic0isyefs8?si=JqCFzHIh2Hz1U2ee

More resources on research data:

- The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-data
- Research Data Alliance https://www.rd-alliance.org/
- Digital Curation Centre (DCC) https://www.dcc.ac.uk/

Identifying essential metadata

Metadata is any **structured information** that accompanies research data (see Identifying data types). It is 'data about data', makes it easier to understand and reuse data and allows search engines and aggregators to index them **Metadata and documentation**.

Metadata is usually structured according to **standard schemas**, often **domain-specific**, implemented by infrastructures for long-term data archiving and access **Repositories**. Using controlled vocabularies ensures that the information included in the metadata is even more understandable and interoperable **Using different types of standards**.

Remember to:

- Identify the most appropriate **metadata schema** as soon as possible. Remember that this choice may be influenced by the repository where you will archive your datasets in the long term.
- Identify the **information** required by the metadata schema you are using and make sure to collect it throughout your research.
- Where possible, remember to use **controlled vocabularies**.
- When you deposit your data in a repository, remember to include the following key information in the metadata: **names and affiliations** (e.g. Alma Mater Studiorum – Università di Bologna) of anyone who contributed to creating the dataset in order to document authorship, a **Persistent Identifier**, or PID in short (e.g. Digital Object Identifier, or DOI), and the **type of licence** associated with the dataset (see also Depositing data according to FAIR principles).
- Remember: any information that cannot be structured in the metadata but that is needed in order to understand your data should be associated with the datasets as additional **documentation** (see Gathering documentation).

Useful links

Videoclip "5 Minute Metadata- What is metadata?" https://www.youtube.com/watch?v=L0vOg18ncWE

Resources to search for metadata schemas:

- FAIRsharing Standard Registry https://fairsharing.org/search?fairsharingRegistry=Standard
- RDA Metadata Standards Directory https://rd-alliance.github.io/metadata-directory/standards/
- DCC Guidance on Disciplinary Metadata https://www.dcc.ac.uk/guidance/standards/metadata

Planning how to organise data into different datasets

A dataset is a **structured set of related data** (see Identifying data types), in any format. Data are usually organised into datasets if they share a common goal (i.e. they answer the same research question) or reflect the results of a research activity **Datasets**. Datasets are crucial to ensure the quality and usefulness of research results. A well-organised and well-managed dataset is always accompanied by informative metadata (see Identifying essential metadata).

Right at the start of your research, remember to:

- **Organise** your data in a meaningful way.
- Use a **clear naming system** for the files and folders that will form your dataset, to clarify as much as possible their content and the relations between them.
- Start **gathering the documentation** you will need to make your datasets as understandable and reusable as possible (see Gathering documentation).
- Start thinking about which repository/ies may be most suitable to **archive your datasets** in the long term (see Choosing the most appropriate repository). The content and size of your dataset may limit your choices.

Drafting a Data Management Plan

A Data Management Plan (or DMP in short) is the main tool for **documenting all data management choices** made during a project. It is usually a text document and should be drafted **at a very early stage**. A DMP is also contractually required by many funding bodies.

- Start writing your DMP as soon as you begin thinking about how you will manage your data.
- Remember that data management strategies **evolve during the research** and that your DMP must **be kept up to date**.

In the document:

- Give a **detailed overview of your research data**, both new and reused data (see Identifying data types).
- Specify the **methodologies, tools and software** you employ to collect, create and analyse your data **Managing software**.
- Identify the strategies you will implement to ensure **data quality** and avoid inaccuracies or inconsistencies (see Ensuring data quality).
- Include **data storage** strategies, e.g. whether you share data with collaborators or make backup copies (see Saving data in appropriate storage spaces).
- Describe **long-term archiving strategies** (see Choosing what data to deposit and Choosing the most appropriate repository) and how you apply **FAIR principles** to your data (see Depositing data according to FAIR principles).
- Describe **roles and responsibilities** within your research team.
- Document data management **costs**, also in terms of time spent.
- Cover the aspects of data management related to **privacy**, **intellectual property rights** and **ethics** **Respecting privacy** **Copyright**.

Useful links

Video “Dati: conoscerli e gestirli per valorizzare la ricerca. Il Data Management Plan” (in Italian)
https://www.youtube.com/watch?v=SlOSrQdrhtQ

Video “RDM101- Module5: Data Management Plan”
https://youtu.be/28rTTRFDq58?si=E1ZGfzU_Q8G7p2c_

DMP preparation guidelines. Science Europe Templates and Guidelines
https://scienceeurope.org/our-priorities/research-data/research-data-management/

Online tools for preparing your DMP:
- Elixir Data Stewardship Wizard https://ds-wizard.org/
- DCC DMPonline https://dmponline.dcc.ac.uk/
- ARGOS https://argos.openaire.eu/home

Data Management Plan templates:
- Horizon Europe Data Management Plan Template
  https://www.openaire.eu/images/Guides/HORIZON_EUROPE_Data-Management-Plan-Template.pdf
- Science Europe Data Management Plan Template
  https://scienceeurope.org/media/411km040/se-rdm-template-3-researcher-guidance-for-data-management-plans.docx

DATA COLLECTION AND ANALYSIS

Main steps in this phase

- **Save** data in appropriate **storage spaces** and make the necessary **backup copies**.
- **Ensure** data **quality** through methodical processes and by
  - **keeping track** of file **versions**;
  - **organising files and folders** hierarchically, and naming them consistently;
  - **choosing the most appropriate formats** for your data, preferring standard and open formats whenever possible to facilitate interoperability and reusability.
- **Include** in the **documentation** all information necessary to understand and interpret your data.

Saving data in appropriate storage spaces

To **store** data (see Identifying data types) means to retain them in the **short or medium term**, e.g., during the active phases of the research process. **Storage systems** include external hard drives, cloud services and servers.

During your research:

- Plan how you will store your data early on, deciding **where to save it** and **budgeting** any potential **costs.**
- Remember that the choice of the most appropriate storage system depends on the **nature of the data**, **its volume** and how often different people **collaborate** on the same files.
- Make **backup copies** of your files on a regular basis, on different storage systems, to avoid data loss.
- If you need to share data with your research team, use storage platforms that allow for remote access and set **each member's access rights to files and folders** right from the start.
- Make sure that the storage infrastructure you choose allows you to monitor any **changes** to the data, who made them, as well as to recover any **previous versions**. These aspects are crucial in collaborative research contexts.
- Make sure to protect your data by **updating** your own personal **passwords** regularly and by keeping your computer's **antivirus** up to date.
- Remember that **managing and storing personal data** requires an additional level of protection, for which you can use a dedicated password or software that encrypts files and folders **Respecting privacy**.

Useful links

More on choosing the storage system:

- The Research Data Management Toolkit https://rdmkit.elixir-europe.org/storage#what-features-do-you-need-in-a-storage-solution-when-collecting-data
- The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-storage

File and folder encryption tools:

- Veracrypt https://www.veracrypt.fr/en/Home.html
- BitLocker https://docs.microsoft.com/it-it/windows/security/information-protection/bitlocker/bitlocker-overview

Data management budgeting tools:

- OpenAIRE RDM Cost Calculator https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs
- UK Data Service RDM Cost Calculator https://ukdataservice.ac.uk/app/uploads/costingtool.pdf

Video "Dati: conoscerli e gestirli per valorizzare la ricerca. Salvare e condividere i dati" (in Italian)
https://youtu.be/VQ0yK0tQ1N4?feature=shared

Ensuring data quality

Ensuring data quality is a domain-specific process and as such goes beyond the scope of this Guidelines. However, it begins with a **set of methodical processes** that allow your data to be traced, used, and reused properly.

During your research:

- Choose a **clear organisation for folders and files** and keep track of ensuing versions.
- Choose a clear and legible file **naming standard**. It is a good idea to enter the name of the author or the origin of the data, creation date and version number, without using spaces or special characters (e.g. FocusGroup1_20240502_v2.rtf).
- Choose the **most appropriate format** for your data, preferring standard and open formats whenever possible to facilitate interoperability and reusability **Research data: types, formats, methods**.
- **Validate and verify your data** to avoid inaccuracies, incompleteness or inconsistencies. Strategies include data entry validation, data interval control, removal/recording of inaccurate or missing variables, control of consistent data scales.
- Define a **standard methodology and workflows** for data analysis and processing, especially in collaborative research contexts. For example, you can define which data to save in which folders, how and when to document the data, or which systems to use to share data with collaborators (see Saving data in appropriate storage spaces).
- If your data is related to any other research results, describe this relationship in the documentation and **provide a full citation** (e.g. a dataset derived from an existing dataset) (see Gathering documentation).

Useful links

More on data quality control strategies:

- The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-data-curation
- The Research Data Management Toolkit https://rdmkit.elixir-europe.org/data_quality

File naming tools:

- Bulk Rename Utility (Free File Renaming Utility for Windows) https://www.bulkrenameutility.co.uk/
- File Naming Conventions https://www.data.cam.ac.uk/data-management-guide/organising-your-data#Naming

Methodology standardisation and sharing tools:

- Protocol Manager https://protocols.io

Gathering documentation

Your data and datasets (see Planning how to organise data into different datasets) can only be **intelligible and interpretable by others** if they are accompanied by additional documentation. For example, a README file is a free text document (i.e. human-readable) included within the dataset that **explains the data origin and how it is organised**.

During your research:

- Draw up the documentation during all active phases of data collection and analysis.
- Document the **data** that makes up the dataset, the **relationships** between data, and their **origin**.
- Provide comprehensive information about the **methodologies** (protocols, technical specifications, tools used) **applied** for data collection and/or reuse and/or generation.
- Document **quality assurance processes** during data generation and analysis (see Ensuring data quality).
- Provide information about any **tools or software needed** to open, read or interpret your data **Managing software**.
- **Archive** the documentation **together with the data** at the time of deposit in a repository (see Choosing what data to deposit).
- Save the documentation in an open and accessible file format (e.g. .rtf, .md).

Useful links

CESSDA Data Management Expert Guide. Documentation and Metadata
https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

Utrecht University. Research Data Management Support Guides. Metadata and Documentation
https://www.uu.nl/en/research/research-data-management/guides/during-research/metadata-and-documentation

DATA PRESERVATION AND SHARING

Main steps in this phase

- **Choose what data to deposit** to ensure the best understanding, transparency and reproducibility of your research.
- **Choose the most appropriate repository** for your data to ensure its preservation and sharing in the long term.
- **Deposit your data** organised in datasets and according to **FAIR principles**.
- **Associate** your data with **a licence** that allows **as wide a reuse as possible**, unless you are limited by commitments towards third parties or by commercial exploitation strategies.

Choosing what data to deposit

To deposit data means to archive them in a digital infrastructure designed for their long-term preservation and called a **repository** (see Choosing the most appropriate repository). At this stage, data are **organised in datasets** (see Planning how to organise data into different datasets). The deposit can occur at the end of a research activity but needs to precede the publication of its results in a scientific article.

Deposit ensures **data preservation** and **visibility** well beyond the end of the research project that generated it.

During your research:

- **Select** which data to deposit to allow for the validation of your conclusions and the reproducibility of your research. Remember that the researcher who generates the data is **responsible for depositing** it.
- **Remember to deposit**, for example, original datasets or software, raw data obtained from the analysis of physical samples, and observational data that cannot be re-obtain **Managing software**.
- **It is not necessary to deposit** data that is easy to re-obtain or that is too large compared to its actual usefulness.
- **Do not deposit** data that is already available, e.g. data you are reusing because someone else deposited it first.
- In any case, carefully **document** the **origin** of the data you deposit, as well as the **methodologies** with which it was produced and managed (see Gathering documentation).
- You can also deposit data that must remain inaccessible to third parties, for privacy, ethical or intellectual property reasons, provided you choose a suitable repository **Respecting privacy** **Copyright.**

Useful links

Video "UGent Open Science. Knowledge clip: Preserving data": https://youtu.be/UaiRAI-fwmw?si=b5YHHxNmXkUwEofM

Video "Dati: conoscerli e gestirli per valorizzare la ricerca. Conservare i dati a lungo termine" (in Italian) https://www.youtube.com/watch?v=J3VyrUzzj_E

More on long-term preservation of data:

- The Research Data Management Toolkit https://rdmkit.elixir-europe.org/data_publication
- Stanford University Library Guidelines on Data Management and Sharing https://laneguides.stanford.edu/DataManagement/
- Digital Curation Centre, How to Appraise and Select Research Data for Curation https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data

Choosing the most appropriate repository

Repositories are infrastructures for the long-term archiving of datasets (see Planning how to organise data into different datasets). They can be disciplinary, institutional or generalist in nature. Repositories can also be officially certified. In all cases, reliable repositories assign a **PID** (such as a DOI) and allow authors to associate **metadata** (see Identifying essential metadata) and licences to datasets at the time of deposit (see also Depositing data according to FAIR principles). **Copyright**

During your research:

- Use a **registry** (see Useful links) to search for one or more repositories that suit your needs and the types of data you wish to deposit.
- Find out if there is a repository **specific to your research domain**. A disciplinary repository allows you to describe your data using a domain-specific metadata schema and makes them more visible to your scientific community.
- Check if your **home institution** has one or more repositories available to its members. The University of Bologna has two institutional data repositories, AMS Acta and AMS Historica, and related support, validation and data curation services. **Repositories**.
- Keep in mind that **generalist** repositories also exist and tend to collect heterogeneous data and materials (e.g., Zenodo).
- Choose a repository with a suitable security profile and that allows for **controlled access** if your data must remain inaccessible to third parties, for privacy, ethical or intellectual property reasons **Respecting privacy** **Copyright**.
- Remember that cloud storage services (see Saving data in appropriate storage spaces), personal or project websites, and social networking platforms such as ResearchGate and Academia.edu are not repositories because they do not ensure long-term preservation of your data.
- Be aware that some publishers and journals, especially in certain domains, have created their own repositories and advise (or sometimes require) authors to deposit their data there. We recommend that you (also) publish your data elsewhere – be it in a disciplinary, institutional or generalist repository.

Useful links

Video "UGent Open Science. Knowledge clip: Data repositories". https://youtu.be/pm_C0U8ByYE?si=-Mv_dqsH66amR6Zs
Video "Dati: conoscerli e gestirli per valorizzare la ricerca. Conservare i dati a lungo termine" (in Italian) https://www.youtube.com/watch?v=J3VyrUzzj_E
Registries of repositories:

- Re3data https://www.re3data.org/
- OpenAIRE Explore https://www.openaire.eu/find-trustworthy-data-repository
- FAIRsharing Repository Database https://fairsharing.org/search?fairsharingRegistry=Database

University repositories:
AMS Acta https://amsacta.unibo.it/ | AMS Historica https://historica.unibo.it/
Further information about the University repositories: "Preserving and disseminating research data in AMS Acta" (https://sba.unibo.it/en/almadl/almadl-services/preserving-and-disseminating-research-data-in-ams-acta);
"Preservation and enhancement of the digital cultural heritage" (https://sba.unibo.it/en/almadl/almadl-services/preservation-and-enhancement-of-the-digital-cultural-heritage).
Generalist repositories:
Zenodo https://zenodo.org/ | Figshare https://figshare.com/ | Open Science Framework https://osf.io

Depositing data according to FAIR principles

Depositing data for long-term archival forms an integral part of responsible research data management, in line with FAIR principles.

First published in 2016, FAIR principles are general recommendations that aim to improve **research data reusability** by individuals and IT systems. To manage data in accordance with FAIR principles means making them **Findable, Accessible, Interoperable and Reusable** **FAIR principles**.

During your research:

- Deposit your data in a **repository**. This is the first step to making your data FAIR since every dataset (see Planning how to organise data into different datasets) is accompanied by its metadata (see Identifying essential metadata), including a PID and a licence. **Repositories** **Copyright**.
- Choose **standard and open formats**. Where possible, use vocabularies, ontologies and taxonomies to make your data understandable, interoperable and reusable **Research data: types, formats, methods** **Using different types of standards**.
- If you publish your data according to FAIR principles, you make your data, their analysis and the derived publications **more citable and exploitable**. Your research becomes more **transparent and verifiable**, in line with the requirements of a growing number of funding bodies, including the European Union.
- While FAIR data is not always freely accessible to anyone, their metadata usually are (see Associating a license with your data).

Useful links

Wilkinson *et al, The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Video "UGent Open Science. Knowledge clip: FAIR data principles". https://www.youtube.com/watch?v=2uZxFu9SFi8

More on the FAIR principles:

- GOFAIR. FAIR Principles https://www.go-fair.org/fair-principles
- FAIRsFAIR Fostering Fair Data Practices in Europe https://www.fairsfair.eu/
- How to FAIR https://howtofair.dk/

Video "Dati della ricerca: la European Open Science Cloud e i principi FAIR" (in Italian) https://www.youtube.com/watch?v=eNiHNaU6MrQ

Video "EOSC Portal. European Open Science Cloud- The New Frontier of Data-Driven Science" https://youtu.be/3HgnIe1Xu8I?si=CjR4tLRYi-vAw1bu

Associating a license with your data

Choosing a licence to associate with your datasets, in accordance with FAIR principles, allows you to specify to what extent they can be reused by others **Copyright**.

Publishing data openly fosters **collaborative research** and supports the Open Science movement. Open data is **distributed under a licence that ensures that they can be freely accessed, used, modified and shared** by anyone, requiring at most some form of attribution and integrity requirements while preserving openness.

Open Science is the movement to make **scientific research accessible without barriers to the scientific community and to citizens**. It is based on transparency, inclusion, correctness, fairness and sharing. It has been a strategic objective of the European Union since 2015, of UNESCO since 2021, and of the Italian Ministry of Research since 2022, under the *Italian National Plan for Open Science*.

Remember that the choice of licence for your data should be inspired by the 'as open as possible, as closed as necessary' principle.

In some cases, it may be appropriate to limit access to data if this is conducive to the exploitation of research for commercial purposes. This is in line with the University's mission of fostering the transfer of research results to make an impact on the economy and society.

During your research:

- Manage your data according to FAIR principles throughout the data lifecycle and prepare a **Data Management Plan** (see Drafting a Data Management Plan) **FAIR principles**.
- **Publish your data in open access unless restrictions apply** due to third-party rights or other legal provisions and provided this does not undermine opportunities for commercial exploitation of research results.
- To distribute your data openly, choose **permissive licences** that allow for any use, with any means and format and for any purpose, including commercial ones. These licences include CC0 1.0, CC BY 4.0, CC BY-SA 4.0.

Useful links

Open Science on the University website https://www.unibo.it/en/research/open-science/open-science

The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/open/open-data

Open Definition "Defining Open in Open Data, Open Content and Open Knowledge" https://opendefinition.org/od/2.1/en/

Italian national plan for open science https://www.mur.gov.it/it/atti-e-normativa/decreto-ministeriale-n-268-del-28-02-2022

Further information about copyright and cultural heritage protection and enhancement:
https://sba.unibo.it/en/almadl/almadl-services/legal-support-for-copyright-management-and-cultural-heritage-protection

Conclusions

Research data management is a set of good practices aimed at making the most of research data throughout their lifecycle, from the planning phases up to deposit and sharing.

It has advantages for researchers in terms of research quality and impact, as well as dissemination and exploitation of research results.

As we have discussed, to ensure a proper, high-quality data management process, it is crucial that each step described in this Guidelines is taken consciously and at the right time.

The first, essential one is to identify the data types you will work with. And while data management at its core is based on a set of fairly standard steps, we recognize that every research project is based on data that are extremely varied in typology, origin and use, according to domain specificities.

Contact

At the University of Bologna, research data management is supported in several ways.

If you need support about managing data and writing your Data Management Plan, please contact the Data Stewards that work within ARIC – Research Division: aric.datasteward@unibo.it.

If you need support about using the University repositories (AMS Acta or AMS Historica), please contact: almadl@unibo.it.

If you need support about copyright and related rights and about cultural heritage protection and exploitation, please contact: almadl@unibo.it.

If you need support about the commercial exploitation of the University of Bologna's research results, please contact the Knowledge Transfer Office: kto@unibo.it.

If you need support about privacy matters, please contact: privacy@unibo.it.

Checklist: The key steps for proper research data management.

PLANNING DATA FLOWS PHASE

- [ ] **Identify data types.**
  - [ ] Decide whether to **generate new data and/or reuse available data.**
  - [ ] Be aware of **ethical principles and privacy and intellectual property regulations.**
- [ ] Identify the **most important metadata**.
- [ ] Plan how you will **organise your data into datasets.**
- [ ] Prepare a **Data Management Plan.**

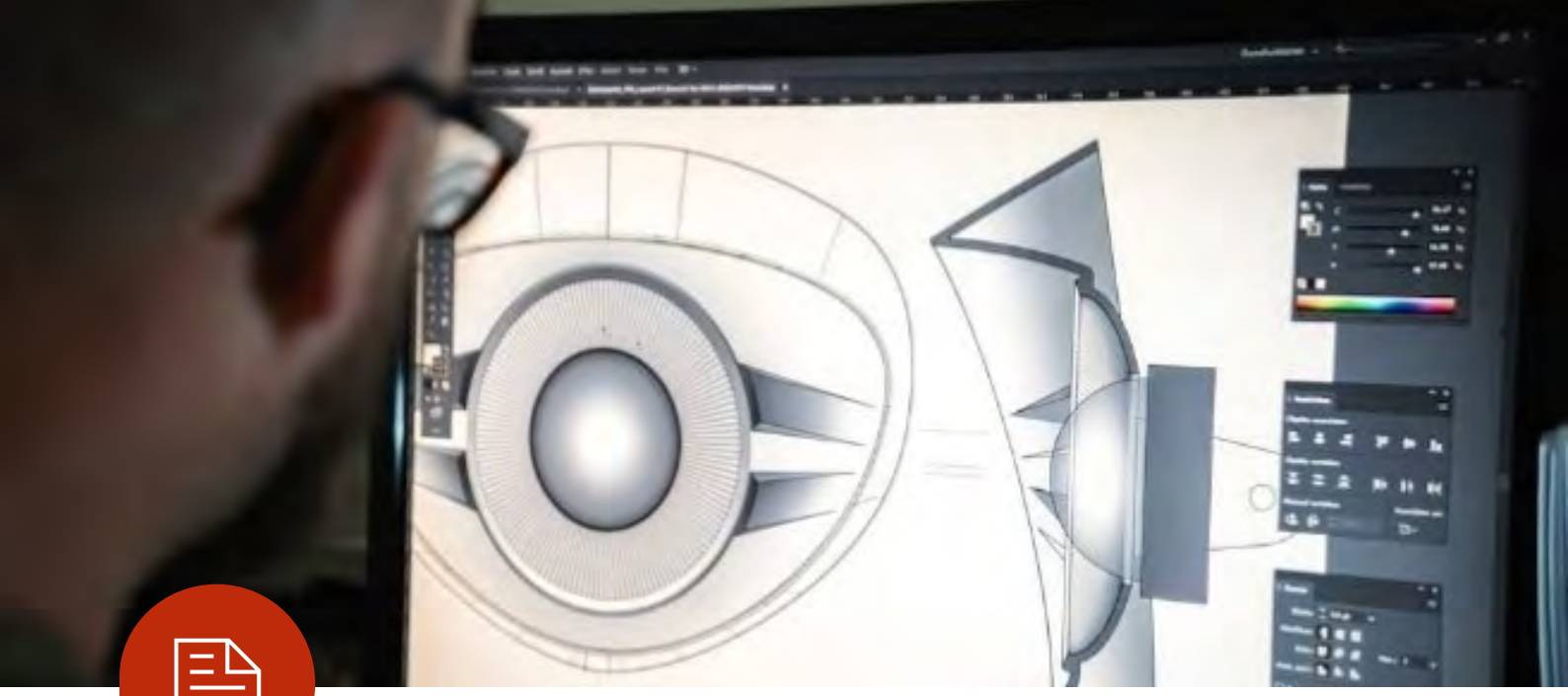
DATA COLLECTION AND ANALYSIS PHASE

- [ ] Save data in **appropriate storage spaces** and make **backup copies.**
- [ ] Ensure data **quality.**
  - [ ] Keep track of **versions.**
  - [ ] **Organise files and folders** consistently.
  - [ ] **Choose** the most appropriate data **formats.**
- [ ] Include the **documentation** necessary to understand and interpret your data.

DATA PRESERVATION AND SHARING PHASE

- [ ] Identify the **data that needs to be preserved in the long term.**
- [ ] Choose **the most appropriate repository** for deposit.
- [ ] **Deposit data** in datasets according to FAIR principles.
- [ ] Associate **the most appropriate licence** with your deposit.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

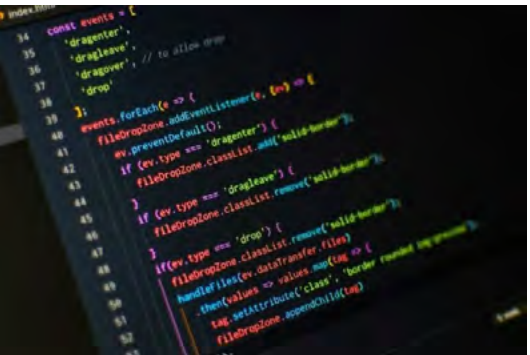


Research data: types, formats, methods

Research data are factual records collected, generated or reused as a basis for analysis, reasoning, discussion or calculation.

Data can be classified in different ways. Examples include observations, experiences, published and unpublished sources, bibliographic references, text, images, all created and/or collected in digital form, as well as other digital outputs of research, such as 3D models and source code.

Data types also vary according to the research **domain.**



In the field!

My research is theoretical in nature, and I do not produce any data. *Do these Guidelines apply to me too?*

Yes. All type of research produces or reuses data (in a broad sense), even though every disciplinary domain has its own specificities. You almost certainly use primary or secondary sources to answer your research questions and, in doing so, you produce a collection of bibliographic metadata, organised more or less systematically. In this context, always remember to use the Persistent Identifiers (PIDs) of the resources you cite and think about how you can exploit this output, for example by publishing it online as open data.

Types of data: how to categorise them

Knowing and classifying your research data allows you to choose the most effective strategies to manage them responsibly, avoid data loss or corruption, and select the most appropriate methods for data collection, archiving and analysis.

Research data can be **digital, digitised or non-digital.** While digital or digitised data management necessarily follows computerised protocols, non-digital data can be managed both digitally and non-digitally.

Whether digital or non-digital, data can be described based on its **content** – numerical, textual, audio, video, etc.

Dati con lo stesso contenuto possono avere forme diverse e quindi la loro struttura dal punto di vista digitale può cambiare. Ad esempio, dati testuali possono essere raccolti tanto nella forma di fogli di calcolo quanto nella forma di documenti di testo.

- Dati con lo stesso contenuto e raccolti nella stessa forma possono avere formati (e quindi estensioni) differenti. Ad esempio, dati numerici possono essere raccolti in un foglio di calcolo che può essere scritto in formato file "comma-separated values" (CSV) con estensione del file .csv, così come in formato OpenDocument Spreadsheet (ODS), con estensione del file `.ods`, o ancora in formato Microsoft Excel, con estensione del file `.xls` o `.xlsx`.

To ensure that your data remains accessible and reusable, it is recommended that you collect, save, share and deposit it in open, non-proprietary formats, rather than in closed, proprietary ones.

- **Proprietary format**: a format that is owned and developed by a specific company or organisation.
  - **Proprietary and closed format**: the developer of the format decides what software can use that format. For example, `.indd` for Adobe InDesign files, a software produced by Adobe for the publishing industry.
  - **Proprietary and open format**: the developer of the format has not restricted its use to a certain software. For example, MP3 exists as an open format for audio files but is patented in some countries. The XLS format was once a closed format – i.e. it could only be run by Microsoft's proprietary software, Microsoft Excel – but has since been opened. The same goes for the `.xlsx` format that is based on XML (open format) and can also be used by other software, such as LibreOffice Calc.
- **Non-proprietary and open format**: the format specifications are openly available, and anyone can create software for it. For example, CSV files for tabular data can be opened by a variety of different software.

Examples of open formats for the most common data types include:

- **Quantitative and qualitative tabular data**: SPSS (`.sav`), Stata (`.dta`), CSV (`.csv`).
- **Geospatial, vector and raster data**: ESRI Shapefile (essential – `.shp`, `.shx`, `.dbf`, optional – `.prj`, `.sbx`, `.sbn`), Geo-referenced TIFF (`.tif`, `.tfw`), CAD data (`.dwg`), e Tabular GIS attribute data.
- **Qualitative textual data**: eXtensible Mark-up Language (XML), Rich Text Format (`.rtf`), Plain text data, ASCII (`.txt`).
- **Images, audio and video**: TIFF (`.tif`, `.tiff`), JPEG (`.jpeg`, `.jpg`), Adobe Portable Document Format

(PDF/A, PDF) (`.pdf`), PNG (`.png`), Free Lossless Audio Codec (FLAC) (`.flac`), MPEG-1 Audio Layer 3 (`.mp3`), Audio Interchange File Format (`.aif`), Waveform Audio Format (`.wav`), MPEG-4 (`.mp4`), MOV (`.mov`), Windows Media Video (WMV) (`.wmv`).

In the field!

I am a researcher working with data organised in tables. *What are the most used formats?*

The most used formats for tabular data are:

- 'Comma-Separated Values' (CSV, `.csv`): a non-proprietary, textual format in which data is usually separated by commas.
- 'OpenDocument Spreadsheet' (ODS, `.ods`): an open standard format for spreadsheets, which stores data in cells arranged into rows and columns. Also, .ods files can be opened in Microsoft Excel and saved as XLS or XLSX files.
- 'Excel Workbook' (XLS/XLSX, `.xls`/`.xlsx`): the Excel format is a proprietary yet very common format that allows users to create, handle and analyse tabular data in a spreadsheet.

For my research, I need to collect data through surveys. *What tool can I use?*

Surveys can be conducted through interviews or questionnaires in person, over the phone or online.

Depending on the population you want to sample, the size of the sample itself and the sample design – which can be simple or complex, longitudinal or cross-sectional – it may be necessary to accompany these techniques and tools with the support services offered for data management, privacy and/or ethics.

Examples of online survey tools include Microsoft Forms, Google Forms, LimeSurvey, SurveyMonkey, Qualtrics. If you collect personal data, you must use a tool such as Microsoft Form, supplied by the University, or check any licences available through your Department (LimeSurvey, SurveyMonkey, Qualtrics), rather than using personal licences.

In the case of a cross-sectional survey for which you won't need to contact the same person twice (or several times), you can choose to implement privacy-by-design techniques to anonymise data at the source, thus avoiding privacy issues.

For my research, I work with biomedical imaging data. *How do I choose the format for saving and archiving it?*

Digital Imaging and Communications in Medicine (DICOM) is the standard for transmission and management of medical images and related information. In addition to the image, a DICOM file includes a heading that contains all the metadata acquired with the image itself (patient data, tumour location, duration and amount of radiation, etc.).

TIFF is another appropriate format to store and share medical images – it is a raster graphic file format that supports lossless compression and, as such, is suitable for archiving and printing high-resolution images and photos. All the relevant metadata can be saved in a separate TXT file.

On data collection and methodologies

In research practice, data collected or generated by third parties can be reused instead of, or in addition to, **generating new data**.

Provided the data is of good quality, **reusing existing data** saves time and resources. **Online digital archives for long-term data preservation**, sometimes specific to a certain disciplinary domain, can be accessed to browse and download relevant data **Repositories**.

Before reusing data, regardless of their origin, you need to make sure you are legally and contractually able to do so **Copyright** **Respecting privacy**.

Generare o raccogliere i dati può comportare pratiche molto diverse tra loro. Ad esempio, i dati possono essere di natura **sperimentale**, quando ottenuti tramite esperimenti e dimostrazioni che seguono un metodo scientifico. Oppure possono essere di natura **osservativa**, quando vengono raccolti attraverso l'osservazione critica, con l'eventuale aiuto di strumenti. Quando la ricerca è **compilativa**, i dati vengono raccolti in forma derivata/compilata da altre fonti.

Indipendentemente dalle pratiche di generazione o raccolta dati, gli **strumenti**, **software e metodi utilizzati devono essere registrati per consentire la riproducibilità della ricerca** **Gestire il software**.

Inoltre, sempre a prescindere dai metodi di raccolta o generazione dei dati, è necessario assicurarsi di essere conformi alle normative sulla privacy e sull'etica.

Se hai intenzione di sfruttare commercialmente i tuoi dati, perché possono essere utili, per esempio, per depositare una domanda di brevetto, pianifica in anticipo delle strategie di gestione dei dati che possano garantirti adeguata protezione.

In the field!

I have developed a software for analysing and displaying the results of my research.
Do I have to manage it in the same way as research data?

Yes, it is advisable that you plan software development and use tools to document it, as this allows you to exploit the software itself as an asset and the main output of your research, as well as facilitating reuse in future research.

Some tools, such as cloud notebooks, can help you document code development and every step of its algorithm. By running your code in cloud, you can view how every single part is run and the corresponding input and output data.

Once your code reaches a stable executable version, it is recommended that your deposit it in a disciplinary repository together with suitable documentation and specific metadata, to ensure that it is preserved in the long term. An example of disciplinary repository for source code is Software Heritage, which uses CodeMeta as metadata schema, and regularly and automatically harvests the most common forges for development, such as GitHub. **Managing software** **Repositories**.

I work with cultural heritage and my data is mostly text and images, often from archives, museums and libraries.
What do I do?
Contact the institution that holds the sources you wish to use in your work to find out what you need to do. Even though they are no longer copyrighted, they may still be protected as cultural heritage, and you may need permission to reproduce them. If the sources you work with are still in copyright, you will need permission from the rights holders **Copyright**.

Useful links

More on formats: https://www.loc.gov/preservation/resources/rfs/TOC.html | https://www.dicomstandard.org/

https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

Useful tools for interviews:

https://forms.office.com | https://www.qualtrics.com/it/ | https://www.limesurvey.org/it

Useful tools for software:

https://datasciencenotebook.org/ | https://www.softwareheritage.org/ | https://codemeta.github.io/codemeta-generator/ | https://github.com/

Datasets: selecting and organising data

A dataset is a **set of data organised in an orderly manner** and structured according to specific criteria. It also needs to include the metadata that describe, locate and relate the data to each other.

Data, datasets and metadata: an everyday example

DATA: a random group of photographs

DATASET: a set of photographs collected for a common purpose

METADATA: information about each photograph and the album as a whole

→ the photos can be found and understood in context

Image credits (respectively): Photo by Roman Kraft for Unsplash; Photo by Laura Fuhrman for Unsplash; Photo by Frelo Design for Unsplash

Datasets play a key role in replicating the analysis carried out by researchers. Proper dataset management and organisation are crucial to ensure the reliability and usefulness of data, allowing others to explore a research topic and deepen their understanding of it.

A well-organised dataset needs to meet specific criteria that guarantee its quality and usability. FAIR principles (Findable, Accessible, Interoperable, Reusable) provide essential guidance in assessing how good a dataset is  **FAIR principles**.

Every dataset should allow other researchers to replicate the analysis and validate results. The research process and methodologies adopted should be **transparent**. An **accessible** dataset is easy to share and, when it is associated with clear conditions for access, fosters collaboration among researchers. Finally, when a dataset is **reusable**, it is available for new research and allows reducing costs and duplications.

Selecting data to be included in a dataset

Depositing all data associated with a research project would often be unsustainable – you need to consider what data is useful for the understanding, verification and reproducibility of research. This decision is ultimately the responsibility of researchers and can be domain specific.

Examples of data that should always be deposited include:

- Original dataset and/or software code.
- Raw data obtained from the analysis of physical samples.
- Observational data that cannot be reproduced.
- Non-original datasets that are not readily available (provided you have permission).

Organising datasets

A single study can produce many different types of data, all contributing to answering the same research question. Structuring them into a dataset, and relating them to each other, can help **clarify the process** that led to the result.

The organisation of data in datasets, if **planned at the start** of research, simplifies data management throughout its lifecycle and is a fundamental investment in the success of a research project because it improves:

- **Efficiency.** It makes it easier to search for and access the data when needed, avoiding time loss and frustration. It also prevents file duplication, saves storage space, simplifies management and allows team members to collaborate more easily and to keep track of the changes made to the data.

- **Reproducibility.** It makes research more transparent and reproducible, allowing other researchers to understand methodologies and results. Giving access to clear and documented raw data and metadata facilitates verification and validation of results by other researchers.
- **Reliability.** It prevents data loss, corruption or unauthorised access. Implementing appropriate security measures protects sensitive data from intrusions and breaches, and data organisation facilitates compliance with domain-specific standards and regulations.

In the field!

I am a researcher, and I need to understand how to structure my data in a dataset. *Where do I start?*

You can start by looking at the data you intend to work with, its characteristics and how the different categories (if any) relate to each other.

Organise them in folders with clear and effective names.

Carefully document your datasets by providing all necessary information and metadata.

Identify and implement storage solutions to protect datasets during the active phases of your research.

Think in advance of possible criticalities and choose a repository that suits the needs of your project, by adopting a long-term perspective aimed at final preservation.

I am a researcher in social sciences, and I want to create a dataset to analyse socio-demographic variables.
– Example

You want to study wealth in Italy, e.g. income in every region, to produce an article in which you show your data graphically using maps. In the planning phase, you decide to structure your dataset as follows:

- The input files, containing the raw data. This data includes information on population and geographical units.
- The code you used to analyse the data. The script loads the data, cleans the data, performs the analysis and generates the output files (e.g. using R, which is an open-source software).
- The output files, containing the results. They include tabular data with the values of the variables derived from the output data, and images, i.e. the maps in your article.

All these data constitute your dataset, which makes your research more understandable and reproducible.

Metadata and documentation

Metadata is any information about data, i.e. **descriptors** that make it easier to classify and find them. Documenting all information necessary to make research data intelligible to others (i.e. anyone not involved in its collection, anyone who wants to reuse it, or even your future self) is crucial to proper research data management. Metadata is a special form of documentation: by presenting information in a **structured** way, they describe, explain and locate data, and facilitate their use. Metadata are **machine-readable** and, as such, make data findable by search engines, allow software/services to access, understand, transform, export and/or import the data to/from other software.

Types of metadata

There are different types of metadata, including:

- **Structural metadata** provide information about how the object (e.g. file, dataset, etc.) is organised. For example, structural metadata may describe the layout of a table or the relationships between its elements.
- **Descriptive metadata** provide information about the content of an object (e.g. file, dataset, etc.) and describe it in an essential manner. For example, the place and the time a photograph was taken are descriptive metadata.
- **Data quality metadata** collect specific elements to define indicators of the quality of an object. It may include quantitative metrics, such as the variance or standard error of a variable, or more qualitative aspects (e.g. a discrete classification).
- **Administrative metadata** include elements such as object identifiers (e.g. the Persistent Identifier associated with a dataset). Administrative metadata also include:
  - **Provenance metadata**, documenting information about the authors and processes that produced the object. This category includes the names and affiliations of the authors.
  - **Legal metadata**, documenting the conditions for (re)using the data. This category includes information about licences.

TITLE: Panorama of Bologna

LOCATION: Bologna

DATE: 01/01/2024

EXPOSURE: 1/30s

Image credits: Photo by Bianca Ackermann for Unsplash

Standard metadata schemas

Metadata are usually structured according to schemas, meaning their elements are organised based on **standardised conventions**. Using a standard offers many advantages in terms of interoperability, making it possible to **assign a unique, unambiguous meaning to each element**, gather essential information about the data, and compare heterogeneous data from several sources, domains and disciplines.

Standard metadata schemas can be either **generic or disciplinary**.

Generic metadata schemas such as Dublin Core collect a minimum set of high-level standard information, are easy to use and widely adopted. However, when more specific information is required, they often need to be expanded.

Domain-specific schemas have a much richer vocabulary and structure, tend to be highly specialised, and tend to be fully understandable only by researchers in that domain.

A further level of standardisation can involve the **vocabularies** employed to fill in the values of the metadata elements **Using different types of standards**.

In the field!

I am a researcher, and I am looking for a disciplinary metadata schema to describe my data. *Where do I start?*

Regardless of your research domain, there are several online portals you can use to look for the metadata schemas that best suit your needs:

1) FAIRsharing, a manually curated resource that gathers data and metadata standards;
2) The Research Data Alliance (RDA) Metadata Standards Directory;
3) The Digital Curation Centre (DCC) List of Metadata Standards, which gathers disciplinary metadata standards.

I am a biomedical researcher, and I am unable to find a metadata schema suitable for my data. *What can I do?*

If you are having trouble finding a specific metadata schema, you should look for the 'minimum information about your topic' (MIAME, MIAPE or MIAPPE) recommended for your specific data type in your specific discipline.

Metadata and repositories

The **repository you choose** to deposit your datasets influences the choice of metadata schema. At the time of deposit, a set of contextual information is required; this is organised according to a standard schema, and it effectively constitutes the metadata of the object deposited **Repositories**.

So, knowing from the early phases of your project what repository you will use gives you a clear idea of what metadata fields you will need to fill in and guides you in adding the right metadata to link your data to other software and systems.

In the field!

I have chosen the repository where I will deposit my data. *How do I find out which metadata schema it adopts?*

This information can be easily found in the repository documentation pages or via the re3data registry. AMS Acta, the University repository, uses the Dublin Core and DataCite schemas, just like Zenodo.

The repository I have chosen to deposit my disciplinary data uses a generic metadata standard.
How can I add more information to my data?

Even though a repository uses a generic schema, you can still add domain-specific metadata to your documentation. If you need to find a suitable schema, see the first point in this section.

Metadata and documentation

While metadata is a special type of documentation, structured and designed so that machines can read and interpret it, other kinds of documentation take the form of **human-readable text documents**, such as README files or codebooks. They gather additional information needed to understand and interpret the data they refer to. As a rule, while a README file contains high-level information, such as the scope and context of research, funding, methodologies. A codebook is designed to document, for example, the meaning of the names of the variables, or the units of measurement, if any.

A README file designed to accompany a dataset must include:

- General information about the project, such as title and goals of the project and dataset, names/roles/contact details of the researchers involved, funding information, Persistent Identifier (PID), etc.
- Folder and file structure and naming system, file names and folder structure, relationships and dependencies between files, description of the content of each main file, etc.
- Information about methods and software used for data collection (including references, documentation, links, experimental conditions, instrument standards and calibration, etc.), methods and software used for data processing, file formats and description of the version control system, quality control procedures applied, etc.

- Information about reuse and links to other materials, i.e. information about licences and restrictions on (parts of) the dataset, links to publications based on the dataset, relationships with other datasets and other resources used as a source for data collection (books, articles, etc.).

The codebook, which may also form part of the README file, is usually designed to document the meaning of the names of the variables and of the units of measurement, if any. At data collection level, it is usually a tabular file that contains:

- Definition of codes, symbols and abbreviations used in the files.
- List of variables with their full name and definition.
- Definition of column headings and row labels for tabular data.
- Units of measurement and data formats (e.g. YYYYMMDD).
- Processing of missing data (code, etc.).

Il file contenente la documentazione che descrive il dataset deve essere archiviato insieme ai dati nel momento in cui questi vengono depositati in un repository. È opportuno che il file di documentazione sia salvato in un formato aperto e accessibile (es .rtf).

Useful links

FAIR and the Notion of Metadata https://faircookbook.elixir-europe.org/content/recipes/introduction/metadata-fair.html

The Research Data Management Toolkit, Documentation and Metadata https://rdmkit.elixir-europe.org/metadata_management

The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-metadata

FAIRify Your Data: Data Documentation and Metadata, Flora D'Anna https://osf.io/wbr7t

Research Data Management: Metadata (University College Dublin Library) https://libguides.ucd.ie/data/metadata

Dublin Core Metadata Standard https://www.dublincore.org/specifications/dublin-core/dces/

Resources to search for metadata schemas:

- FAIRsharing Standard Registry https://fairsharing.org/search?fairsharingRegistry=Standard
- RDA Metadata Standards Directory https://rd-alliance.github.io/metadata-directory/standards/
- DCC Guidance on Disciplinary Metadata https://www.dcc.ac.uk/guidance/standards/metadata

Minimum Information Standards:

- MIAME https://www.fged.org/projects/miame
- MIAPE https://www.psidev.info/miape
- MIAPPE https://www.miappe.org/

FAIR principles

The main goal of FAIR principles is to provide the scientific community with a framework in which data, methodologies and tools are used for validating (and, in certain cases, for replicating) the conclusions of a research project.

The first step is **finding** the data. Then, a user needs to know how to **access** them (either through a free download or through an authentication and authorisation mechanism). The data must be in the condition to be **integrated** with other data and **interoperate** with different applications or workflows.

The goal of FAIR principles is to **optimise the reuse** of data. They describe how research results should be organised to make it easier for anyone to **access, understand, exchange and reuse** the data. Not all FAIR data is open data.

Data and FAIR principles

We have seen that repositories are archives that preserve data and make it findable online. The data deposited by a researcher in a repository is **more likely to abide by FAIR principles**. In fact, in order to be findable, data and datasets must be accompanied by a Persistent Identifier (PID), i.e. a long-term, unique and unambiguous reference. Examples include DOIs, Handles and URNs. Data and datasets must also be accompanied by meaningful metadata and keywords (containing the PID).

Accessible data is not necessarily open data; rather, it is **data associated with clear access conditions**. Data may be openly accessible (usually the default) or accessible through an authentication and authorisation system, when their nature prevents them from being open. Privacy and intellectual

property protection are only two of the reasons for restricting data access. Remember that the metadata associated with closed data should be kept openly accessible via standard protocols.

In order to be interoperable, data need to be combinable and usable with other data or tools. This is achieved by using **common and preferably open formats** as well as **standardised languages** that are shared internationally by the various indexing services. Repositories support **interoperability standards** and may recommend using **specific ontologies or vocabularies** **Repositories** **Using different types of standards**.

To guarantee maximum reusability, data must also be **described and documented** thoroughly, to ensure compliance with the standards adopted by the relevant scientific communities and allow reuse and combination across different contexts.

Last but not least, accompanying a dataset with a **clear, accessible, possibly open licence** is key to establish and declare how the dataset can be reused **Copyright**.

Metadata and FAIR principles

FAIR principles apply to both data and metadata.

Rich, descriptive metadata make a significant impact on data utility by improving findability, accessibility, interoperability and reusability. To comply with FAIR principles, metadata must be always accessible and associated with a permissive licence (CC0 or equivalent), even when the data is not **Respecting privacy**.

Sul campo!

How do I know if my data is FAIR?

To easily self-assess compliance with FAIR principles in research data management, please find below a checklist prepared by EUDAT and reworded by AlmaDL

Findable

- ☐ Has a persistent identifier (e.g. DOI, Handle, URN) been assigned to your dataset?
- ☐ Are there rich metadata, describing your dataset?
- ☐ Are your metadata recorded in an online searchable resource e.g. a catalogue or data repository
- ☐ Does your metadata record specify the persistent identifier?

Accessible

- ☐ Does the persistent identifier take you directly to the dataset or associated metadata?
- ☐ Does the protocol by which data can be retrieved follow recognised standards?
- ☐ Are metadata public, wherever possible, even if the data are not?

Interoperable

- ☐ Have your data been provided in commonly understood and preferably open formats?
- ☐ Do your metadata follow relevant standards?
- ☐ Are controlled vocabularies, keywords, thesauri or ontologies used where possible?
- ☐ Are qualified references and links provided to other related data, such as publications, technical reports or software applications?

Reusable

- ☐ Are your data accurate and well described with many relevant attributes?
- ☐ Has your dataset been assigned a clear and accessible data usage licence?
- ☐ Have the scientific responsibility for and purpose of the data been made clear in the metadata and attached documentation?
- ☐ Do your data and metadata meet relevant domain standards?

Useful links

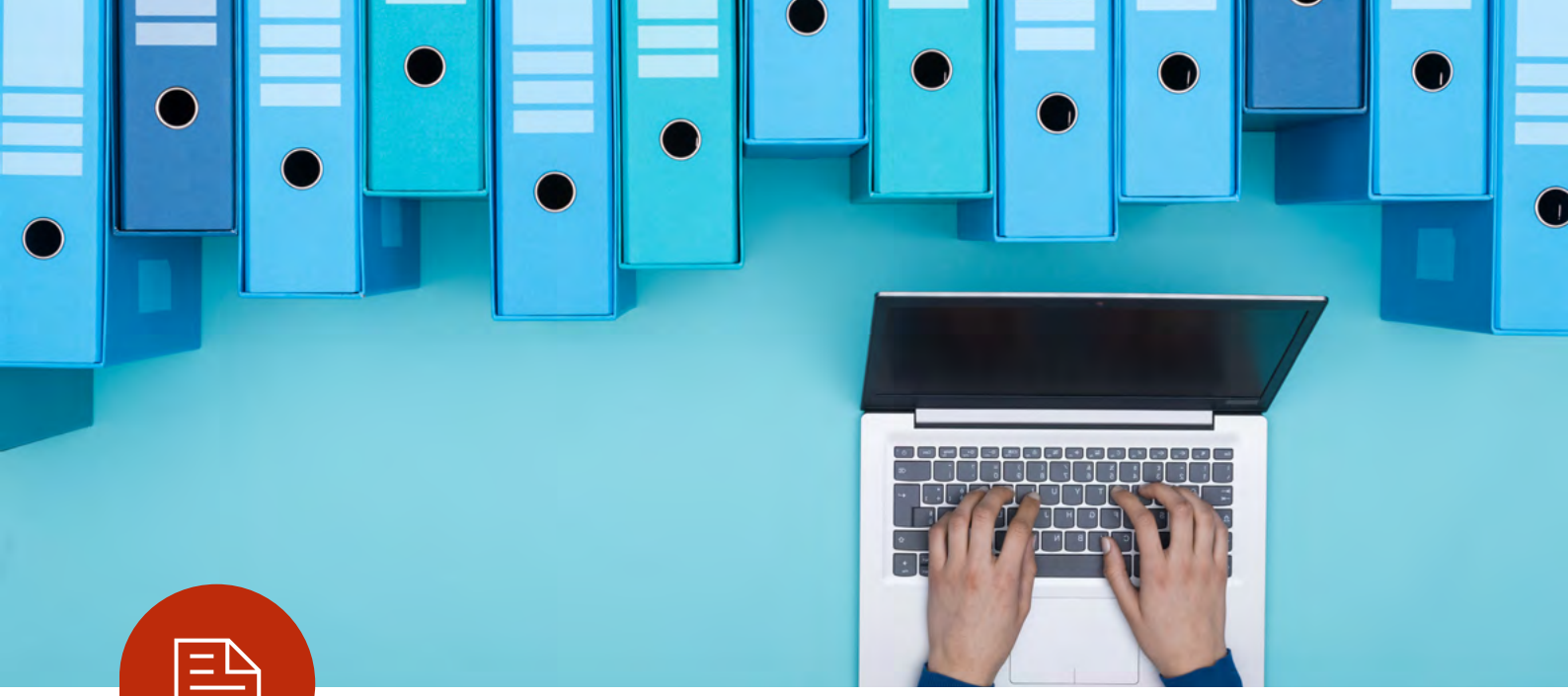
Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Video “UGent Open Science. Knowledge clip: FAIR data principles”. https://www.youtube.com/watch?v=2uZxFu9SFi8

More on the FAIR principles:

- GOFAIR. FAIR Principles https://www.go-fair.org/fair-principles
- FAIRsFAIR Fostering Fair Data Practices in Europe https://www.fairsfair.eu/
- How to FAIR https://howtofair.dk/

Video “Dati della ricerca: la European Open Science Cloud e i principi FAIR” (in Italian) https://www.youtube.com/watch?v=eNiHNaU6MrQ



Repositories for data deposit

Repositories are **digital infrastructures for archiving and preserving data, publications, software** and, more generally, research results **in the long term**.

The data archived in a repository is **associated with metadata structured** according to standard schemas **Metadata and documentation**. Repositories also play a key role in promoting open and accessible research, allowing researchers to **abide by FAIR principles** as they make it possible to associate a set of descriptive metadata, a **Persistent Identifier (PID)** and a **licence** with the data deposited **FAIR principles; Copyright**. Repositories further enable researchers to **choose an access level** for their data and allow users to download deposited data they wish to reuse.

Data volume and repositories

The choice of a repository must consider not only the type of data managed, but also its volume. The need to pay for depositing data is specified in the **terms and conditions** of each repository, which you should always check. In most cases, data can be deposited free of charge, but some repositories may require the payment of a fee to deposit large datasets (usually over tens of GB).

Types of repositories

Repositories can be **disciplinary, institutional or generalist** (multi-purpose) in nature. Each type has its specific advantages, so it is important that you choose based on your needs.

- **Disciplinary** repositories use metadata schemas specifically designed for a certain discipline, can offer greater visibility and make it easier to share data within the relevant scientific community.
- **Institutional** repositories are made available to the members of an academic or research institution and usually provide validation and support services to ensure the quality of the datasets deposited. AMS Acta and AMS Historica are the repositories that the University of Bologna provides to its researchers (see below).
- **Generalist** repositories gather data and materials from different disciplines and research contexts. They provide a solid platform for data preservation, visibility and accessibility. Zenodo is one of the main generalist repositories in use.

In the field!

I am a researcher, and I need to choose a repository to deposit my datasets. *Where do I start?*

Carefully review any ethical, privacy and intellectual property constraint of the datasets you intend to produce to establish long-term preservation strategy and dataset access level.

Estimate the volume of your data to find out whether you can deposit it for free.

Decide if you need support or are happy to deposit your data by yourself.

Check the repository documentation to make sure that your selected repository meets the requirements to ensure that the data you deposit is FAIR (especially: PID, metadata, licence).

Repositories and accessibility levels

The choice of the access level is key: you should choose a repository that allows you to deposit your data in accordance with Open Science practices but also restricts access where necessary.

The access level can be:

- **Open**, when the dataset is openly accessible to anyone who wishes to browse, download and reuse it.
- **Restricted**, when those who wish to browse or download the deposited dataset must ask for permission. This may be obtained directly from the researcher who deposited the dataset, or, for certain specific repositories, from a committee in charge of assessing the legitimacy of the request for access.

- **Embargoed**, a temporary restriction that keeps the deposited datasets private for a limited time, after which the embargo expires, and the dataset becomes open.

In the field!

I am a researcher in social sciences, and I want to deposit a dataset I produced, containing a survey. *What can I do?*

Use re3data to search for a repository that makes your data visible to your scientific community.

Consider whether the survey you intend to conduct is cross-sectional or longitudinal, as they pose different challenges.

Longitudinal surveys usually include personal data, as the same individuals are interviewed more than once over time. In this case, you will have to choose a repository that also allows you to restrict access to the deposited data.

I am a researcher working with sensitive data, which cannot be deposited in open access. *What can I do?*

As regards the management of sensitive data that cannot be made anonymous, it is crucial that you choose a restricted-access repository. This will make your data available to authorised users only, while publishing metadata and support documentation.

Another option is to deposit the methodology for data analysis in open access, but keep the data private where necessary, so that at least the research process can be replicated.

My data could be involved in a patent procedure. *Can I deposit it?*

Remember that, if your data is necessary for a patent, it should be kept confidential until the procedure is completed, in order not to jeopardise novelty. Do not share it with anyone and do not upload it online without restrictions. Deposit your data in a repository that allows you to place a temporary embargo on it.

University repositories: AMS Acta and AMS Historica

AMS Acta is an institutional repository for the collection, preservation and dissemination of the University of Bologna's research data. It allows professors, researchers, research fellows, and students of the University of Bologna to archive research data in compliance with FAIR principles and Open Science:

- It ensures preservation and access to the outputs deposited over time.
- It assigns a DOI (Digital Object Identifier).
- It implements various access levels (open, closed, embargoed).
- It implements the Dublin Core and DataCite descriptive metadata standards; metadata are always accessible under a Creative Commons Zero (CC0 1.0 Universal) licence.
- It implements various licences for the data, including Creative Commons licences.
- It complies with international standards for interoperability and metadata

transmission, is registered in the re3data catalogue, and is indexed by the main catalogues (OpenAIRE, BASE, WorldCat) and search engines (Google, Google Scholar, etc.).

- It provides statistics about accesses and downloads for each output.

AMS Historica is an institutional repository that gathers the digital reproductions of the University of Bologna's ancient and valuable sources of scientific and cultural importance. It is built for browsing the digital reproductions of rare and unique documents, including works of art, monuments, archaeological finds, manuscripts, papyri, books, journals, newspapers, maps, drawings, photographs, audio and video sources of scientific, historical and cultural importance, which are kept in the museums, libraries and archives of the University or which are the result of national and international research projects.

The content is published in accordance with the national and international guidelines and standards that foster preservation and enhancement of digital collections over time:

- It associates metadata and licences that enable discovery, study, sharing and reuse according to the principles of Open Science.
- It operates on an open-source platform, DSpace-GLAM, that offers new and powerful features for browsing and studying heterogeneous digital content thanks to its digital services based on IIIF (International Image Interoperability Framework).
- It is indexed by national and international catalogues and aggregator services such as Cultura Italia, Europeana, OpenAIRE, WorldCat and BASE.

Useful links

Registries of repositories:

- Re3data https://www.re3data.org/
- OpenAIRE Explore https://www.openaire.eu/find-trustworthy-data-repository
- FAIRsharing Repository Database https://fairsharing.org/search?fairsharingRegistry=Database

University repositories:

AMS Acta https://amsacta.unibo.it/ | AMS Historica https://historica.unibo.it/

Further information about the University repositories:

- "Preserving and disseminating research data in AMS Acta" (https://sba.unibo.it/en/almadl/almadl-services/preserving-and-disseminating-research-data-in-ams-acta);
- "Preservation and enhancement of the digital cultural heritage" (https://sba.unibo.it/en/almadl/almadl-services/preservation-and-enhancement-of-the-digital-cultural-heritage).

Using different types of standards

The good practice of using standards (e.g., formats, vocabularies, processes) during research can take many forms, but always leads to consistent and interoperable data. You can use **standards to codify the methodologies** for generating or reusing data, or to **describe data itself**.

You can reuse existing standards or – more sporadically and only if strictly necessary – create new ones, specific to your research.

To make sure that all researchers involved in the project follow a standard **methodology**, for example, you can use the **ISO standards** specific to your disciplinary domain.

Standards in the form of vocabularies, taxonomies and ontologies can be used to univocally structure data and make it interoperable with both existing and future data. Using at least one vocabulary when collecting information can help you to correctly identify data and facilitate their reuse, by making them understandable to others who share the same vocabulary.

'**Vocabularies**' or 'thesauri' are lexical resources comprised of controlled terms used to describe a set of items, knowledge, data, theories from a certain scientific field or disciplinary domain. They provide standard terminology, improving the value of data and making them machine-readable.

Items in '**taxonomies**' are not only described by controlled vocabularies of selected terms, but also ordered in a system, usually a hierarchical one.

In '**ontologies**', the types and therefore the names of the relationships between the items that make up a taxonomy or a vocabulary are made explicit.

On the other hand, specific metadata standards should be used to describe data **Metadata and documentation**.

In the field!

I work on a collaborative project that also involves companies and research centres outside of the University of Bologna. *How do I make sure that my results are of high quality and interoperable?*

To achieve this goal, common standards need to be shared within the collaborative project.

ISO standards are rules that describe the best way of doing something, as agreed by international experts in the field. Standards cover a wide range of activities – from making a product to managing a process, to providing a service or supplying materials.

For example, quality management standards are designed to help work more efficiently and minimise product defects. Health and safety standards can reduce accidents in the workplace. IT security standards contribute to protecting sensitive information.

I want to use vocabularies and taxonomies to describe the information I gather in my data. *What benefits does this bring and where do I find the most suitable ones?*

A vocabulary can contain standard terms from a specific disciplinary domain, which can be used to identify the variables in an analysis, to then include them in a tabular file as column headings, for example.

In research with qualitative data, using terms from a standard vocabulary can help express relationships between variables within a text such as a technical report, so that all those involved in the project know exactly what event or type of data is being referred to.

There are both generic schemas, such as the Information Artifact Ontology (that describes information entities, understood as pieces of information encoded in digital or physical entities), as well as extremely specific ones, such as the IUPAC Compendium of Chemical Terminology (which contains some 7,000 chemical concepts derived from the IUPAC Recommendations).

You can use a registry, like FAIRsharing (see box below), to look for schemas, vocabularies and thesauri that suit your needs.

Useful links

ISO Standards
https://www.iso.org/standards.html

Search tool for metadata schemas, vocabularies and thesauri
https://fairsharing.org/search?fairsharingRegistry=Standard

Managing software

Software is developed across a variety of disciplinary domains and should always be treated with care. Unlike data, software is executable, it continuously changes its form over time and hence requires specific steps to be managed properly.

Let us identify the steps and tools relevant to each phase of code development.

In the **planning** phase, roles and responsibilities for designing, developing and maintaining the code should be defined. From architecture and concept design, to development, up to revision and debugging.

During **development** it is highly recommended to use online platforms such as GitHub and GitLab. They are built to monitor development through version control and ensure effective collaboration thanks to Git's distributed system.

When your software reaches a satisfactory working version, it is also a good idea to **deposit** it in Zenodo or in another repository specific to this type of digital object, such as Software Heritage. They both allow the software to be semi-automatically updated in case of further development.

Deposit both **machine-readable and human-readable documentation** with your code **Metadata and documentation**. On the one hand, this could simply consist of metadata and a citation file – two examples specifically designed for software are CodeMeta and CITATION.cff. On the other hand, human-readable documentation usually includes a README file and inline documentation, i.e. comments within the code. Both make your code easier to understand for anyone else (or your future self!).

Once deposited, the software requires a licence to clarify which uses are permitted. The choice of a certain licence over another depends on your

specific situation – whether you need to work in a community, you want a simple and permissive licence, or you wish to share improvements.

After deposit in a repository such as Software Heritage, the code is assigned a Persistent Identifier (PID), a licence, and disciplinary metadata describing it. This means that it can be cited as a research output, in the same way as a publication.

Sometimes, however, archiving your software (even with the related input and output data) is not sufficient to ensure its **reusability** by contemporaries, and especially by the developers of the future. Programming languages, libraries and plug-ins change version often, and even code developed a few months prior may no longer be up to date. Special solutions, such as containers or cloud-based systems, can simulate the system environment variables in which the code was developed and to run it again.

In the field!

The ultimate research goal of my competitive project is to develop a data analysis software.
What workflow should I follow?

Arrange a kick-off meeting with the entire development team to identify roles and responsibilities, as well as setting a work schedule and deadlines.

Development can start directly in GitHub – creating a new repository shared by all developers facilitates collaborative work by keeping track of changes to the code, input data and other support material.

Some cloud-based tools, such as electronic notebooks, can help document every step of the code algorithm, keeping track of and visualising the input and output data of each function.

Linking the repository to Software Heritage before your code reaches the first operating version enables regular and automatic harvesting and ensures that the code is properly preserved (a feature not guaranteed by GitHub!) with an appropriate metadata schema, i.e. CodeMeta.

Various interactive online tools – such as Choose a Licence – summarise the entire documentation of the main licences in use for software, allowing you to choose a licence that suits your project requirements, which you can include in the GitHub repository.

To ensure proper recognition of your work, CITATION.cff files can also be generated semi-automatically, which saves a lot of time, and then uploaded directly to the GitHub repository and thus automatically saved to Software Heritage.

Useful links

Licence chooser tool https://choosealicense.com/

Citation file generator https://citation-file-format.github.io/cff-initializer-javascript

Examples of cloud notebooks for code development https://datasciencenotebook.org/

- Software Heritage https://www.softwareheritage.org/
- CodeMeta https://codemeta.github.io/codemeta-generator/
- GitHub https://github.com/

Copyright and research data

Research data may be protected by copyright when they are creative intellectual works. These may be texts, creative images, processed tables, databases, software **Managing software**.

As a matter of fact, copyright protects **creative intellectual works** in the fields of literature, music, art, architecture, drama, cinema, science, regardless of the mode or form of expression. However, copyright protects the tangible form of an intellectual work; this means that copyright does not protect ideas, procedures, methods of operation or mathematical concepts as such.

Applicable regulations

Italian Law no. 633 of 22 April 1941 "Protezione del diritto d'autore e di altri diritti connessi al suo esercizio (Protection of copyright and related rights)".

In the field!

I am a researcher, and I need to use copyrighted works. *How do I know if copyright protection has expired?*

Check if the work is in the public domain, i.e. the term of copyright protection has expired, by counting 70 years from the death of the author or 70 years from the first publication of a collective work.

If this is the case, make sure that no cultural heritage protection restrictions apply.

I am a researcher, and I need to use copyrighted works in my research. *How can I avoid infringing copyright?*

First, check if these works come with a reuse licence and make sure that the uses you intend to make are in line with the terms of the licence. For example, free use is permitted under a Creative Common Attribution (CC BY) licence.

If this is not the case, consider if you can only use fragments or parts of these works in your research. If so, make sure that the use you intend to make complies with Article 70 of the Italian Copyright Law, under which the abridgment, quotation or reproduction of fragments or parts of a work and their communication to the public for the purpose of criticism or discussion is permitted within the limits justified for such purposes, provided such acts do not conflict with the commercial exploitation of the work. If it is for teaching or research, this use must have the sole purpose of illustration, and be non-commercial.

In all other cases, you will need to obtain permission from the rights holder (often the publisher) to use the work.

Authors' rights

Authors' rights include **so-called moral rights**, such as the right to claim authorship of a work, the right to publish unpublished works and the right to the integrity of a work. These rights may not be waived or transferred and are not subject to the statute of limitations. They may be claimed at any time after the author's death. Authors' rights also include **so-called property rights**, which entitle authors to commercially exploit their works on an exclusive basis. Publishing, digitisation, communication (also online), modification and translation rights, among others, are property rights that may be exercised exclusively by the author. These rights may be transferred for a fee or free of charge, on an exclusive or non-exclusive basis, and may be exercised up to 70 years after the author's death.

Databases: between copyright and related rights of database creators

Databases are defined as "*collections of works, data or other independent materials which are systematically or methodically arranged and can be individually accessible electronically or by other means*". Databases are protected by copyright when the choice or arrangement of materials are regarded as an intellectual creation of their authors. The copyright protection for databases does not extend to their contents and is without prejudice to any third-party rights on said contents.

The investments made to create, verify or present a database, requiring financial means, time or labour, are protected independently from authorship. A related right (**so-called *sui generis* right**) is thus recognised to the "database creator" (Italian: "costitutore"), preventing the extraction or reuse of the whole or of a substantial part of the database.

The term of the *sui generis* right is shorter than that of authors' rights and applies for 15 years from the first of January of the year following the date of completion of the database or the date on which it is first made available to the public.

The conditions for using a database are governed by the rights holders by means of specific licences for use. Remember to check the terms of use.

Licensing intellectual property

In order to use works and materials protected by copyright and related rights, you need the **prior consent of the rights holders**.

Property rights cover the work as a whole and each of its parts; this means that the author's exclusive right also extends to any partial uses and must be authorised.

Licensing agreements allow authors to transfer their property rights and third parties to use their work, under the agreed conditions. Authors retain ownership of the rights, which are returned to them when the licence expires.

Property rights are independent of each other, i.e. each right can be transferred separately from the others. The transfer of rights must be made in writing.

A work may only be used without permission of the rights holders in the exceptional and limited cases expressly provided for by the law.

Remember to check the terms of the licence associated with a work.

Creative Commons licences

Creative Commons (CC) licences are the most popular licences for digital works. They are licensing agreements under which the author grants permission to use a work to an indefinite number of people, under certain conditions, by deciding which rights to retain and which to license for use.

There are six licensing schemes available, based on four basic clauses that authors can select and combine, thus expressing the ways in which end users will be able to use their work.

Each basic clause has a graphic symbol that makes it easier to recognise:


BY – Attribution: always included


NC –Non-commercial


SA – Share Alike


ND – No derivative works

CC licences are available in three forms:

- Commons Deed (the user-friendly symbols summarising the terms of the licences);
- Legal Code (the actual, full licensing agreement);
- CC REL – Creative Commons Rights Expression Language (the set of machine-readable information).

The Creative Commons licences and often associated with datasets, in line with the principles of Open Science, are:

- CC BY, "Attribution": enables to freely reuse and modify the work, so long as attribution is given.
- CC BY-SA, "Attribution, ShareAlike": enables to freely reuse and modify the work, so long as attribution is given, and the modified work is licensed under the same terms as the original work.
- CC0, "No Rights Reserved": enables authors to put their work into the public domain and/or give up their rights over it.

Image credits:
JoKalliauer; foter, CC BY-SA 3.0
https://foter.com/blog/how-to-attribute-creative-commons-photos/
via Wikimedia Commons

In the field!

I am a researcher, and I would like to use materials I found online in my research. *How do I know if I can do that?*

Remember that the Internet is subject to rules and restrictions, too. Always check the terms of use of each website and of the materials published in it. As a rule of thumb, the fact that a work is available online free of charge does not mean that it can be used freely without permission.

When extracting data from an online database or reading a scientific article in an e-journal, remember to comply with the terms of the associated licence.

If no Creative Commons licences are associated with a piece of content, look for the "Terms of use" page (or similar) on the website. If the terms of use are not expressly stated, it means that all rights are reserved, and you must obtain permission from the rights holder.

I am a researcher, and I would like to use images I downloaded from a digital library in my research.
How do I know if I can do that?

Check the associated licence for use; Creative Commons licences are often used.

I am a researcher, and I want to associate a CC0 licence with the data I produced. *What does it mean in practice?*

Check that the licence can be applied to the data contained in your work without prejudice to third parties' rights and in compliance with the law or other agreements.

Remember that you are waiving all your rights to the work worldwide under copyright law, including all related and neighbouring rights, to the extent allowed by the law.

Remember that a CC0 licence applied to a dataset allows anyone to copy, modify, distribute and use the dataset and data contained therein, even for commercial purposes, all without asking for permission.

Useful links

Regulatory framework: Italian Law no. 633 of 22 April 1941 "Protezione del diritto d'autore e di altri diritti connessi al suo esercizio (Protection of copyright and related rights)" (in Italian)
https://www.gazzettaufficiale.it/eli/id/1941/07/16/041U0633/sg

Regolamento in materia di proprietà industriale e intellettuale dell'Università di Bologna (in Italian):
https://normateneo.unibo.it/regolamento-in-materia-di-proprieta-industriale-e-intellettuale-delluniversita-di-bologna

Useful resources:

- Creative Commons licensing schemes https://creativecommons.org/
- Extended version of the CC0 licence https://creativecommons.org/publicdomain/zero/1.0/legalcode.it

Respecting privacy in research data management

Carrying out ethical scientific research is essential to ensure the reliability, quality and transparency of research. Personal data protection is an important element of an ethical scientific research.

Many types of research projects involve people – from clinical trials on patients to demographic data collections, from anthropological to linguistic studies. Lawful personal data processing (see box) implies foreseeing and dealing with privacy issues in accordance with the **'privacy by design' principle**, which requires focusing on these matters right from the planning phase. Another key principle is that of **data minimisation**, pursuant to which only personal data that are strictly necessary to achieve a certain purpose (here: a scientific purpose) should be collected and processed.

'Processing' means any operation or set of operations which is performed on personal data or on sets of personal data. Examples include collection, use, organisation, storage and destruction of personal data.

Personal data

'Personal data' means any **information that identifies** or makes it possible to identify **a natural person**, either directly (e.g. name and surname, online identifier, or personal images) or indirectly (e.g. data relating to their habits, lifestyle, health, financial status, or a code assigned to them within a scientific research project).

'Special' categories of personal data include data revealing racial or ethnic origin, religious or philosophical beliefs, political opinions, trade union

Applicable regulations

"Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica (Ethics standards for data processing for statistical or scientific research purposes)"; "Autorizzazione generale al trattamento dei dati genetici (Aut. Gen. 8/2016) (General Authorisation for the processing of genetic data (Gen. Aut. 8/2016))" and "Prescrizioni relative al trattamento dei dati personali effettuato per scopi di ricerca scientifica (Aut. Gen. 9/2016) (Prescriptions on the processing of personal data for scientific research purposes (Gen. Aut. 9/2016))" issued by the Italian Privacy Authority; Regulation (EU) 2016/679 "General Data Protection Regulation".

membership, genetic data, biometric data, data concerning health, sex life or sexual orientation, legal data and data relating to criminal convictions and offences. These types of data require an extra level of protection because they can lead to discriminations.

Planning privacy management

Start by asking yourself if you really need to collect and process personal data for your specific research activity and by deciding whether to collect new data or reuse existing data collected in the past (also by third parties).

All research participants must receive clear, transparent and appropriate **information on the processing of personal data (privacy notice)** including information about the research project, and the purposes and methods of personal data processing. This document must also indicate the legal basis for the processing (e.g. consent, which can be collected in a number of ways – paper, online, videorecording, etc.) and the rights of data subjects over their own personal data.

To draft this privacy notice, you must **define beforehand** for how long the personal data collected will be retained in an identifying (i.e. non-anonymous) form, with whom such data will be shared and for what purposes.

Remember that the Italian law mandates approval by a dedicated ethical committee for all clinical research involving the recruitment of patients. In the case of the University of Bologna, the committee responsible for approving clinical research projects is the independent ethical committee of Area Vasta Emilia Centro (CE-AVEC).

For non-clinical research projects that involve the collection of personal data, it is in some cases appropriate to ask for the opinion of the Bioethics Committee. It is not yet mandatory, unless expressly requested by, for example, the funding body or the publisher of a publication containing the data.

In the field!

I am a researcher in the humanities and social sciences, and I collect observational data and data from surveys.
How can I manage cross-cutting issues such as privacy?

If you collect personal data, prepare a privacy notice as early as the planning phase, and have it signed by the persons you will interview or from whom you will receive information. Follow current legislation, such as the GDPR, and ask the competent offices for help in drafting your informed consent form.

Personal data security measures

Assess the technical and organisational measures necessary to ensure personal data protection during the active phases of research (collection, analysis, storage).

The first aspect to consider is **choosing an appropriate data storage system.** In order to comply with the General Data Protection Regulation (GDPR), if you need to use a cloud-based solution to facilitate collaboration with third-party partners in personal data processing, this needs to have servers based in a country that ensures an appropriate level of personal data protection, as established by the adequacy decision of the European Commission. Again, with a view to striking a balance between collaboration and protection, you must clarify who needs to access identifying data to conduct the research and **manage access rights to the folders** in which personal data are stored accordingly. Depending on the sensitivity of the data you process, consider encrypting the folders by using dedicated tools.

Another aspect to consider during short-term storage and data analysis is the possibility of anonymising or pseudonymising data. Anonymisation and pseudonymisation modify personal data to make the data subject unidentifiable or less identifiable, respectively. Both techniques involve removing or modifying direct and indirect personal identifiers and may reduce the quality and usefulness of research data, as they imply a loss of information.

Anonymisation is the processing of data so that the data subjects can no longer be identified. Anonymous data are such for everyone, including the researchers that collected them in the first place. To anonymise data, you need to select all possible direct and indirect identifiers and modify them using the most appropriate strategies. Pay special attention to attribute combinations that can lead to the identification of certain individuals, and to small population samples. To avoid inference disclosure, you can use measures such as generalisation, aggregation, top and bottom coding to hide identifiable outliers, data perturbation, etc. Anonymised data are no longer regarded as personal data under the GDPR.

Pseudonymisation is the processing of identifying data so that they can no longer be attributed to a specific person in the absence of additional information, such as an encryption key. Pseudonymised data are still regarded as personal data under the GDPR. Data pseudonymisation involves removing or encrypting any directly identifiable piece of information. For encryption, you should use random codes for each person and store the encryption key, possibly encrypted, separately from the encrypted data file.

In the field!

I need to pseudonymise two datasets, a quantitative one and a qualitative one. *What can I do?*

To pseudonymise quantitative data:

- Remove or replace all information that enables direct identification (e.g. name, surname, address, telephone number, email address, IP address, etc.) using a random code for each person.
- Encrypt the key that allows to re-identify each record and store it separately from the dataset.
- Generalise or remove indirect identifiers (e.g. age, occupation, etc.) from the dataset.

To pseudonymise qualitative data:

- For text, e.g. interview transcripts, use pseudonyms and general descriptions and mark replacements with [square brackets]. Example: [Person 1] works for [a financial organisation] in Belgium.
- For audio and/or video, use dedicated tools to blur faces and modify voices.

I am a medical researcher, and I need to anonymise my quantitative data.
What strategies can I use and what results should I expect?

- Generalising or removing indirect identifiers reduces the level of detail in the data. E.g. change "Age 27" into "Age group 21-30"; change "Schizoid personality disorder" into "Mental and behavioural disorder".
- Top and bottom coding hides outliers in the data. E.g. "Age group above 70", "Salary below 1,658 euros/month", etc.
- Data perturbation modifies the value of numerical data by adding 'noise' and replacing real values with simulated or average values.

Long-term preservation of personal data

The retention period for identifiable personal data must be decided at the start and disclosed to research participants in the privacy notice. It is unlawful to retain data longer than it is necessary to achieve the purpose for which they were collected. If you need to deposit identifiable personal data in open access to ensure the transparency and reproducibility of research, be aware that this is only possible if there is an appropriate legal basis to do so (e.g. the consent of the data subject).

If this is not the case, you must consider depositing them in a secure repository that allows for restricted access to the data and requires authorisation. Sometimes there may even be a committee responsible for evaluating requests for data access and reuse.

As already mentioned, anonymised data are no longer regarded as personal data under the GDPR. For this reason, careful anonymisation is the best strategy to make data available in a repository at the end of research.

Useful links

Intranet page on Personal data processing for scientific research (in Italian)
https://intranet.unibo.it/Ateneo/Web1/Pagine/PrivacyRicerca.aspx

Laws and regulations:

- Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica (Ethics standards for data processing for statistical or scientific research) (in Italian)
  https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9069637
- Autorizzazione generale al trattamento dei dati genetici (General Authorisation for the processing of genetic data) (in Italian)
  https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/5803688
- Prescrizioni relative al trattamento dei dati personali effettuato per scopi di ricerca scientifica (Aut. Gen. 9/2016) (Prescriptions on the processing of personal data for scientific research purposes (Gen. Aut. 9/2016)) (in Italian)
  https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9124510#5
- Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR) https://eur-lex.europa.eu/eli/reg/2016/679/oj
- Adequacy decision https://ec.europa.eu/newsroom/article29/items/614108 |
  https://www.garanteprivacy.it/temi/trasferimento-di-dati-all-estero

Useful tools:

- Encrypting data https://www.veracrypt.fr/en/Home.html |
  https://docs.microsoft.com/it-it/windows/security/information-protection/bitlocker/bitlocker-overview
- Blurring faces in a video https://coehelp.uoregon.edu/using-openshot-to-blur-a-face-in-a-video/
- Modifying recorded voices https://www.qualitative-research.net/index.php/fqs/article/view/512/1106
- Anonymising data https://amnesia.openaire.eu/ | https://arx.deidentifier.org/ | https://github.com/sdcTools/sdcMicro