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.