



How to Organize Research Data?

Data organization begins in the planning phase but becomes especially important when data collection starts. Well-organized file names and folder structures make it easier to find and keep track of data files, and to share and reuse research data.

Documentation of the file naming convention and folder structure is also important metadata to provide when sharing or publishing research data.

 Make sure that your data organization system is practical and used consistently.

-  **Good to know:** To organize data properly, make sure to:
- Name your files in a structured and consistent manner.
 - Choose appropriate file formats.
 - Organize files in a workable folder structure.
 - When available, adopt data organization standards from your discipline.

File Naming Conventions

Information to Include in a File Name

Good file names can provide useful cues to the content and status of a file, uniquely identify a file and help in classifying files. You might consider including some of the following information in your file names, but you can include any information that will allow you to distinguish your files from one another.

- Project or experiment name or acronym
- Location/spatial coordinates
- Researcher name/initials
- Date or date range of experiment
- Type of data
- Version number of file

Version Control

Version control is a system that keeps track of changes made to a file over time so that earlier versions can be recalled later. This can be either done automatically (e.g. via Git) or manually by adding version numbering to a file. Each new version indicates file revisions or edits, which is especially useful in collaborations.

Examples for versioning:

File name	Changes to file
LC_InterviewSchedule_1.0	Original document
LC_InterviewSchedule_1.1	Minor revisions
LC_InterviewSchedule_1.2	Further minor revisions
LC_InterviewSchedule_2.0	Substantive changes

Some types of text documentation (e.g. a "terms and conditions" document) may include a changelog at the end of a document, describing the main changes in each version.

Examples of Good File Names

- 20100212_Transcript_FG1_CONS is the file that contains the transcript of an interview with the first focus group of a consumer study, that took place on 12 February 2010.
- 20080605_Int024_R01 is an interview with participant 024, interviewed by researcher R01 on 5 June 2008. (It can be useful to mask the researchers instead of using their names or initials if you want to conduct blind analyses.)

💡 **Good to know:** When naming files, keep in mind:

- Choose short but meaningful names.
- Avoid special characters such as ~ ! @ # \$ % ^ & * () ` ; < > ? , [] { } ' " and | in file names.
- Avoid spaces and use other options instead (e.g. underscore, dash, no separation, CamelCase).
- File names with year-month-day start in the format YYYYMMDD or YYMMDD so that the files are listed chronologically.
- When possible, follow naming conventions that are standard or broadly used in your field.

🔗 **Learn more:**

- Primer on file naming conventions by the Center for Reproducible Science and Research Synthesis: van de Wiel, H., Fraga Gonzalez, G., Furrer, E., & Held, L. (2024). Primer: File Naming Conventions. Zenodo. <https://doi.org/10.5281/zenodo.13322390>

File Formats

File formats may be either proprietary or free and may be either unpublished or open. The file formats you use have a direct impact on your ability to open those files at a later date and on the ability of other people to access those files.

When selecting file formats for archiving, the formats should ideally be:

- non-proprietary
- unencrypted
- uncompressed
- in common usage by the research community

File Formats for Reusability and Preservation

The file formats listed in the table below can be categorized into “appropriate”, “acceptable” and “not suitable” for long-term preservation. If possible, avoid file formats from the column “not suitable”.

Type of Data	Appropriate	Acceptable	Not suitable
Tabular data with extensive metadata	.csv / .tsv / .hdf5	.txt / .html / .tex / .por	
Tabular data with minimal metadata	.csv / .tsv / .tab / .ods / SQL	.xml if appropriate DTD / .xlsx	.xls / .xlsb
Textual data	.pdf / .txt / .odt / .odm / .tex / .md / .htm / .xml	.pptx / .pdf with embedded forms / .rtf	.doc / .ppt
Code	.m / .R / .py / .iypnb / .rstudio / .rmd / NetCDF	.sdd	.mat / .rdata
Digital image data	.tif (uncompressed) / .png / .svg / .jpeg	.jpg / .jp2 / .tif (compressed) / .tiff / .pdf / .gif / .bmp	.indd / .ait / .psd
Digital audio data	.flac / .wav / .ogg	.mp3 / .mp4 / .aif	

Digital video data	.mp4 / .mj2 / .avi / .mkv	.ogm / .webm	.wmv / .mov
Geospatial data	NetCDF, tabular GIS attribute data, .shp / .shx / .dbf / .prj / .sbx / .sbn / PostGIS / .tif / .tfw / GeoJSON	.mdb / .mif	
CAD / vector and raster data	.x3d / .x3dv / .x3db / PDF3D .pdf	.dwg / .dxf	
Generic data	.xml / .json / .rdf		

Source: EPFL Library. 2018. Recommended data formats. Available at: https://www.epfl.ch/campus/library/wp-content/uploads/2018/05/Recommended_DataFormats_-2018_03_05_Final.pdf; last accessed 09 June 2026.

 Good to know: A note on textual data

The encoding format UNICODE UTF-8 is recommended for interoperability. If formatting is not important, convert the file to plain text. If formatting is important, the recommended format for archiving is .pdf (following the standardized PDF/A version).

 Good to know: Compressing files

If you need to use compressed files, favor lossless compression as opposed to lossy compression formats that involve permanently discarding original data. Depending on the data features that are scientifically relevant, compressed formats of any type may not be acceptable.

 Learn more:

- [ETHZ: File formats for archiving](#)
- [CLARIN: Format recommendations for language data](#)

Folder Structure

When organizing the folder structure, it helps if the data is divided into categories, e.g. by project/subproject, time, date, file type, location, etc. When possible, follow folder and file organization standards in your field and organize folders in a way that allows an external collaborator to find and reuse the files with minimal instructions.

 Good to know:

- The folder structure should not be more than three to four levels deep.
- Keep source (original) data folders clearly separated from folders with processed data.
- Clearly label any folder that contains identifying or particularly sensitive information. Such folders may have additional encryption and they may be excluded from archiving, hence they should be easy to distinguish from others.
- Balance informative names and concision.
- In general, avoid numbering in folder names (e.g., 01_data, 02_documents, etc.). This is discouraged as it can lead to ambiguities, a frequent need to rename folders or to an inflexible folder structure. However, numbering script folders or file names can be helpful to indicate the intended order, e.g. for the execution order of the scripts for analysis.

Examples

Example 1

```
Project_name
├── README.txt*      # Basic project description
├── LICENSE*         # License information clarifying conditions for reuse
├── Analysis
│   └── pipeline_01
├── Data
│   ├── Source/raw  # Data as-collected.
│   └── Derivatives # Preprocessed for analysis.
│       └── pipeline_01
├── Docs
│   ├── Ethics      # Submitted to Ethics Committee
│   ├── Procedures  # Standard Operating Procedures (SOPs), protocols, etc.
│   ├── Publications # Research reports
│   └── Recruitment
└── Scripts

Δ Contains no identifying information
≡ Synchronized with a Git version controlled code repository
* Additional README and LICENSE files can be found at subfolder level (e.g., in the Scripts folder)
```

Example 2

```
Project_name
├── README.txt*      # Basic project description
├── LICENSE*         # License information clarifying conditions for reuse
├── Data
│   ├── Source/raw  # Data as-collected
│   └── Derivatives # Preprocessing and analysis pipelines
│       └── pipeline_01
│           └── analysis_01
├── Docs
│   ├── Ethics      # Submitted to Ethics Committee
│   ├── Procedures  # Standard Operating Procedures (SOPs), protocols, etc.
│   ├── Publications # Research reports
│   └── Recruitment
└── Scripts

Δ Contains no identifying information
≡ Synchronized with a Git version controlled code repository
* Additional README and LICENSE files can be found at subfolder level (e.g., in the Scripts folder)
```

Learn more:

Fraga González, G., Clark, A., Furrer, E., & Held, L. (2024). Primer: Long-term Archiving of Experimental Data (Version v3). Zenodo. <https://doi.org/10.5281/zenodo.13880988>

Editor

University of Zurich
Open Science Services & Center for Reproducible Science and
Research Synthesis

Editorial Office

Melanie Röthlisberger (<https://orcid.org/0000-0002-1959-174X>),
Gorka Fraga González (<https://orcid.org/0000-0002-1857-8607>)
& Stefanie Strebel (<https://orcid.org/0000-0002-3336-4199>)