nccr evolving language
NATIONAL CENTRE OF COMPETENCE IN RESEARCH

Open Research Data (ORD) Guidelines

Background

SNF (hence the NCCR) requires all research data to be maximally freely accessible to the public. More specifically, **research data** should be deposited and published on an **open repository** that complies with the FAIR principles.

Do you have research data for your publication?

Follow the steps on Page 2 to publish your research data and obtain a DOI for it!

Good to know

What is *research data*?

Research data denotes essential data for **reproducing** the results in your publication. Such data may contain personal data or copyright data, in which cases at least the **metadata** should be published, together with anonymized data and other unrestricted data.

What is *metadata*?

Metadata denotes the data **about** your research data, such as a general description, the time period of the data collection or study, collection methods, licensing information, and so on. In principle, such metadata are unrestricted and should always be made openly accessible.

Which repositories shall I choose?

The NCCR **strongly recommends** researchers to choose among the three repositories, **SwissUBase**, **Yareta**, and **Zenodo**, in Table 1, for different institutions, respectively. These repositories are hosted in Switzerland.

Table 1: Overview of repositories

Repository	Supporting institution(s)	Size limit	Access control and licenses
SwissUBase	UZH, UniNe, UNIL	1TB per file	- Dataset access configurable - General open and restricted licenses available - An NCCR special license available
Yareta	UniGe	Free up to 50GB No limit for paid storage	- Dataset access configurable - Open licenses available - Restricted licenses created upon request
Zenodo	N/A	50GB per dataset >50GB upon request	- Dataset access configurable - Open licenses available - Restricted licenses customizable

Other repositories? If you plan to use a different repository other than SwissUBase, Yareta and Zenodo, you can look up in https://www.re3data.org/ to find such repositories that comply with the FAIR principles (Findable, Accessible, Interoperable, Reusable). **OSF** and **Figshare,** for example, are accepted FAIR repositories.

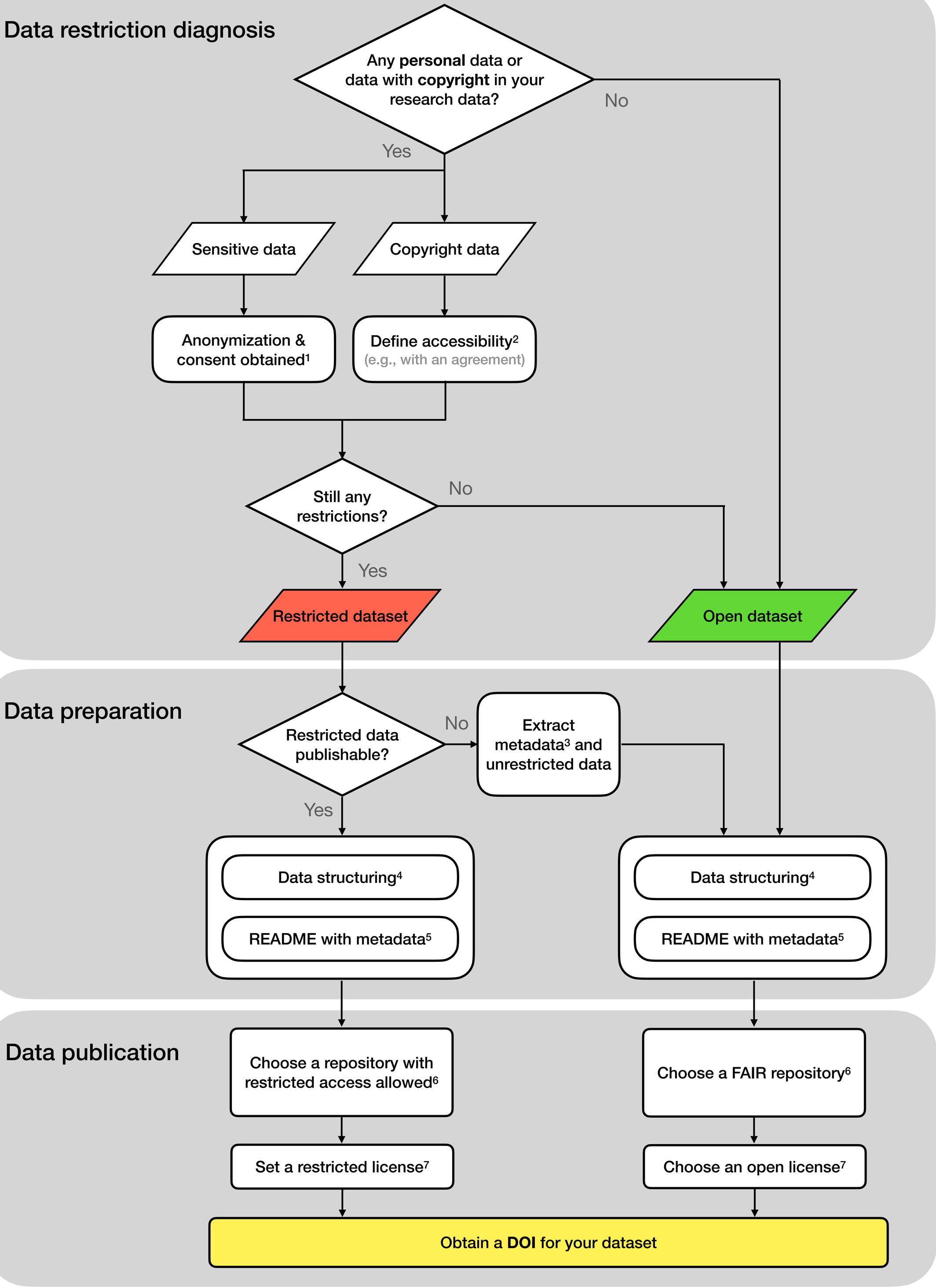


Figure 1. Flow chart of data publication

1. The condition for publishing personal/confidential data should be defined in detail in the consent form before the data collection. By default, such data should always be fully annoymized or destroyed before being published, according to the article 6.3 in the Federal Act on Data Protection of September 25, 2020 ("FADP") . In exceptional cases (e.g., further processing of the personal data is necessary), pseudonymization (i.e., personal data record is securely preserved elsewhere) or even non-anonymized data can be accepted.
2. Essential research data as a small subset of a larger set of copyright data may become accessible and sharable under conditions defined in a license/agreement/contract.
3. Here, the metadata refers to the top-level information about your research data. See more on Page 1 under "What is *Metadata*".
4. See Page 3 for the full guidelines.
5. See Page 4 for the full guidelines.
6. See Page 1 for more details about the recommended repositories.
7. SwissUBase, Yereta and Zenodo all have the option of choosing the license, open or restricted, for your data upon submission. For an open dataset, a **CC** (Creative Commons) license is commonly applied. See more on CC licenses and other licenses for open research data. Restricted data can be licensed in the three recommended repositories (see Table 1 on Page 1)

Data structuring

We stress three major aspects of data structuring: 1) file format, 2) naming convention, and 3) folder structure.

Special note: for neuroimaging and behavioral data, we recommend that you comply with the Brain Imaging Data Structure (BIDS) standard to organize and structure your data.

File format

Ideally, research data should be converted to formats that are suitable for long-term data archiving/preservation. In principle, research data files should always use non-proprietary formats to ensure interoperability.

If the goal is to preserve the data for more than 10 years, only certain formats are suitable, such as **.txt* for textual data, **.csv* for tabular data, **.png* for image data, and **.wav* for audio data. For archiving up to 10 years, the range of accepted formats is wider. Examples include:
- Textual data: *.docx, *.pdf
- Tabular data: *.xlsx, *.ods
- Image data: *.jpeg, *.bmp
- Video/audio data: *.mp4

More examples and instructions can be found **here**, provided by ETH Library.

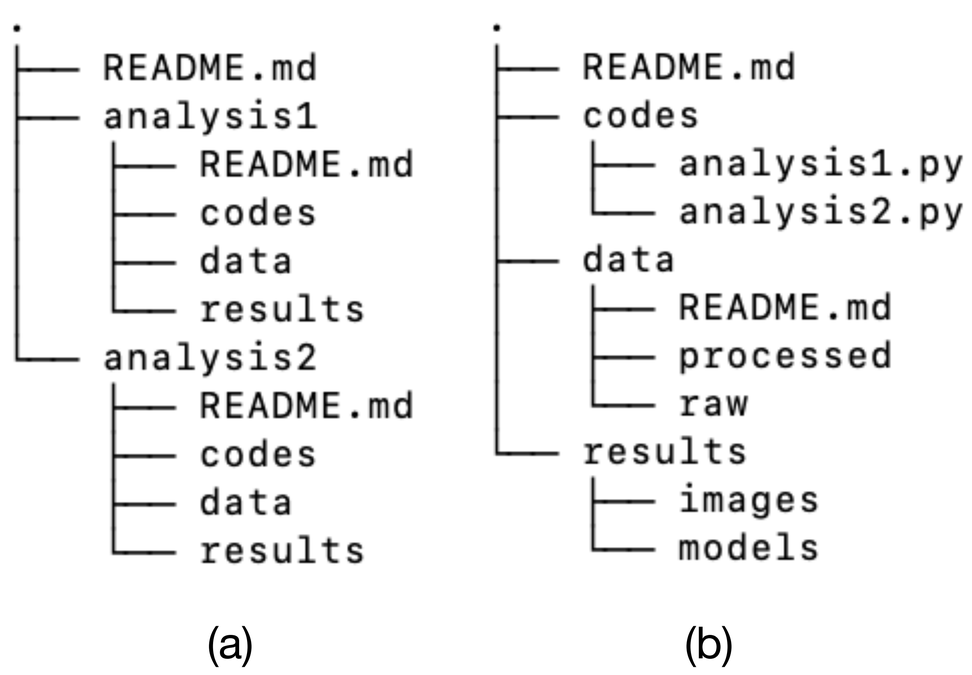
Naming convention

Although there is no universal rule for naming data files, it is important to keep a set of naming conventions for the files within your dataset. File names should be **descriptive**, **consistent**, and **interoperable** with such conventions.
- **Descriptive**: File names should include useful information that can help identify the file content and facilitate the data processing later, such as date, time, location, subject, condition, analysis type, and the correct file extension in the end.
- **Consistent**: File naming should be consistent across similar files to facilitate the organization and reuse of data. Consistency is crucial for efficient description of the dataset.
- **Interoperable**: Some characters and patterns of filenames are not compatible in certain operating systems or software, and thus could result in errors or other issues in processing. Below are a couple of general recommendations:
  - Use ASCII characters and interoperable special characters such as _ and **-**
    - These characters should be avoided: **\ / ? : * “ > < | : # % " { } | ^ [ ]** `` ` ~ and **blanks**
    - Period (**.**) is, in principle, exclusively used to precede the extension.
  - Consider using one of the case patterns below to facilitate machine processing:
    - snake_case: an_example_filename.csv
    - kebab-case: an-example-filename.csv
    - camelCase: anExampleFilename.csv
    - PascalCase: AnExampleFilename.csv

Folder structure

Data files should be hierarchically and/or logically organized in folders. How the files are organized should be described in detail as part of the metadata (e.g., in a README file) of the dataset. In case of a complex folder structure, each folder or subfolder can have its own README file to describe the contents.

Fig. 2 shows two examples of folder structure. Data files are grouped either by project hierarchy (see Fig. 2a) or by contents (see Fig. 2b; see also an example dataset by NCCR researchers in Phaniraj et al., 2023).



(a) (b)

Figure 2. Examples of folder structure

References

Phaniraj, N., Wierucka, K., Zürcher, Y., & Maria Burkart, J. (2023). Data and codes: Who is calling? Optimising source identification from marmoset vocalisations with hierarchical machine learning classifiers (Version V1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8367132

How to write a README file?

Fancy a quick generic template?

https://cornell.app.box.com/v/ReadmeTemplate

Need to see some examples?

- A concise and aesthetically pleasing README for a dataset with source codes by Sarkar et al. (2023). View the README file directly on https://github.com/idiap/ssl-caller-detection.
- A README with a comprehensive metadata description for the AUTOTYP database by Bickel et al. (2023). View the README file directly on https://github.com/autotyp/autotyp-data.

Write your own README?

It is generally advised that you should consider adding the information below to the README.

General information

- A title for the dataset
- Name/institution/address/email information for all researchers involved
- Contact person for questions
- Date of data collection (can be a single date, or a range)
- Information about geographic location of data collection

Data and file overview

- For each filename, a short description of what data it contains
- Dataset/folder structure
- Date that the file was created

Sharing and access information

- Licenses or restrictions placed on the data
- Links to publications that cite or use the data
- Recommended citation for the data

Methodological information

- Description of methods for data collection or generation (include links or references to publications or other documentation containing experimental design or protocols used)
- Description of methods used for data processing (describe how the data were generated from the raw or collected data)

Data-specific information

- Variable list, including full names and definitions (spell out abbreviated words) of column headings for tabular data
- Specialized formats or other abbreviations used

References

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., Bierkandt, L., Zúñiga, F., & Lowe, J. B. (2023). The AUTOTYP database (v1.1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7976754

Eklavya Sarkar. (2023). idiap/ssl-caller-detection: v0.1.0 (v0.1.0). Zenodo. https://doi.org/10.5281/zenodo.10022610