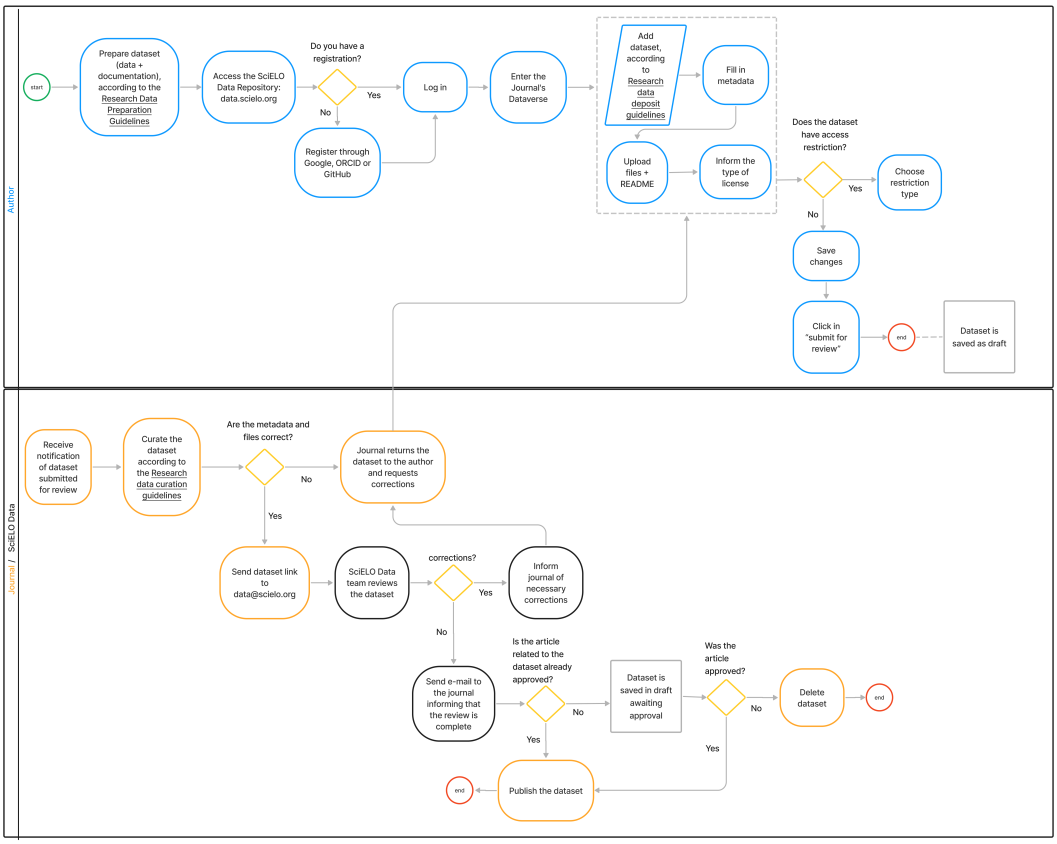
Research data preparation guidelines

April 2023

This is an Open Access document distributed under the terms of the Creative Commons Attribution License **(CC-BY)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Flowchart SciELO Data



The adoption of best practices in the preparation of data for deposit is of key importance for accessing, sharing, and reusing research data, in addition to contributing to making them as FAIR as possible.[1]

To better understanding the FAIR Principles and how to make your data as FAIR as possible, we recommend using the FAIR-Aware tool: https://fairaware.dans.knaw.nl/.

1. File naming

Adopting best file naming practices prevents them from being overwritten and facilitates the location and reusing by other researchers.

- Use meaningful/descriptive names, with 100 characters maximum.
- Do not use accentuation.
- Use alphanumeric characters, underline (my_data) or hyphen (my-data).
- Avoid using blank spaces (my data), periods (my.data), capital initials (MyData) and special characters (such as \ / ? : * ” > < | : # % ” { } | ^ ` ~ @ & ; ° æÆ øØ åÅ äÄ öÖ);
- Use the format YYYY-MM-DD (my_data_2021-01-07) or YYYYMMDD (my_data_20210107) for dates.

[1] We recommend the webinar "*Cómo hacer que los datos sean FAIR? buenas prácticas para datos (abiertos) de investigación*" available from: https://www.youtube.com/watch?v=I14SwZxIRHY.

- Include version number in naming, where appropriate (my-data001.csv, my-data002.csv, ..., my-data010.csv, etc.).
- Use the same naming for files with the same content but different formats (my-data.doc and my-data.txt).

2. File format

SciELO Data accepts any type of file; however, the following formats are recommended:

- Not proprietary.
- Open, with documented international standards.
- With standard character encoding, preferably Unicode (for ex., UTF-8).

Type of document	Recommended formats	Not recommended, however accepted, formats
Compressed files	.zip*	.rar
Statistical analysis	R (.r, .rdata) SPSS (.dat/.sps) STATA (.dat/.do)	SPSS Portable (.por) SPSS (.sav)
Tabular data	Comma Separated Values (.csv) Text file (.txt)	Excel (.xlsx)
Textual data	Text file (.txt) OpenOffice (.odt, .ods or .odp) PDF (.pdf)	Word (.doc or .docx)

* Files compressed with the .zip extension will be unzipped after uploading the files.

3. Files size

The size limit for individual files is 2GB. To add files above this limit, please contact data@scielo.org.

4. Data description

To help ensure that the deposited data can be correctly interpreted and reused, both by you later and by other researchers, it is essential that they be described in as detailed and understandable a way as possible.

This detailed description must be provided by filling in the fields during deposit and a README file, which acts as a guide for the user, and must be deposited with the data files.

The README file should be written as plain text with Unicode UTF-8 character encoding (.txt) or as PDF if you need to illustrate or format the data description.

The README file should contain at least the following information:

- Dataset title.
- Contact information (name, institution, and email) of the corresponding/main researcher or person responsible for data collection.
- Data collection date (single date or time interval).
- Overview of data and files (brief description of the data each file contains, date of creation of each file and how they relate to each other, etc.).
- Description of data collection or generation methods.
- Description of the methods used to process the data.
- Specific data information (list of variables, measurement units, definitions of codes or symbols, equipment calibration, etc.).

For examples of other information that can be added to the README file, see:

- Guide to writing "readme" style metadata

Examples of templates for the README file:

General	• https://drive.google.com/file/d/167cJdaRy4sxQWA5qEEp2cgfWqKV-smZV/view • https://cornell.app.box.com/v/ReadmeTemplate
Social Science	• https://social-science-data-editors.github.io/template_README/template-README.html
Software code	• https://drive.google.com/file/d/1VlDF489DDr044Uta8z1G7EuLj-Wm_UtG/view

For further information on data preparation see also:

Computer codes and data	• Experiences on reproducibility of paper experiments • Research Code
Social Science data	• Guide to Social Science Data Preparation and Archiving
Tabular data	• Preparing tabular data for description and archiving

5. Data anonymization

Must be anonymized: Personal data, sensitive or not,[2] information that violates the right to privacy of the people involved, or puts them at risk, as well as coordinates of protected areas, geographic location of areas that protect species under extinction risk, or information that violates commercial agreements, patents or belong to third parties.

Example of anonymized data[3]:

Information not anonymized	Answer mot anonymized
Name	Juan Pérez
Original country	Argentina
Age	54
Years of experience	25
Aircraft model	Boeing 777 Boeing 747
Last flight date	05/01/2022

Anonymized information	Anonymized answer
-	-
Continent	South America
Age Range	50-60
Years of experience	10-20
Aircraft model	Commercial
Last flight date	01/2022

In cases where the practice of anonymization is impossible, try to use pseudonyms or consider not publishing.

References

Cornell University. Guide to writing "readme" style metadata. *Cornell University* [online]. [viewed 12 February 2021]. Available from: https://data.research.cornell.edu/content/readme.

[2] The following may be considered personal data: name and surname; home address; e-mail address (if it contains elements that help identify the owner, such as first and last name); gender; date of birth; number of registration documents, such as RG, CPF and social security numbers; geolocation data from a cell phone; personal phone number. https://portal.fiocruz.br/noticia/entenda-melhor-lei-geral-de-protecao-de-dados-pessoais. Portuguese. Access 21 mar 2023.
Sensitive personal data: "personal data about racial or ethnic origin, religious conviction, political opinion, union affiliation or organization of a religious, philosophical or political nature, data referring to health or sexual life, genetic or biometric data, when linked to a natural person". https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Portuguese. Access 30 jan 2023.

[3] Example from: Gestión de Datos de Investigación - Parte I. Available from: https://www.youtube.com/watch?v=BM-lZ2XCCN0

DataverseNO. Prepare your data. *DataverseNO* [online]. [viewed 12 February 2021]. Available from: https://site.uit.no/dataverseno/deposit/prepare/.

Nanyang Technological University. DR-NTU (Data) User Guides and Policies. *Nanyang Technological University* [online]. [viewed 12 February 2021]. Available from: https://libguides.ntu.edu.sg/drntudataguidespolicies/depositor#s-lg-box-21651979.

UC Santa Barbara Library. The Dos and Don'ts of file naming. *UC Santa Barbara* [online]. [viewed 30 March 2021]. Available from: https://www.library.ucsb.edu/sites/default/files/dls-n01-2021-filenaming.pdf.

University of Illinois at Urbana-Champaign. Dataset Documentation. *Illinoius Data Bank* [online]. [viewed 12 February 2021]. Available from: https://databank.illinois.edu/help#dataset_documentation.

How to cite this document

SciELO. *Research data preparation guidelines* [online]. SciELO, 2023 [cited DD Month YYYY]. Available from: _____.