Research data preparation guidelines

May 2025
Version 1.1

1. Research data preparation **2**
2. Data anonymization **4**
2.1 Anonymization of transcripts 5
2.2 Image anonymization 5
3. File format **6**
4. File naming **8**
5. Data description **8**
6. README file **9**
7. Files size **11**
8. Changes to the manuscript **11**
References 12
Annex 1. SciELO Data Flowchart **13**
Annex 2. Suitable formats for preservation **14**

1. Research data preparation

The adoption of best practices in the preparation of data for deposit is of key importance for accessing, sharing, and reusing research data. Preparing data appropriately involves organizing and documenting information in a way that promotes access and reuse by other researchers, significantly impacting research progress and scientific transparency, and helping researchers follow the FAIR Principles and be as FAIR as possible.[1]

FAIR data is data that meets standards of findability, accessibility, interoperability and reusability. Data curation, based on the FAIR principles, aims to promote these standards. To achieve these goals, it is essential that data is stored in appropriate repositories, with clear and consistent metadata, facilitating its discovery and access.

To better understand 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/. After publishing the datasets, it is possible to check the FAIRness level using the Fuji tool: https://f-uji.net/.

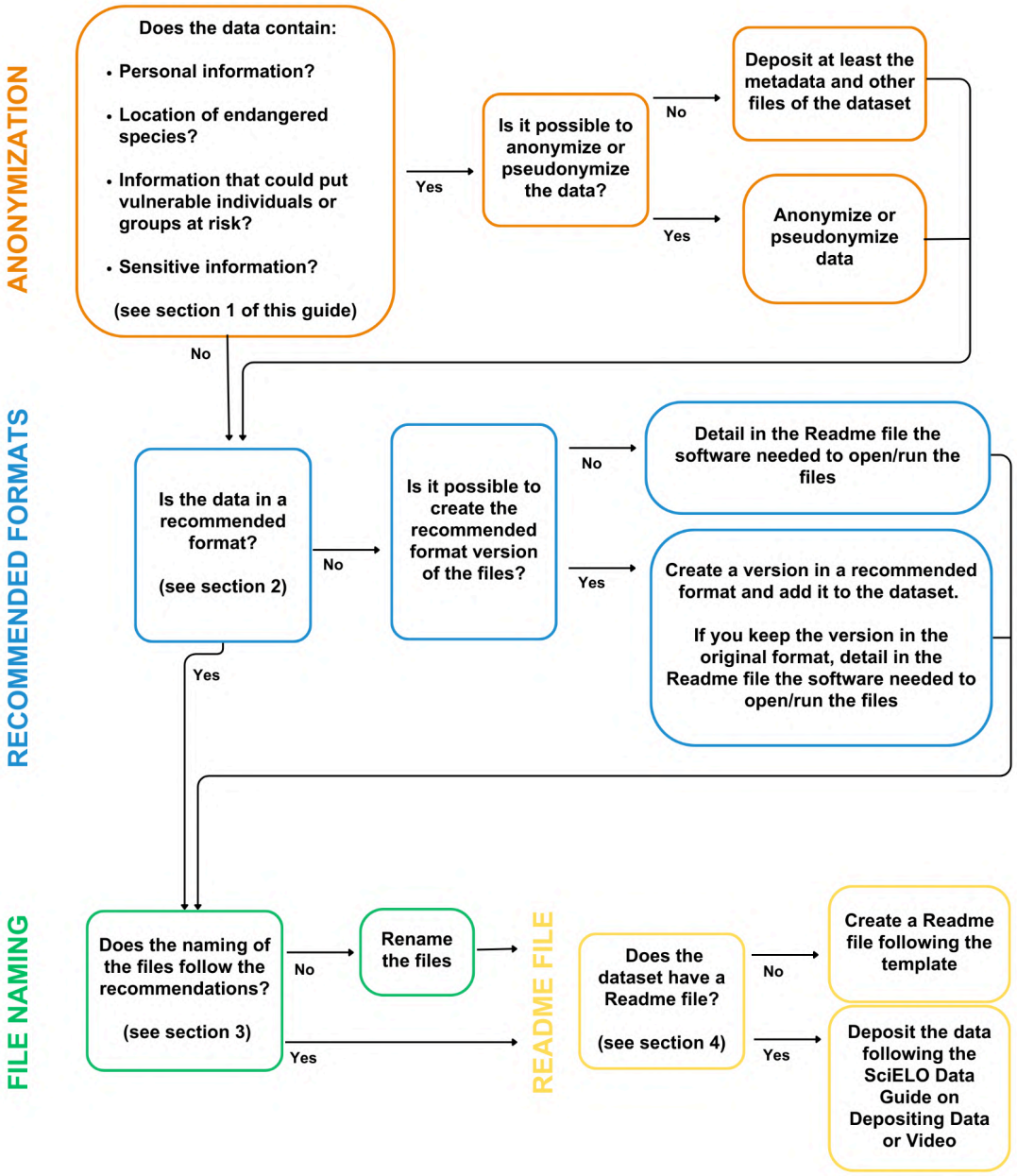
In addition to the FAIR principles, it is important to consider ethical and cultural guidelines related to research data, particularly when dealing with information that involves communities, such as indigenous peoples, or vulnerable populations and their rights. It was in this context that the *CARE*[2] (Collective Benefit, Access, Respect, Ethics) that seek to ensure that the use of data respects ethical and cultural norms.

[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=l14SwZxIRHY.

[2] VIDOTTI, Silvana Aparecida Borsetti Gregorio; TORINO, Emanuelle; CONEGLIAN, Caio Saraiva. #SejaJUSTOeCUIDADOSO: princípios FAIR e CARE na gestão de dados de pesquisa. In: SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). Princípios FAIR aplicados à gestão de dados de pesquisa. Rio de Janeiro: Ibict, 2021. p. 201 - 214. DOI: 10.22477/9786589167242.cap15

Applying the CARE principles in conjunction with the FAIR principles helps to create a more fair, responsible and transparent research environment, promoting not only the reusability of data, but also its ethical and responsible use.

When preparing your research data for deposit in a repository, a series of decisions and procedures must be made considering the information contained in the data, as well as the format of its files and related documentation. Below is a flowchart of decisions that can assist in the process of adapting research data:



The video version of this guide is available on Youtube: SciELO Data Deposit Guide (Portuguese only). For information about the SciELO Data data flow, see Annex 1.

2. Data anonymization

The following must be anonymized: Personal data, whether sensitive or not[3], information that exceeds the privacy rights of the individuals involved, or puts them at risk, as well as coordinates of protected areas, under threat of extinction or information that violates commercial agreements, patents or belongs to third parties.

Reduce the presence of direct identifiers in the files that make up the data set to reduce the precision and detail of people, locations or information that cannot be identified through aggregations such as:

- Year or decade of birth instead of precise date of birth;
- Age range rather than specific age;
- Region instead of city;
- Urban/rural or general location (e.g.: North Zone, South Region of the municipality, business building in the city center, etc.) instead of the name of places;
- Occupation or area of expertise rather than specific job title;
- Period of time rather than specific date or time.

Example of anonymized data[4]:

Information not anonymized	Answer mot anonymized
Name	Juan Pérez
Original country	Argentina
Age	54
Years of experience	15
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	Comercial
Last flight date	01/2022

[3] Personal data: The following can be considered personal data: first and last name; residential address; email address (if it contains elements that help identify the owner, such as first and last name); gender; date of birth; registration document numbers such as ID, CPF (individual taxpayer registry), and work permit; geolocation data from a mobile phone; personal phone number. Available at: https://portal.fiocruz.br/noticia/entenda-melhor-lei-geral-de-protecao-de-dados-pessoais. Accessed on March 21, 2023.
Sensitive personal data: “Personal data concerning racial or ethnic origin, religious beliefs, political opinions, membership in a trade union or a religious, philosophical, or political organization, data concerning health or sex life, genetic or biometric data, when linked to a natural person.” https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Accessed on January 30, 2023.
[4] Example from: Gestión de Datos de Investigación - Parte I. Available from: https://www.youtube.com/watch?v=BM-lZ2XCCN0

2.1 Anonymization of transcripts

To anonymize transcripts, it is not enough to remove the name of the person interviewed. Carefully analyze the interview responses and remove information that allows direct or indirect identification, such as:

- Names of places mentioned that can identify where the person was born, lives or works;
- Telephone number, document number, date of birth or address;
- Position held or specific work performed that could identify the person or institutional affiliation;
- Citations of other people's names (e.g., names of teachers or coworkers).

Avoid deleting or replacing items without indicating that an edit was made. Use pseudonyms or replacements in square brackets (e.g.: “Ana lives in Brasília” -> “[E1] lives in [city in the Brazilian Midwest]”) to indicate the location where the edit was made.

Ana lives in **Brasília** → [**E1**] lives in [**city in the Brazilian Midwest**]

2.2 Image anonymization

Images and videos should also be anonymized. Blur or cover them to prevent identification of people or other information that cannot be shared in open access (street names, license plates, document numbers, etc.):

Example of anonymized images:

Source: https://25.scielo.org/fotos/. Photo: Carla Formanek

The same treatment should be applied to audio files, paying attention to the possibility of identification through voice, accent and language defects.

In cases where anonymization is impossible, try to use pseudonyms or, if the participants are still identifiable, do not deposit the file in question, only the other files that will form the dataset (such as the questions asked and analysis of the responses), as well as a Readme file or

other documentation file for the dataset (see section 4), explaining that not all files are available and the reason for preventing sharing in open access.

Some tools that can help with anonymization are: Amnesia from Openaire, anonymoUUs from Utrecht University and Text Anonymisation helper tool from UK Data Service.

3. 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).

Files in non-proprietary formats do not depend on specific software to be opened. Therefore, give preference to the formats recommended below, in order to facilitate access to the files by other people, as well as data preservation and long-term access.

Table 1. Recommended Formats[5]

Type of document	Recommended formats	Not recommended, however accepted, formats
Statistical analysis	R (.r, .rdata) SPSS (.dat/.sps) STATA (.dat/.do)	SPSS Portable (.por) SPSS (.sav)
Compressed files	.zip *	.rar
Tabular data	Comma Separated Values (.csv) Text file (.txt) Excel (.xlsx)	Excel (.xls)
Textual data	Microsoft Word (docx) Text file (.txt) HTML, XHTML or XML, containing reference to the DTD or format validation schema (.html, .xhtml, .xml) OpenOffice (.odt, .ods or .odp) PDF (.pdf)	Microsoft Word (.doc) Postscript (.ps) Encrypted PDF (.pdf) WordPerfect (.wpd) DVI (.dvi)
Images	PNG (.png) JPEG2000 using lossless compression (.jp2) Uncompressed TIFF	Macromedia Flash (*.swf) Photoshop (.psd) JPEG 2000 Part 2 (.jpf, .jpx) MrSID (.sid)

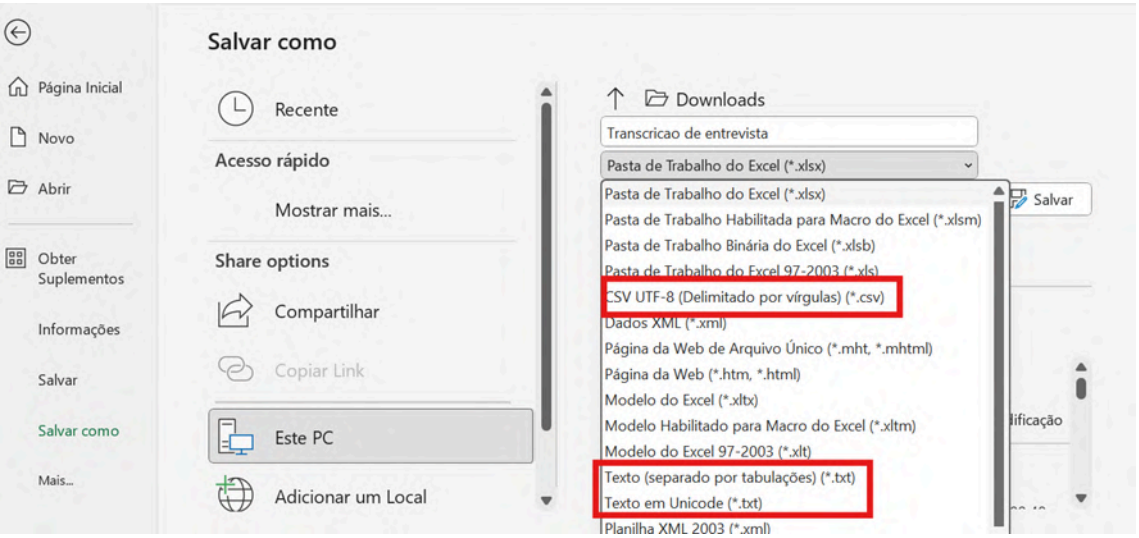
[5] Source: Formatos adequados para preservação. [S.l.]: Arquivo.pt, [s.d.]. Available from: https://sobre.arquivo.pt/pt/colabore/recomendacoes/formatos-adequados-para-preservacao/. Access: 30 apr. 2025.

	SVG (.svg)	TIFF in Planar format (.tiff) FlashPix (.fpx) RAW Encapsulated Postscript (.eps)
Audio	AIFF with Pulse-code modulation (.aif, .aiff) WAV with Pulse-code modulation (.wav, bwf) Ogg Vorbis (.ogg, .oga)	RealNetworks 'Real Audio' (.ra, .rm, .ram) Windows Media Audio (.wma) Compressed WAV (.wav) AIFC with compression (.aifc) NeXT SND(.snd)
Video	Uncompressed QuickTime Movie (.mov) Uncompressed AVI (.avi) Motion JPEG 2000 (ISO/IEC 15444-4) (mj2) Motion JPEG (.avi, .mov)	Windows Media Video (.wmv) Compressed AVI (.avi) QuickTime Movie with compression (.mov) RealNetworks 'Real Video' (.rv, .rm)

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

Once files are deposited in formats that are not recommended, or are not listed as High or Medium suitability for preservation according to Annex 1, there is no guarantee that the files will be properly preserved in the long term.

File format conversion can usually be done through the program used to create the documents. For files created with the Office Suite, for example, simply open the file, click “Save as” and select a corresponding open format, such as the examples highlighted in the image below.



If you convert files from one format to another, please report this in the Readme or supporting documentation deposited with the dataset.

4. File naming

Adopting best file naming practices prevents them from being overwritten and makes it easier for other researchers to find and reuse them.

- Use self-explanatory/descriptive names, with a maximum of 100 characters (e.g.: “interview-child-1”).
- Do not use accentuation.
- Use alphanumeric characters, underscores or hyphens (e.g.: interview-child_1).
- Avoid spaces (e.g.: interview child 1), periods (interview.child.1), capital letters (InterviewChild1) (such as \ / ? : * ” > < | : # % ” { } | ^ ` ~ @ & ; ° æÆ øØ åÅ äÄ öÖ);
- Use the format YYYY-MM-DD (interview_child_1_2021-01-07) or YYYYMMDD (interview_child_1_20210107) for dates.
- Include the version number in the naming when appropriate (interview–child-1-001.csv, interview-child-1-002.csv, …, interview-child-1-010.csv, …);
- Use the same naming for files with the same content but different formats (interview-child-1-data.doc and interview-child-1.txt).

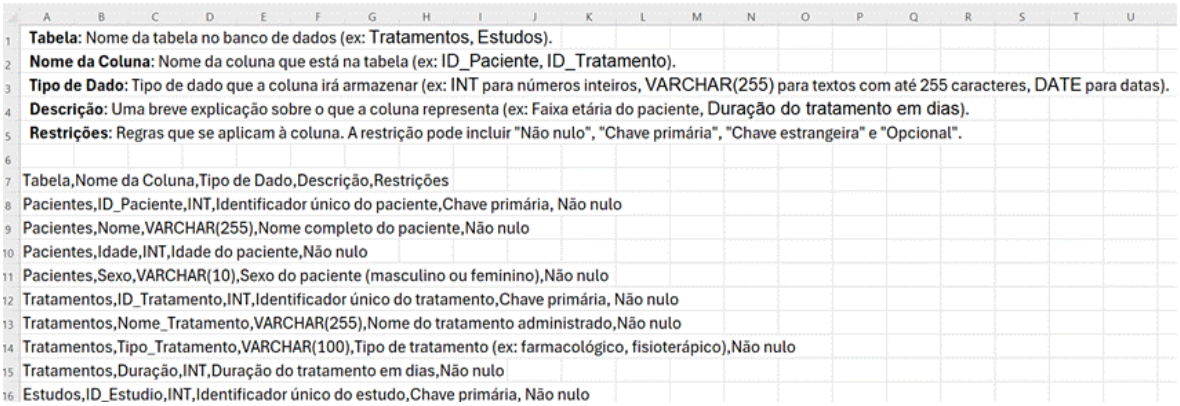
5. Data description

In order for the deposited data to be correctly interpreted and reused, both by you in the future and by other researchers, it is essential that they are described in as detailed and comprehensible a way as possible.

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

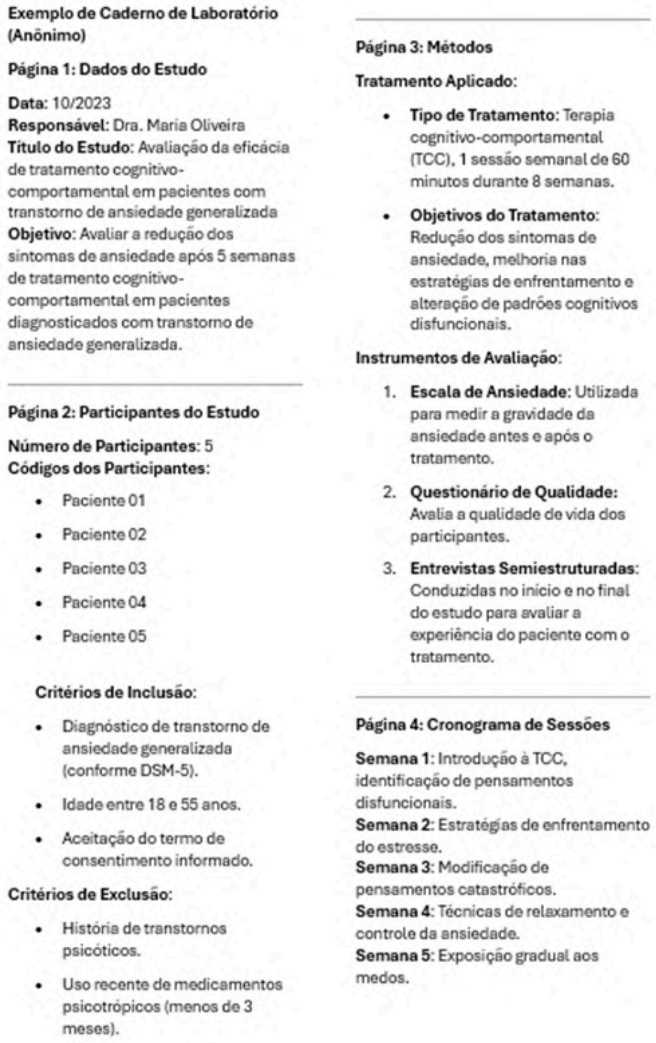
Alternatively, other documentation can be submitted, such as lab notebooks and data dictionaries (which provide detailed information about tables, columns, data types, constraints, and other data characteristics), which contain descriptive and contextual information about the deposited data. Data Dictionary Creator is an app that helps you create a data dictionary from your research files.

Data Dictionary Example



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**Tabela**: Nome da tabela no banco de dados (ex: Tratamentos, Estudos).																				
2	**Nome da Coluna**: Nome da coluna que está na tabela (ex: ID_Paciente, ID_Tratamento).																				
3	**Tipo de Dado**: Tipo de dado que a coluna irá armazenar (ex: INT para números inteiros, VARCHAR(255) para textos com até 255 caracteres, DATE para datas).																				
4	**Descrição**: Uma breve explicação sobre o que a coluna representa (ex: Faixa etária do paciente, Duração do tratamento em dias).																				
5	**Restrições**: Regras que se aplicam à coluna. A restrição pode incluir "Não nulo", "Chave primária", "Chave estrangeira" e "Opcional".																				
6																					
7	Tabela,Nome da Coluna,Tipo de Dado,Descrição,Restrições																				
8	Pacientes,ID_Paciente,INT,Identificador único do paciente,Chave primária, Não nulo																				
9	Pacientes,Nome,VARCHAR(255),Nome completo do paciente,Não nulo																				
10	Pacientes,Idade,INT,Idade do paciente,Não nulo																				
11	Pacientes,Sexo,VARCHAR(10),Sexo do paciente (masculino ou feminino),Não nulo																				
12	Tratamentos,ID_Tratamento,INT,Identificador único do tratamento,Chave primária, Não nulo																				
13	Tratamentos,Nome_Tratamento,VARCHAR(255),Nome do tratamento administrado,Não nulo																				
14	Tratamentos,Tipo_Tratamento,VARCHAR(100),Tipo de tratamento (ex: farmacológico, fisioterápico),Não nulo																				
15	Tratamentos,Duração,INT,Duração do tratamento em dias,Não nulo																				
16	Estudos,ID_Estudio,INT,Identificador único do estudo,Chave primária, Não nulo																				

Example of a Laboratory Notebook



Exemplo de Caderno de Laboratório (Anônimo)

Página 1: Dados do Estudo

Data: 10/2023
Responsável: Dra. Maria Oliveira
Título do Estudo: Avaliação da eficácia de tratamento cognitivo-comportamental em pacientes com transtorno de ansiedade generalizada
Objetivo: Avaliar a redução dos sintomas de ansiedade após 5 semanas de tratamento cognitivo-comportamental em pacientes diagnosticados com transtorno de ansiedade generalizada.

Página 2: Participantes do Estudo

Número de Participantes: 5
Códigos dos Participantes:

- Paciente 01
- Paciente 02
- Paciente 03
- Paciente 04
- Paciente 05

Critérios de Inclusão:

- Diagnóstico de transtorno de ansiedade generalizada (conforme DSM-5).
- Idade entre 18 e 55 anos.
- Aceitação do termo de consentimento informado.

Critérios de Exclusão:

- História de transtornos psicóticos.
- Uso recente de medicamentos psicotrópicos (menos de 3 meses).

Página 3: Métodos

Tratamento Aplicado:

- **Tipo de Tratamento**: Terapia cognitivo-comportamental (TCC), 1 sessão semanal de 60 minutos durante 8 semanas.
- **Objetivos do Tratamento**: Redução dos sintomas de ansiedade, melhoria nas estratégias de enfrentamento e alteração de padrões cognitivos disfuncionais.

Instrumentos de Avaliação:

1. **Escala de Ansiedade**: Utilizada para medir a gravidade da ansiedade antes e após o tratamento.
2. **Questionário de Qualidade**: Avalia a qualidade de vida dos participantes.
3. **Entrevistas Semiestruturadas**: Conduzidas no início e no final do estudo para avaliar a experiência do paciente com o tratamento.

Página 4: Cronograma de Sessões

Semana 1: Introdução à TCC, identificação de pensamentos disfuncionais.
Semana 2: Estratégias de enfrentamento do estresse.
Semana 3: Modificação de pensamentos catastróficos.
Semana 4: Técnicas de relaxamento e controle da ansiedade.
Semana 5: Exposição gradual aos medos.

6. README file

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.

The README file, in particular, is a file that is mandatory for datasets deposited in SciELO Data. It must be written as plain text with Unicode UTF-8 character encoding (.txt) or as a PDF if it is necessary to illustrate or format the data description.

The README file must 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.).

SciELO Data has its own README template, which covers the above topics and can serve as a basis for authors, and is available for download.

When creating your dataset, insert as much information and descriptions of the data as possible, such as a Readme according to the SciELO Data template, together with a data dictionary specifying the columns, rows and variables of the tables present.

Examples of templates for the README file:

General	• https://drive.google.com/drive/folders/1LNxx4YKxN1a2DF10bbVoRW-ioZwNz-9M • 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 examples of other information that can be added to the README file, see:

- Guide to writing "readme" style metadata

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

7. Files size

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

8. Changes to the manuscript

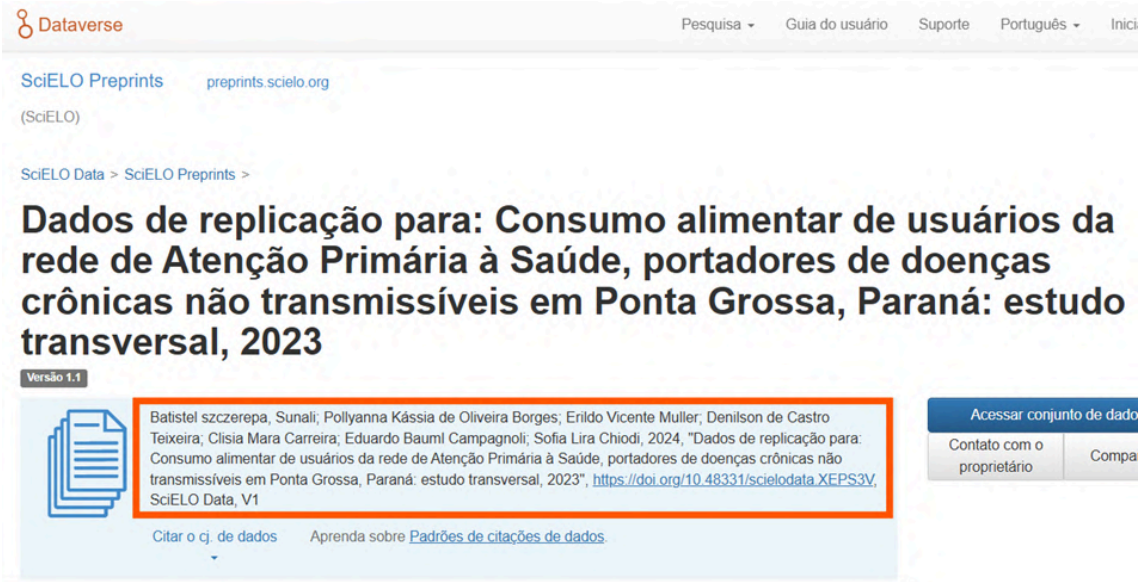
When depositing research data related to an article, the DOI of the dataset must be included in the Data Availability section (or equivalent) as per the journal standard. If you are publishing a preprint, be sure to add this section to the document.

As an example, we suggest:

“The entire anonymized dataset supporting the results of this study has been made available on SciELO Data and can be accessed at [**dataset URL or DOI**].”

It is also possible to cite your own dataset in the body of the text (examples of dataset citation standards are available in the Research Data Citation Guide - Portuguese only), which involves inserting the dataset in the article's reference list following the normative reference standard determined by your journal.

To access the full citation and DOI of the dataset, after submitting the submission to SciELO Data, go to the dataset's home page. The information will be in the blue box below the dataset title. The DOI will only become active after the dataset has been curated by the editorial team and published.



References

Arquivo.PT. Formatos adequados para preservação. [S.l.]: Arquivo.pt, [s.d.]. Available from: https://sobre.arquivo.pt/pt/colabore/recomendacoes/formatos-adequados-para-preservacao/.

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

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.

Syracuse University. De-identification. Qualitative Data Repository, 2023. Available at: https://qdr.syr.edu/guidance/human-participants/deidentification. Access: 30 Apr. 2025

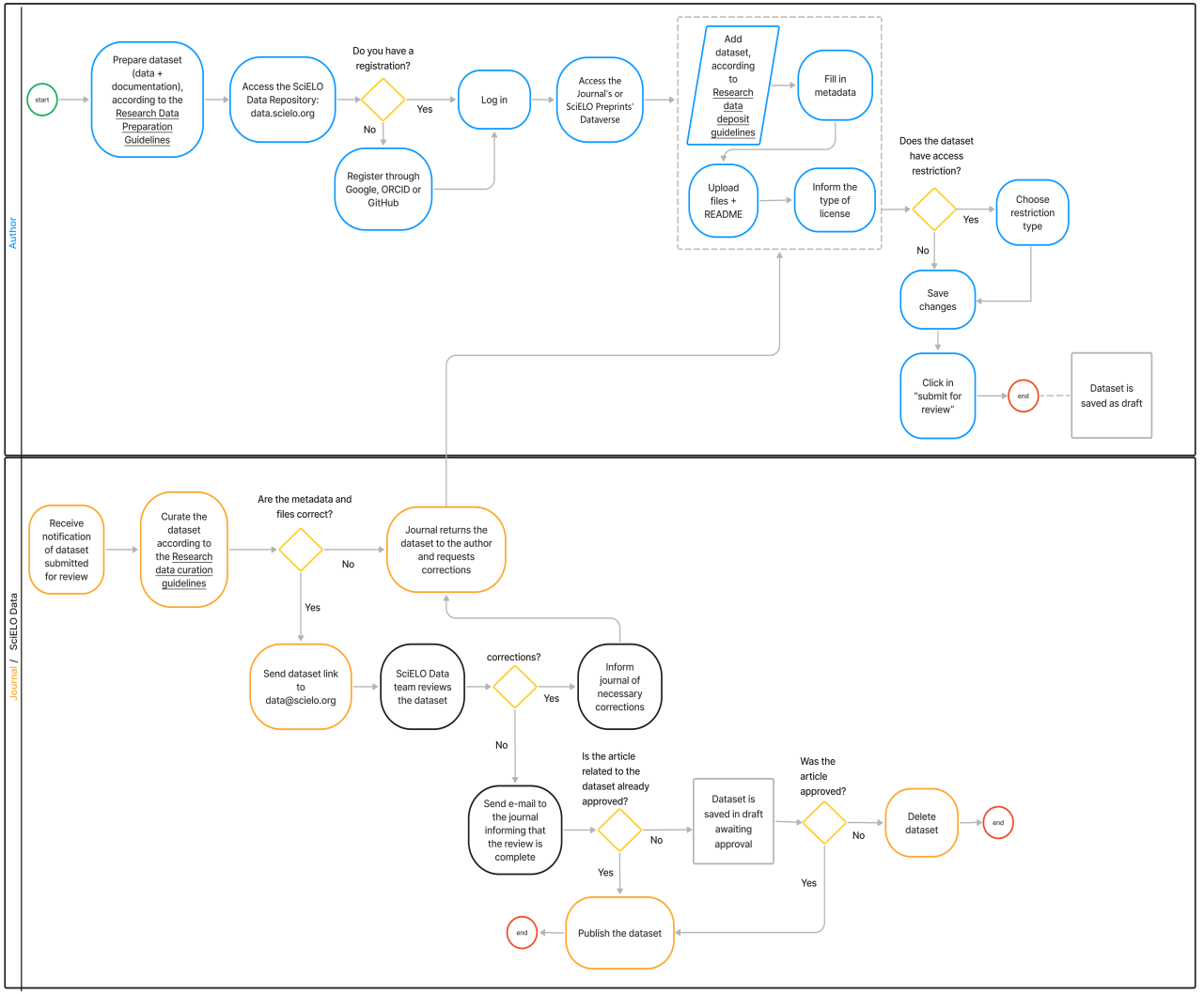
UC Santa Barbara Library. The Dos and Don'ts of file naming. *UC Santa Barbara* [online]. [viewed 30 March 2021]. Available from: https://www.library.ucsb.edu/sites/default/files/dls-n01-2021-filenaming.pdf.

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, 2025 [cited DD Month YYYY]. Available from: _____.

Annex 1. SciELO Data Flowchart



***Only for datasets deposited in the SciELO Preprints Dataverse:** curation is done by the SciELO Data Team. If editing or corrections are required, the SciELO team will contact the depositing author via email to request corrections. The dataset will only be published once the preprint has been approved and the corrections have been made.

Annex 2. Suitable formats for preservation

Some file formats are considered more suitable for long-term preservation because they do not rely on specific software to be opened and accessed. To find out whether your file formats are classified as **High**, **Medium**, or **Low** suitability for preservation, see the list below. Whenever possible, save your files in High suitability for preservation formats.

Text

High suitability for preservation

- HTML, XHTML or XML, containing reference to the DTD or format validation schema (.html, .xhtml, .xml)
- Plain texts using UTF-8, USASCII or UTF-16 character encoding with Byte Order Mark (.txt)
- PDF/A-1 according to ISO 19005-1 (.pdf)
- Open Document Text (.odt)

Average suitability for preservation

- HTML, XHTML or XML, without reference to the DTD or format validation schema (.html, .xhtml, .xml)
- Cascading Style Sheets (.css)
- Plain text with ISO-8859-1 character encoding (.txt)
- PDF with embedded fonts (.pdf)
- Rich Text Format 1.x (.rtf)
- HTML 4.x including DOCTYPE (.html) declaration
- Open Office Text Document (.sxw)
- Office Open XML (.docx)
- DTD (.dtd)
- SGML (.sgml)

Low suitability for preservation

- Microsoft Word (.doc)
- Postscript (.ps)
- Encrypted PDF (.pdf)
- WordPerfect (.wpd)
- DVI (.dvi)

Image

High suitability for preservation

- PNG (.png)
- JPEG2000 using lossless compression (.jp2)
- Uncompressed TIFF
- SVG (.svg)

Average suitability for preservation

- JPEG2000 using lossy compression (.jp2)
- GIF (.gif)
- JPEG/JFIF (.jpg)
- TIFF with compression
- BMP (.bmp)
- Digital Negative (.dng)
- Computer Graphic Metafile and WebCGM (.cgm)

Low suitability for preservation

- Macromedia Flash (*.swf)
- Photoshop (.psd)
- JPEG 2000 Part 2 (.jpf, .jpx)
- MrSID (.sid)
- TIFF in Planar format (.tiff)
- FlashPix (.fpx)
- RAW
- Encapsulated Postscript (.eps)

Audio

High suitability for preservation

- AIFF with Pulse-code modulation (.aif, .aiff)
- WAV with Pulse-code modulation (.wav, bwf)
- Ogg Vorbis (.ogg, .oga)

Average suitability for preservation

- MP3 (MPEG-1/2, Layer 3) (.mp3)
- Free Lossless Audio Codec (.flac)
- SUN Audio Uncompressed (.au)
- Standard MIDI (.mid, midi)
- Advance Audio Coding (.mp4, .m4a, .aac)

Low suitability for preservation

- RealNetworks 'Real Audio' (.ra, .rm, .ram)
- Windows Media Audio (.wma)
- Compressed WAV (.wav)
- AIFC with compression (.aifc)
- NeXT SND(.snd)

Video

High suitability for preservation

- Uncompressed QuickTime Movie (.mov)
- Uncompressed AVI (.avi)
- Motion JPEG 2000 (ISO/IEC 15444-4) (mj2)
- Motion JPEG (.avi, .mov)

Average suitability for preservation

- MPEG-1,MPEG-2 (.mpg, .mpeg)
- MPEG-4 (.mp4)
- Ogg Theora (.ogg, .ogm, .ogv)

Low suitability for preservation

- Windows Media Video (.wmv)
- Compressed AVI (.avi)
- QuickTime Movie with compression (.mov)
- RealNetworks 'Real Video' (.rv, .rm)

Other formats

High suitability for preservation

- Comma Separated Values (.csv)
- SQL DDL

Average suitability for preservation

- OpenOffice (.sxc/.ods, .sxi/.odp))
- OOXML according to ISO/IEC DIS 29500 standard (.xlsx, .pptx)

Low suitability for preservation

- Microsoft Excel (.xls)
- Microsoft PowerPoint (.ppt)
- Microsoft Access (.mdb)
- Microsoft Visio(.vsd)

Source: Formatos adequados para preservação. [S.l.]: Arquivo.pt, [s.d.]. Available from: https://sobre.arquivo.pt/pt/colabore/recomendacoes/formatos-adequados-para-preservacao/. Access: 30 apr. 2025.