



Scientific Electronic Library Online

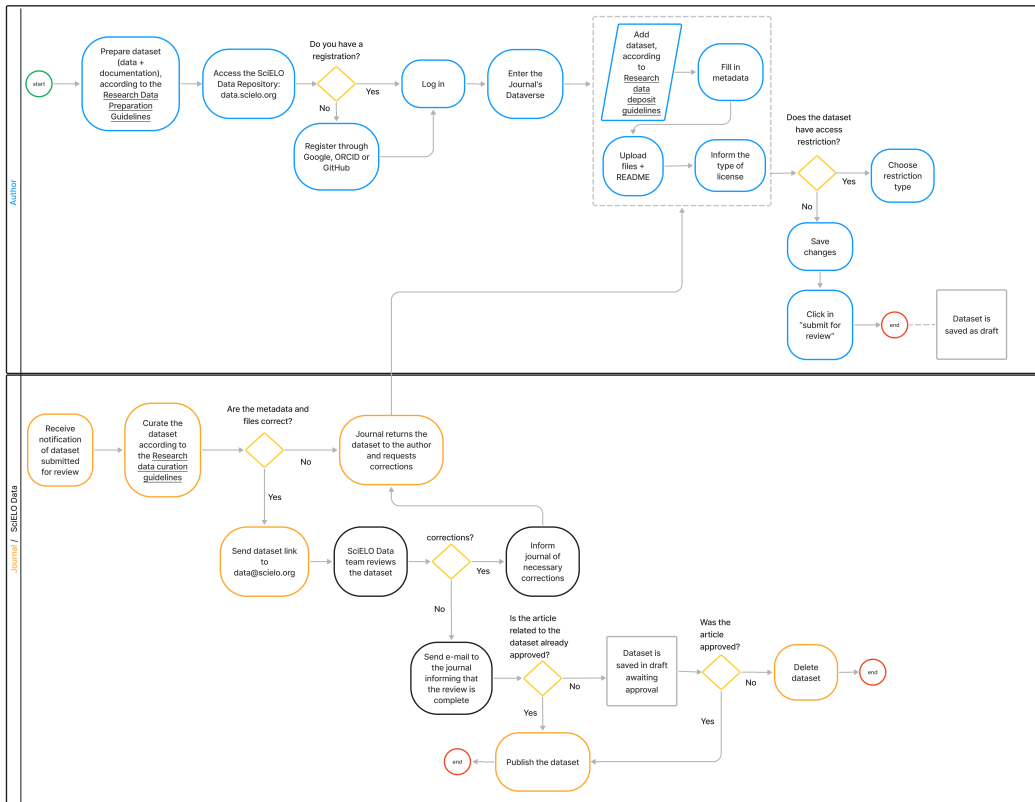
Research data curation guidelines for editorial teams

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



According to CoreTrustSeal¹ (2019), data curation is:

The activity of managing and promoting the use of data from their point of creation to ensure that they are fit for contemporary purpose and available for discovery and reuse. For dynamic datasets this may mean continuous enrichment or updating to keep them fit for purpose. Higher levels of curation will also involve links with annotation and with other published materials.

In other words, Data curation involves the entire Data Life Cycle, from collection planning to preservation for long-term access and reuse. It can also be more specifically defined as the checks and actions carried out by curators aiming to ensure that the dataset is structured and documented as thoroughly as possible and following best practices.

In the context of this guide, the term “curation” will be used considering the second definition.

¹ CoreTrustSeal is an initiative of the International Science Council's (WDS) World Data System and the Data Seal of Approval (DSA). It is a community-based, non-governmental and not-for-profit international organization that promotes reliable and sustainable data infrastructure.

1. Citation of the dataset in the linked article

For there to be a connection between the published article and the underlying dataset, it is recommended that the instructions to authors be documented that the articles must contain the section “Data availability” informing whether the dataset referring to the research is available and, if so, where to access it. it.

Example of content for the section recommended by SciELO:

Non-available data

The dataset supporting the results of this study is not publicly available (Does not apply to articles with datasets in SciELO Data).

Available data

The entire dataset supporting the results of this study was published in the article itself (Does not apply to articles with datasets in SciELO Data).

The entire dataset supporting the results of this study was published in the article and in the section “Supplementary materials” (Does not apply to articles with datasets in SciELO Data).

The entire dataset supporting the results of this study was made available in SciELO Data and can be accessed in [URL or DOI].

The entire dataset supporting the results of this study was made available in SciELO Data with identifiers [list of identifiers].

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

Data available upon request²

The entire dataset supporting the results of this study is available upon request to the corresponding author [name of the corresponding author]. The dataset is not publicly available due to [details of the reason for the restriction, e.g., contain information that compromises the privacy of the research participants] (Does not apply to articles with datasets in SciELO Data).

The entire dataset supporting the results of this study is available upon request to [name of organization]. The dataset is not publicly available due to [details of the reason for the restriction, e.g., contain information that compromises the privacy of the research participants] (Does not apply to articles with datasets in SciELO Data).

² In SciELO Data it is possible to deposit datasets but keep them restricted. When someone wants to access them, they will contact the depositor author of the set through the platform. See more details in item 5.1 of the [Research data deposit guidelines](#).

For detailed information on the application levels of data criteria, codes and research materials in the [Guide for promoting openness, transparency and reproducibility of research published by SciELO journals](#).

2. Data anonymization

Data curation is also important to verify if the data to be published do not violate the ethics and research committees' rules of their respective areas, and to verify if there is data that needs anonymization (such as personal data, sensitive or not³, information that exceeds the privacy rights of people involved, or puts them at risk, as well as coordinates of protected areas, under threat of extinction or information that infringes commercial agreements, patents or belonging to third parties).

The information below already appears in the [Research data preparation guidelines](#), however it is repeated below for verification by the editorial team.

Sensitive data (data that, if exposed without authorization or lost, could result in legal problems), or personal data must be anonymized.

Example of anonymized data⁴:

Information not anonymized	Answer not 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.

³ 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.

⁴ Example from: Gestión de Datos de Investigación - Parte I. Available from: <https://www.youtube.com/watch?v=BM-lZ2XCCN0>

3. Data curation levels

Aiming for transparency regarding the checks and actions carried out by the curators on the deposited datasets, SciELO will reference the levels of curation, from 1 to 3, used by CoreTrustSeal as a requirement in evaluating trusted data repositories:

- A. Basic curation: quick metadata check / content, adding basic metadata or documentation → **Level 1.**
- B. Detailed curation: basic curation + conversion of data files to new formats, documentation enhancement → **Level 2.**
- C. Data-level curation: basic curation + detailed curation + editing the deposited data for greater accuracy → **Level 3.**

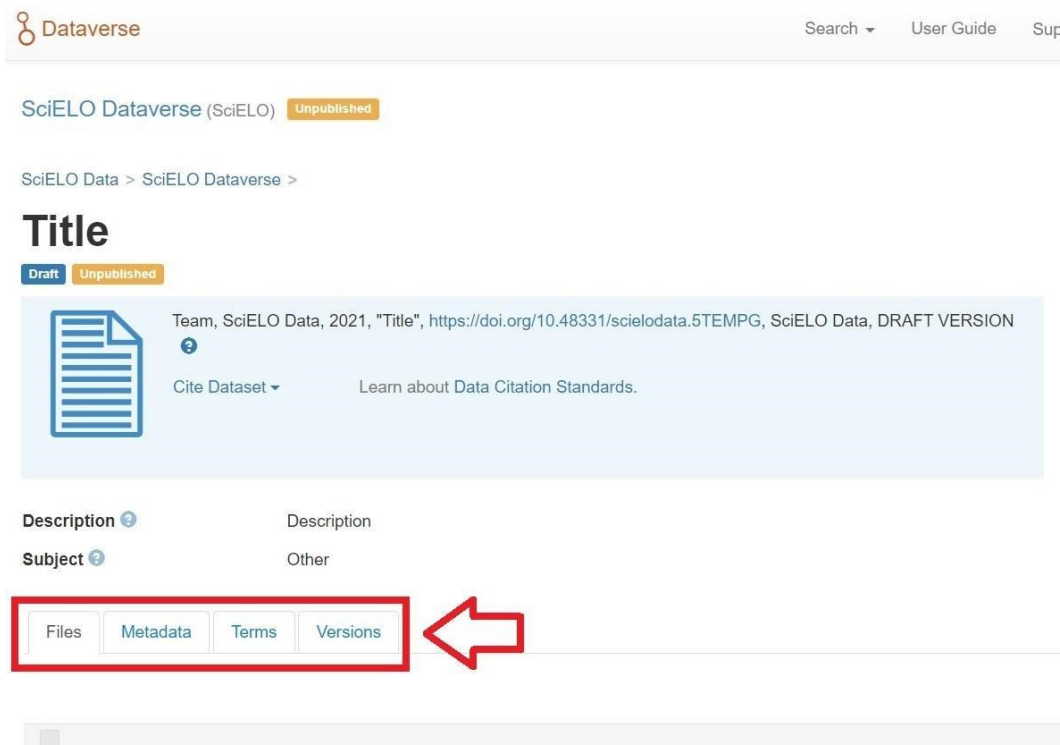
Regardless of the level of curation, it is mandatory to check the files for personal or potentially sensitive data as per item 2 of this guide.

4. Data curation check checklists and actions

At SciELO Data there are 4 sections in which metadata / data files need to be curated considering the curation level adopted and carried out by the journal:

- Files
- Metadata
- Terms
- Versions

You can browse these sections by clicking on the corresponding tabs:



The screenshot shows the SciELO Dataverse interface. At the top, there is a search bar and navigation links for 'User Guide' and 'Sup'. Below this, the page title is 'SciELO Dataverse (SciELO)' with an 'Unpublished' status. The breadcrumb trail is 'SciELO Data > SciELO Dataverse >'. The main heading is 'Title', also with 'Draft' and 'Unpublished' status. A document icon is shown next to the title 'Team, SciELO Data, 2021, "Title", https://doi.org/10.48331/scielodata.5TEMPG, SciELO Data, DRAFT VERSION'. Below the title, there are options to 'Cite Dataset' and a link to 'Learn about Data Citation Standards.'. Further down, there are sections for 'Description' and 'Subject'. At the bottom, there are four navigation tabs: 'Files', 'Metadata', 'Terms', and 'Versions'. A red box highlights these tabs, and a red arrow points to the 'Files' tab.

Level 1 - Basic curation

- Make sure that the dataset is related to any manuscript submitted to the journal. Note: Only approved article datasets can be published.

On the “Files” tab:

- Check if the dataset has been documented in a file named README. The presence of this file is mandatory.
- Check if the naming of files is adequate (see topic 1 of “[Research data preparation guidelines](#)”). If it is not named properly, consider asking the authors to edit following the recommendations.
- Check if files can be opened (not corrupted). If they do not open, request a new deposit from the authors.
- Check the files for data that needs anonymization. If it is necessary to anonymize, refer to item 2 of this guide.
- Check if the files are in the recommended formats (see topic 2 of “[Research data preparation guidelines](#)”).
- Check if the files are in recommended formats (see topic 2 of the “Research data preparation guide”). If you are not recommending that authors edit according to the recommendations.
- Check if the dataset has files with restricted access.
 - If so, verify that the “Terms of Access” field has been filled in with information about users' access to restricted files and how to gain it.

File Restrictions

Limit access to published files by marking them as restricted. Provide users Terms of Access and allow them to request access.

Terms of Access

Request Access Enable access request

Continue Cancel

On the “Metadata” tab:

- **Title:** Check if it is filled in with the article title to which data are related, or with its own title that must be significant/descriptive for the dataset. If not recommend authors to edit.

- **Author:**
 - Check if the authors' names were entered in reverse order (Last Name, First Name).
 - Check if the authors informed their affiliation (mandatory) and ORCID (recommended).
- **Subject:** Check if the selected subject area is the most suitable. Avoid selecting “Other” whenever possible.
- **Keyword:** Check if each keyword has been entered separately (each word in separate fields). If not, edit them on the metadata editing screen and add keywords by pressing the “+” sign.
- **Related Publication:**
 - If the dataset is related to a manuscript under review insert the **manuscript title**.
 - If the dataset is related to a published article, insert an **article citation with DOI**.
- **Funding Information:** If the survey has a funding source, click on “Edit Dataset” and then “Metadata”. Find the field and fill it in.

On the “Terms” tab:

- Check if the chosen license type for the dataset was CC BY 4.0. If you want to use another license, please contact data@scielo.org.

On the “Versions” tab:

- Check whether it is a new dataset or a new version of an already published dataset.

Level 2 - Detailed curation → Perform basic curation and:

On the “Files” tab:

- Rename files in the most suitable way (see topic 1 of the “[Research data preparation guidelines](#)”).
- Evaluate if files are in recommended formats. If necessary, convert files to recommended formats (see topic 2 of the “[Research data preparation guidelines](#)”).
- Evaluate whether or not the documentation provided (README file, codebook, etc.) is complete and understandable (see topic 4 of the “[Research data preparation guidelines](#)”). If not, request the necessary changes from the authors.

On the “Metadata” tab:

- Evaluate the information provided to determine if it is complete and understandable. Request corrections or make edits if necessary.

On the “Terms” tab:

- Evaluate the information provided to determine if it is complete and understandable. Request corrections or make edits if necessary.

Level 3 - Data-level curation → Perform basic curation + detailed curation and:

On the “Files” tab:

- Download data files.
- Open the data files and check if they need any additional treatment. If necessary, ask the authors for corrections or make changes and let them know.
- Open the data files and evaluate them for possible issues such as: appropriate variables and value definitions, out-of-range values, descriptions of programs used for coding files and preferred data structures. If necessary, ask the authors for corrections or make changes and let them know.
- Run and troubleshoot code files.
- Check consistency (checksum⁵) of dataset files to ensure data integrity at bit level.

If the dataset is not properly structured and/or documented, the curator can return it to the author (click on “Publish Dataset” and then on “Return to Author”).

If the dataset is suitable for publication, send an email to data@scielo.org informing the dataset URL and request verification. The SciELO curatorship is temporary. After the SciELO Data team returns, the journal can continue with the publication (click on “Publish Dataset” and then on “Publish”). The author will receive an email informing that the dataset has been published.

IMPORTANT:

Once published, the dataset cannot be deleted. It is recommended that the set be edited, thus generating a 2nd version of the data set. The DOI will remain the same.

For exceptional cases that require exclusion, it is not possible to “disappear” with the dataset, only “disable it”. The DOI of the data set will lead to the same page that will display the citation of the set with the information “DEACCESSIONED VERSION” and the reason for the unavailability.

If, after publication, any dataset needs changes or access is lost, please contact data@scielo.org.

SciELO Data also allows sharing the dataset with peer reviewers through a private URL⁶.

⁵ A checksum is a sequence of numbers and letters used to verify data integrity, that is, whether a file is the same after a transfer, verifying that it has not been altered by a third party or is not corrupted.

⁶ The creation of a private URL allows sharing (for viewing and download files) an unpublished dataset with a group of people who may not have a SciELO Data user account, in other words, anyone who receives the private URL will be able to access the dataset with no need to register or log in to SciELO Data.

To create a private URL, go to dataset → click “Edit” → “Private URL” → in the popup box choose “Create Private URL” or “Create URL for Anonymized Access” (allows anonymous review by removing author names and other potentially citation-identifying information) → copy the created URL and share it with reviewers. When you want to disable the private URL, go to the dataset → click “Edit” → “Private URL” → “Disable Private URL” → “Yes, Disable Private URL”.

For information on curating data files in specific formats see:

Excel (.xlsx)	<ul style="list-style-type: none">• Excel CURATED checklist
Google Docs	<ul style="list-style-type: none">• Google Docs CURATED Checklist
R (.r, .rmd)	<ul style="list-style-type: none">• Filetype CURATED checklist

To assist in the management of draft datasets, the journal's Dataverse administrator or curator can assign labels for a dataset to indicate its status:

- Journal curation: pending journal curation / in progress.
- Author contacted: awaiting corrections from authors.
- Privacy Review: pending peer review / in progress.
- SciELO curation: pending SciELO curation / in progress.
- Awaiting article approval: awaiting approval of related article for dataset publication.

To add a label, click "Publish Dataset" → "Change Curation Status" and then choose the suitable status. The labels will automatically be removed from the dataset when it is published.

References

Abbott, D. What is Digital Curation? *Digital Curation Centre* [online]. [viewed 20 October 2021]. Available from: <https://www.dcc.ac.uk/guidance/briefing-papers/introduction-curation/what-digital-curation>.

CoreTrustSeal Standards and Certification Board. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022. *CoreTrustSeal* [online]. [viewed 20 October 2021]. Available from: <https://doi.org/10.5281/zenodo.3638211>.

CoreTrustSeal Standards and Certification Board. CoreTrustSeal Trustworthy Data Repositories Requirements: Glossary 2020–2022. *CoreTrustSeal* [online]. [viewed 20 October 2021]. Available from: <https://doi.org/10.5281/zenodo.3632563>.

DataverseNO. Curator Guide. *DataverseNO* [online]. [viewed 05 October 2021]. Available from: <https://site.uit.no/dataverseno/admin-en/curatorguide/>.

Lafferty-Hess, S., et al. Conceptualizing Data Curation Activities Within Two Academic Libraries. *Journal of Librarianship and Scholarly Communication* [online]. 2020, **8**, eP2347 [viewed 20 October 2021]. <https://doi.org/10.7710/2162-3309.2347>.

How to cite this document

SciELO. *Research data curation guidelines for editorial teams* [online]. SciELO, 2023 [cited **DD Month YYYY**]. Available from: _____.

This and other SciELO Data documents are available at:

<https://www.scielo.org/en/about-scielo/scielo-data-en/>