



Scientific Electronic Library Online

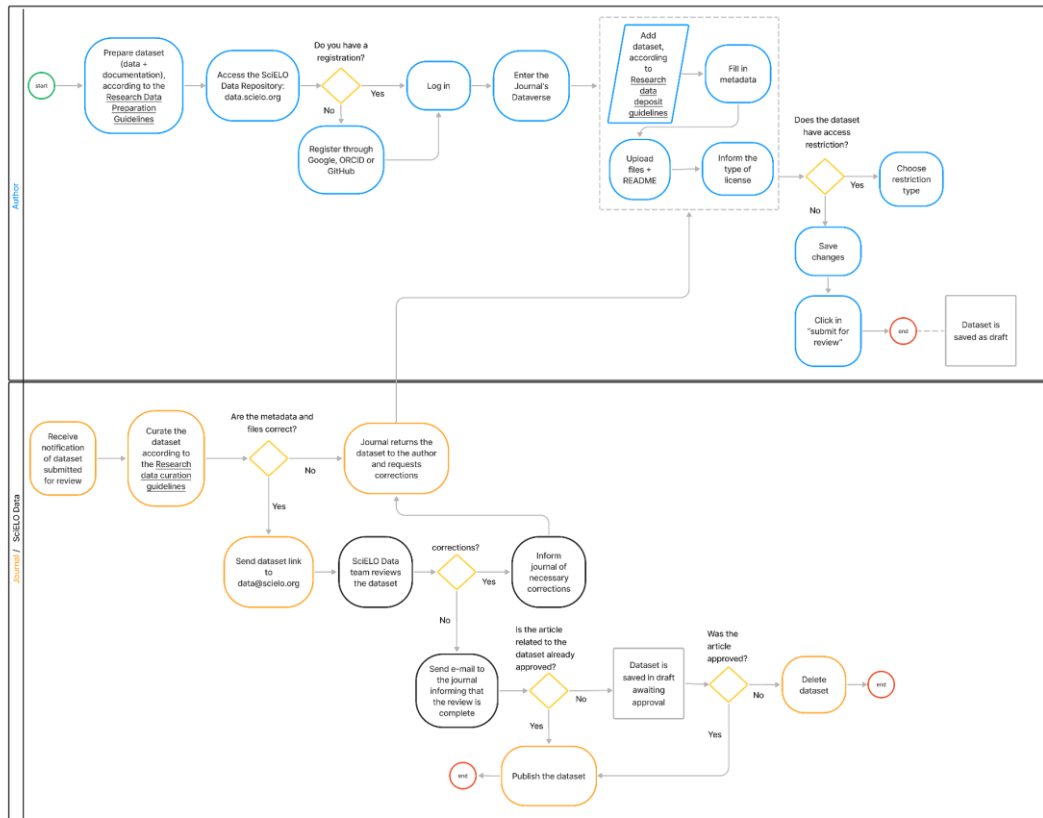
Research data curation guidelines for editorial teams

Junio 2022



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 CoreTrustSeal¹¹ (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. 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 used by CoreTrustSeal as a requirement in evaluating trusted data repositories:

- A. Content published as deposited → **Level 0**;
- B. Basic curation: quick metadata check / content, adding basic metadata or documentation → **Level 1**;
- C. Detailed curation: basic curation + conversion of data files to new formats, documentation enhancement → **Level 2**;
- D. Data-level curation: basic curation + enhanced curation + editing the deposited data for greater accuracy → **Level 3**.

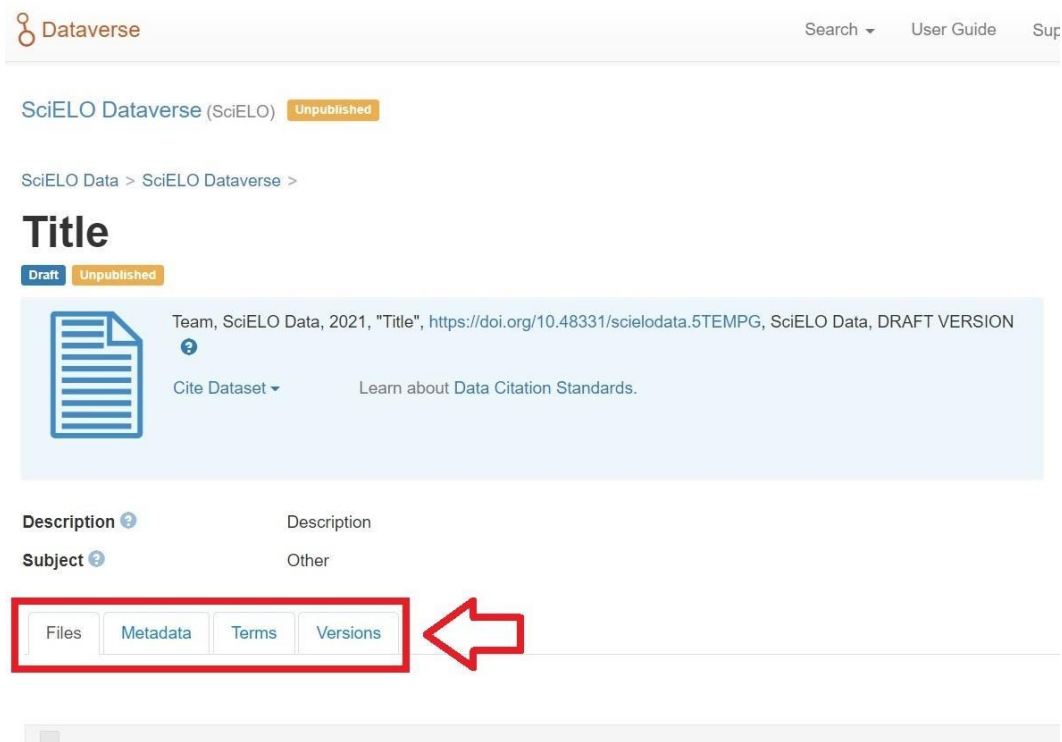
The level of curation adopted and carried out by the journal team must be explained in the journal's Dataverse description in SciELO Data and in their instructions to the authors.

2. 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) Unpublished'. The breadcrumb trail shows 'SciELO Data > SciELO Dataverse >'. The main heading is 'Title', with a 'Draft Unpublished' status. Below the heading, there is a document icon and a description: 'Team, SciELO Data, 2021, "Title", https://doi.org/10.48331/scielodata.5TEMPG, SciELO Data, DRAFT VERSION'. There is a 'Cite Dataset' button and a link to 'Learn about Data Citation Standards.'. Below the description, there are two tabs: '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

- Check if the dataset is related to any manuscript submitted to the journal.

On the “Versions” tab:

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

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;
- **Author:**
 - Check if the authors' names were entered in reverse order (Last Name, First Name);
 - Check if the authors informed their affiliation and ORCID (recommended).
- **Subject:** Check if the selected subject area is the most suitable;
- **Keyword:** Check if each keyword has been entered separately (each word in separate fields);
- **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**.

On the “Files” tab:

- Check if the dataset has been documented in a file named README;
- Check if the naming of files is adequate (see topic 1 of “[Research data preparation guidelines](#)”);
- Check if files can be opened (not corrupted);
- Check if the files are in the recommended formats (see topic 2 of “[Research data preparation guidelines](#)”);
- 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

On the “Terms” tab:

- Check if the chosen license type for the dataset was CC0 - "Public Domain Dedication".
 - If so, confirm with authors whether or not they wish to publish the dataset under this license;
 - If not, check if the “Terms of Use” field has been filled in with any license of use (see topic 7 of the [“Research data deposit guidelines”](#)).

Level 2 - Detailed curation → Before performing basic curation

On the “Metadata” tab:

- Evaluate the information provided to determine if it is complete and understandable.

On the “Files” tab:

- Download data files.
- Rename files in the most suitable way (see topic 1 of the [“Research data preparation guidelines”](#));
- Evaluate if files are in recommended formats (see topic 2 of the [“Research data preparation guidelines”](#));
- 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”](#)).

On the “Terms” tab:

- Evaluate the information provided to determine if it is complete and understandable.

Level 3 - Data-level curation → Before performing basic curation + enhanced curation

On the “Files” tab:

- Download data files;
- Open the data files and check if they need any additional treatment;
- Open the data files and check for sensitive data that needs to be de-identified / anonymized (personal identifiable information, protected health information, etc.);
- 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;
- Run and troubleshoot code files;
- Check consistency (checksum²) of dataset files to ensure data integrity at bit level;

² A checksum is a sequence of numbers and letters used to verify data integrity, that is, whether a file is exactly the same after a transfer, verify that it has not been altered by a third party or is not corrupted.

- If necessary, share the dataset with peer reviewers via a private URL³.

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 |

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

If the dataset is suitable for publication, the curator may publish it (click on “Publish Dataset” then “Publish”). The authors will receive an email informing them that the dataset has been published.

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

³ 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”.

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, 2021 [cited **DD Month YYYY**]. Available from: _____.