Research Data Curation Guide for Editorial Teams

May 2025
Version 1.1

1. What is data curation?

According to *CoreTrustSeal* [1](2019), data curation is:

> *Manage and promote the use of data from its inception to ensure that it is fit for contemporary purposes and available for discovery and reuse. For dynamic datasets this involves continuous enrichment or updates so that it continues to serve its intended purpose. Higher levels of curation will include links to annotations and other published materials.* (Our translation)

In other words, data curation involves the entire data life cycle, from planning for collection to preserving it for long-term access and reuse. It can also be defined more specifically as the checks and actions performed by curators to ensure that the dataset *is* structured and documented as completely as possible and in accordance with 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 *World Data System* (WDS) and the *Data Seal of Approval* (DSA). It is an international, community-based, non-governmental, non-profit organization that promotes trustworthy and sustainable data infrastructures.

1.1. Research data and the FAIR principles

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

It is with these aspects in mind that the curation of datasets in SciELO Data is mandatory and must be carried out by editorial teams following the instructions contained in this Guide, with the support of the Data Preparation Guide.

Publishing organized and well-documented datasets and metadata facilitates the access and reuse of research linked to them, causing a significant impact on the progress of science and scientific transparency [2].

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 level of FAIRness using the Fuji tool: https://f-uji.net/.

1.2. Data curation in SciELO Data

Curation in SciELO Data is the responsibility of each editorial team, which must manage its own repository, exclude invalid datasets (e.g., those related to articles rejected for publication or whose files do not contain data), request corrections from authors, edit metadata when necessary, and publish the datasets.

New journals in SciELO Data must **request curation support** from the SciELO Team after performing the curation as set out in this document. The editorial team, after performing the curation in accordance with this Guide, must send an email to data@scielo.org informing the URL of the dataset and requesting curation support, which is temporary in nature. This measure seeks to mitigate errors in datasets before publication.

The workflow with SciELO Data, as presented in the Flowchart in Annex 1, involves the following steps: initially, the author deposits the dataset following the guidelines in the Data Preparation Guide and Research Data Deposit Guide. The journal then receives notification that a new dataset is ready for review, curates the dataset following this guide, contacts the authors or makes edits if necessary, then requests curation support from the SciELO Data Team and, finally, publishes the dataset.

During this process, the editorial team may, if desired, share the dataset with reviewers for

[2]We recommend the webinar "How to make data FAIR? Good practices for investigative data" available at: https://www.youtube.com/watch?v=l14SwZxIRHY.

peer review through a private URL[3]. Creating a private URL allows sharing (for viewing and downloading files) of an unpublished dataset with people who do not necessarily have a SciELO Data user account.

If authors report that research data has already been published in another repository, do not deposit them in SciELO Data to avoid duplication. In cases where authors of an already published article wish to publish data related to the article, it is necessary to publish an Addendum together with the article, inserting the Data Availability section in the linked article. More information about this document is available in the Guide for registering, marking and publishing Addendums (Portuguese only).

If you have any questions regarding the workflow or data received, please contact the SciELO Data team by email at data@scielo.org.

2. Data Availability section in articles

Datasets published in SciELO Data must be linked to articles approved for publication in SciELO journals.

For this reason, in order to establish a connection between the published article and the underlying dataset, it is recommended that the instructions to authors document that the articles must contain the "Data availability" section, informing whether the dataset related to the research is available and, if so, where to access it.

Example content for the section:

Data Statement	Example text for the section
Data not available	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 has been made available on SciELO Data and can be accessed at [URL or DOI]. The entire anonymized dataset supporting the results of this study has been made available on SciELO Data and can be accessed at [URL or DOI]. 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).

[3]To create a private URL, go to the dataset and click on "Edit" → "Private URL" → in the pop-up box choose "Create a Private URL" or "Create URL for anonymous access" (allows anonymous review by removing author names and other potentially identifying information from citations and metadata filled in SciELO Data). Copy the created URL and share it with the reviewers. To disable the private URL, go to the dataset → click on "Edit" → "Private URL" → "Disable private URL" → "Yes, disable the private URL".

	The entire dataset supporting the results of this study was published in the article and in the "Supplementary Materials" section (Does not apply to articles with datasets in SciELO Data).
Data available upon request[4]	The full dataset supporting the findings of this study is available upon request to the corresponding author [name of corresponding author or organization holding the data]. The dataset is not publicly available due to [detail reason for restriction, e.g., containing information that compromises the privacy of research participants] (Does not apply to articles with datasets in SciELO Data).

Further examples of the Data Availability section and information about its markup are available in the Data Availability Markup and Publishing Guide.

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

3. Levels of data curation

Aiming at transparency regarding the verifications and actions carried out by curators on deposited data sets, SciELO adopts as a reference the curation levels used by *CoreTrustSeal* [5] as a requirement in the evaluation of reliable data repositories:

Level	Title	Description
Level 0	Content published as deposited	**Not** employed by SciELO
Level 1	Basic curation	Reviewing metadata and content, adding basic metadata or documentation, suggesting file naming and format changes to authors
Level 2	Detailed curation	Level 1 Curation + conversion of data files to new recommended formats for greater accessibility, improvement of documentation, making changes to file naming and format
Level 3	Data-level curation	Level 2 curation + editing of deposited data for greater accuracy

[4] In SciELO Data, it is possible to deposit datasets, but keep them restricted. When someone wants to access them, they will contact the author who deposited the dataset through the platform. See more details in item 5.1 of the Research Data Deposit Guide.
[5] https://zenodo.org/records/11476980

Regardless of the level of curation, it is **mandatory** to check whether the files contain personal or potentially sensitive data. If there is, it is essential to anonymize or pseudonymize them.

4. Data anonymization

The following must be anonymized: Personal data, whether sensitive or not [6], information that exceeds the right to privacy of the people 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, places or information that cannot be identified through aggregation as in the examples:

- 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 data anonymization[7]:

Information without anonymization	Response without anonymization
Name	Juan Perez
Country of origin	Argentina
Age	54

Anonymized information	Anonymous response
-	-
Continent	South America
Age Range	50-60

[6]Personal data: The following may 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; number of registration documents, such as ID, CPF and work card; geolocation data of a cell phone; personal telephone number. https://portal.fiocruz.br/noticia/entenda-melhor-lei-geral-de-protecao-de-dados-pessoais. Accessed on March 21, 2023.
Sensitive personal data: "personal data about racial or ethnic origin, religious beliefs, political opinions, membership of a trade union or organization of a religious, philosophical or political nature, data relating 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. Accessed on January 30, 2023.
[7]Example taken from: Research Data Management - Part I. Available at: https://www.youtube.com/watch?v=BM-lZ2XCCN0

Years of experience	15
Model airplanes	Boeing 777 Boeing 747
Last flight date	05/01/2022

Years of experience	10-20
Model airplanes	Commercials
Last flight date	01/2022

4.1 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;

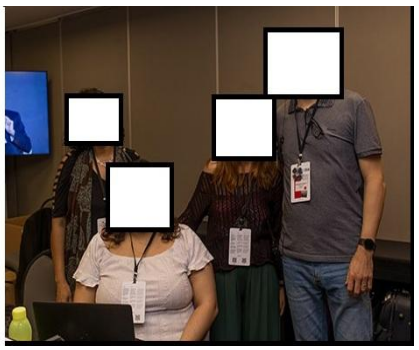
- Telephone number, document number, date of birth or address;
- Position held or specific work performed that could identify the person or institutional affiliation;
- Citations of other people's names (e.g., names of teachers or coworkers).

Avoid deleting or replacing items without indicating that an edit was made. Use aliases or replacements in square brackets to indicate where the edit was made, as in the following example:

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

4.2 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.):



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 publish only other files

in the dataset (such as the questions asked and analyses of the responses), as well as a README file or other documentation file for the dataset, explaining that not all files are available and the reason for preventing sharing in open access.

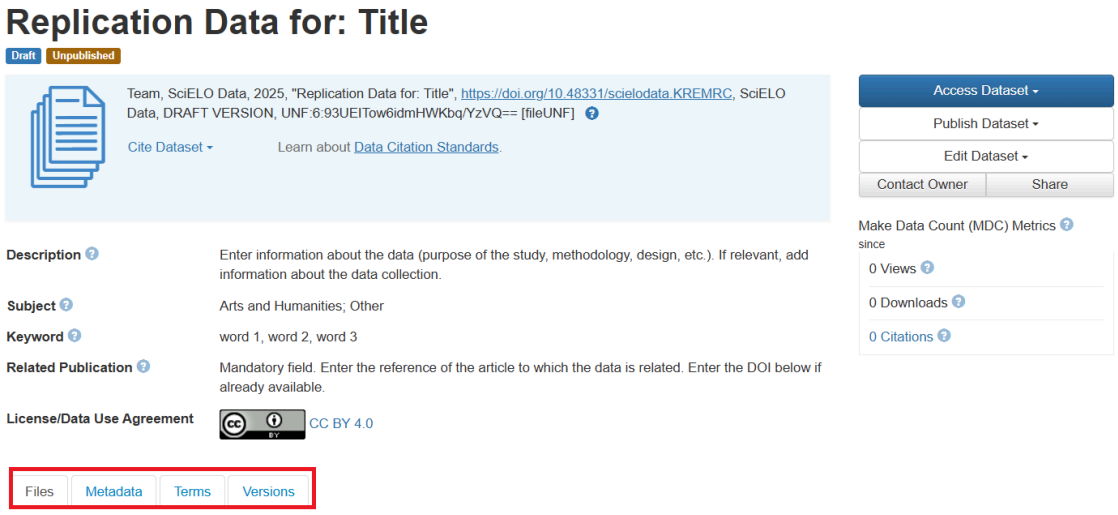
Some tools that can help with anonymization are: Amnesia from Openaire, anonymoUUs from Utrecht University and Text Anonymisation helper tool from UK Data Service.

5. Data curation verification and action checklists

In the SciELO Data repository there are four tabs in which metadata and files need to be curated, according to the level of curation adopted and carried out by the journal. The tabs are:

- Archives
- Metadata
- Terms
- Versions

You can navigate through these areas by clicking on the corresponding buttons:



To help manage draft datasets, users with administrator or curator permissions can assign tags to a dataset to indicate its current status:

- Journal curation: pending/ongoing journal curation;
- Author contacted: awaiting corrections from authors;
- Privacy Review: review by reviewer(s) pending/in progress;
- SciELO curation: SciELO curation pending/in progress. **The use of this tag does not replace the email that should be sent to SciELO Data to perform the support curation**.
- Awaiting article approval: awaiting approval of the related article for publication of the dataset.

To add a tag, click “Publish dataset” →“Change curation status” and then choose the appropriate status. Tags will be automatically removed from the dataset when it is published.

5.1. Checklist Level 1 - Basic Curation

Optionally, the Level 1 curation tutorial is [available as a video on Youtube] (Portuguese only).

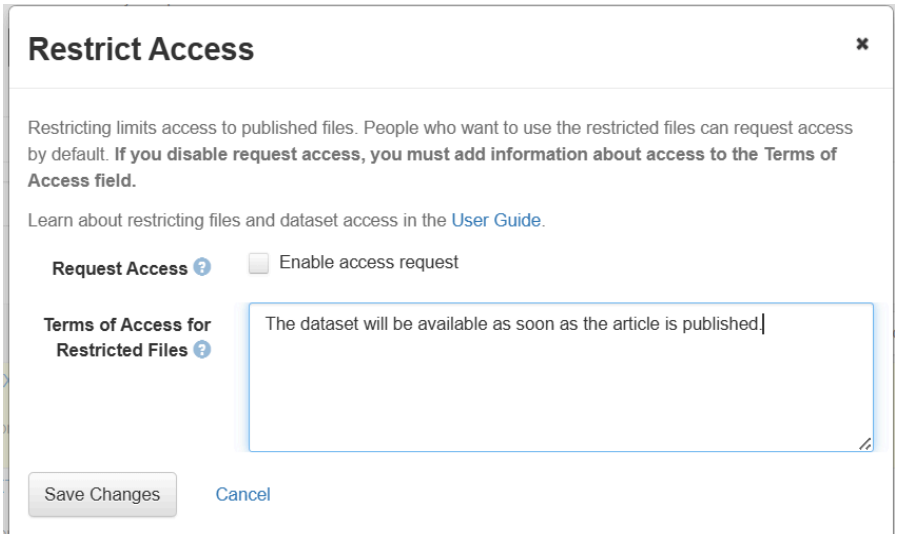
On the dataset home page:

- Make sure the dataset is related to a manuscript submitted to the journal.

Reminder: Only datasets from articles approved for publication by the journal can be published.

In the "Files" tab:

- Check if the dataset has been documented in a file named **README**. **The presence of this file is mandatory**;
- Check that the files that make up the set are not present in the article (figures, tables and charts that are already in the manuscript do not constitute research data for SciELO Data, as well as the dissertations and theses from which the article originates. For Supplementary Material, see the Annex 2.)
- Check if the file names are appropriate (see topic 1 of the "Research Data Preparation Guide"). If they are not, recommend that authors edit them following the recommendations;
- Check if the files can be opened (if they are not corrupted). If they do not open, request a new deposit from the authors;
- Check the files for data that needs to be anonymized. If anonymization is necessary, see item 4 of this guide;
- Check if the files are in recommended formats (see topic 2 of the "Research Data Preparation Guide"). If they are not, recommend that authors edit them following the recommendations;
- Check if the dataset has files with restricted access.
    - If so, the "Access Terms" field must be filled in with information about how users can gain access to restricted files.



In the “Metadata” tab:

- **Title**: check if it is filled in with the title of the article whose data is related or with its own title that is self-explanatory - so that the user does not need to open the files to know what it is about, and in accordance with the Data Preparation Guide. If it is not, recommend that the authors edit it;
- **Author**:
    - Check if the authors' names were entered in reverse order (Last Name, First Name);
    - Check if the authors have informed their affiliation (mandatory) and ORCID (recommended).
- **Subject**: check whether the selected subject area is the most appropriate. If you select “Other”, you must also select another subject;
- **Keyword**: check if each keyword has been entered separately (i.e. each in a field). If not, edit the metadata editing screen and add keywords by pressing the “+” sign;
- **Related post**:
    - If the dataset is related to a published article, insert the **article citation with DOI**.
    - If all the information is not available to enter a full citation, **enter it as completely as you can**.
- **Funding Information**: If the research has a funding source, click on “Edit” and then on “Metadata”. Find the field and fill it in.

In the “Terms” tab:

- Check if the license is CC BY 4.0. If you wish to use another license, please contact data@scielo.org.

In the “Versions” tab:

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

5.2. Level 2 Checklist - Detailed Curation

Perform Level 1 curation + In the “Files” tab:

- Rename the files as appropriate (see topic 1 of the “Research Data Preparation Guide”);
- Evaluate file formats to determine whether they are in a recommended format and convert them if necessary (see topic 2 of the “Research Data Preparation Guide”);
- Assess whether the documentation provided (README file, *codebook*, etc.) is complete and understandable (see topic 4 of the “Research Data Preparation Guide”). If it is not, request the necessary changes from the authors.

In the “Metadata” tab:

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

In the “Terms” tab:

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

5.3. Level 3 Curation - Data-level curation

Perform Level 1 + Level 2 curation + In the “Files” tab:

- Download data files;
- Open the data files and check if they require any additional processing. If necessary, request corrections from the authors or make the changes and inform them;
- Open the data files and evaluate them for potential issues such as: appropriate variable and value definitions, out-of-range values, program descriptions used for code files, and preferred data structures. If necessary, request corrections from the authors or make the changes and inform them;
- Run and troubleshoot code files;
- Perform consistency checks (*checksums* [8]) on the dataset files to ensure data integrity at the bit level.

6. Actions after curation

If the dataset is not properly structured and/or documented, the curator can return it to the author (click on “Publish dataset” → “Return to author”). Without this action, the author will not be able to edit the dataset.

Journals new to SciELO must curate the dataset and, before publishing, send an email to data@scielo.org informing the dataset URL and requesting curation support. During this temporary curation, the SciELO Data team will review the dataset to mitigate any errors before publication. After the SciELO team responds, if there are no corrections to be made, the editorial team must publish the dataset.

To publish the dataset, on the dataset page, click on “Publish dataset” → “Publish”. The author will receive an email informing them that the dataset has been published.

[8] *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, whether it has not been altered by third parties or whether it is not corrupted.

7. Edits to the dataset after publication

Datasets in SciELO Data are version-capable, meaning that corrections can be made after publication (such as editing metadata, replacing or adding files), resulting in a 2nd version of the set. Any updates, and who made the changes, are automatically recorded in the "Versions" tab. The DOI of the dataset is not changed.

However, once published, the **dataset cannot be deleted**. In exceptional cases where edits would not solve the problem (e.g., all the files in the dataset are invalid), it is not possible to "disappear" the dataset, only to "deactivate" it. If someone accesses the dataset's DOI, they will be taken to the same page, which will display the dataset's citation with the information "Deactivated Version" and the reason for its unavailability. If, after publication, a dataset requires changes of this nature, please contact data@scielo.org.

For information on curating file data in specific formats, see:

Excel (.xlsx)	• Excel CURATED checklist
Google Docs	• Google Docs CURATED Checklist
R (.r, .rmd)	• Filetype CURATED checklist

8. Updates to journal information pages

As journals begin to handle data, it is important for editorial staff to update the journal's policy on the deposit of research data in order to promote transparency in the process with authors and provide them with appropriate instructions.

Remember to inform the SciELO team:

1. When will the data be deposited (along with or after manuscript submission)?

2. Will data submission be encouraged or mandatory?

3. What level of curation will be adopted?

Below are suggested topics to be added to journal information pages:

- **Enter the journal's direct Dataverse link**: It is important that authors are clear about the journal repository address. Disclosing the link also allows other researchers to view previously published data;
- **Include definition of research data**: defining what research data is helps researchers identify and share only data that allows validation or replication of the results of the article to which the data are related;
- **Provide guidance on the availability policy**: clearly state whether making data available in data repositories is a recommendation (incentive) or a requirement (obligation);
- **Inclusion of the "Data Availability Statement" section/topic**: inform about the mandatory inclusion of the "Data Availability Statement" section or equivalent in all

articles and, if applicable, provide examples of text for the section. There are examples in item 2 of this guide and in the TOP Guide;

- **Recommend data repositories**: include a list of other repositories, in addition to SciELO Data, recommended by the journal, including community/discipline-specific data repositories and generalist data repositories. For repositories that follow best practices, we recommend visiting FAIRsharing and Re3Data;
- **Provide examples of how to cite data**: In the journal's reference examples, include research data as citable material. Examples of how to do this are available in the SciELO Research Data Citation Guide;
- **Include other standards adopted in the area** if necessary.

References

Abbott, D. What is Digital Curation? *Digital Curation Center* [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/.

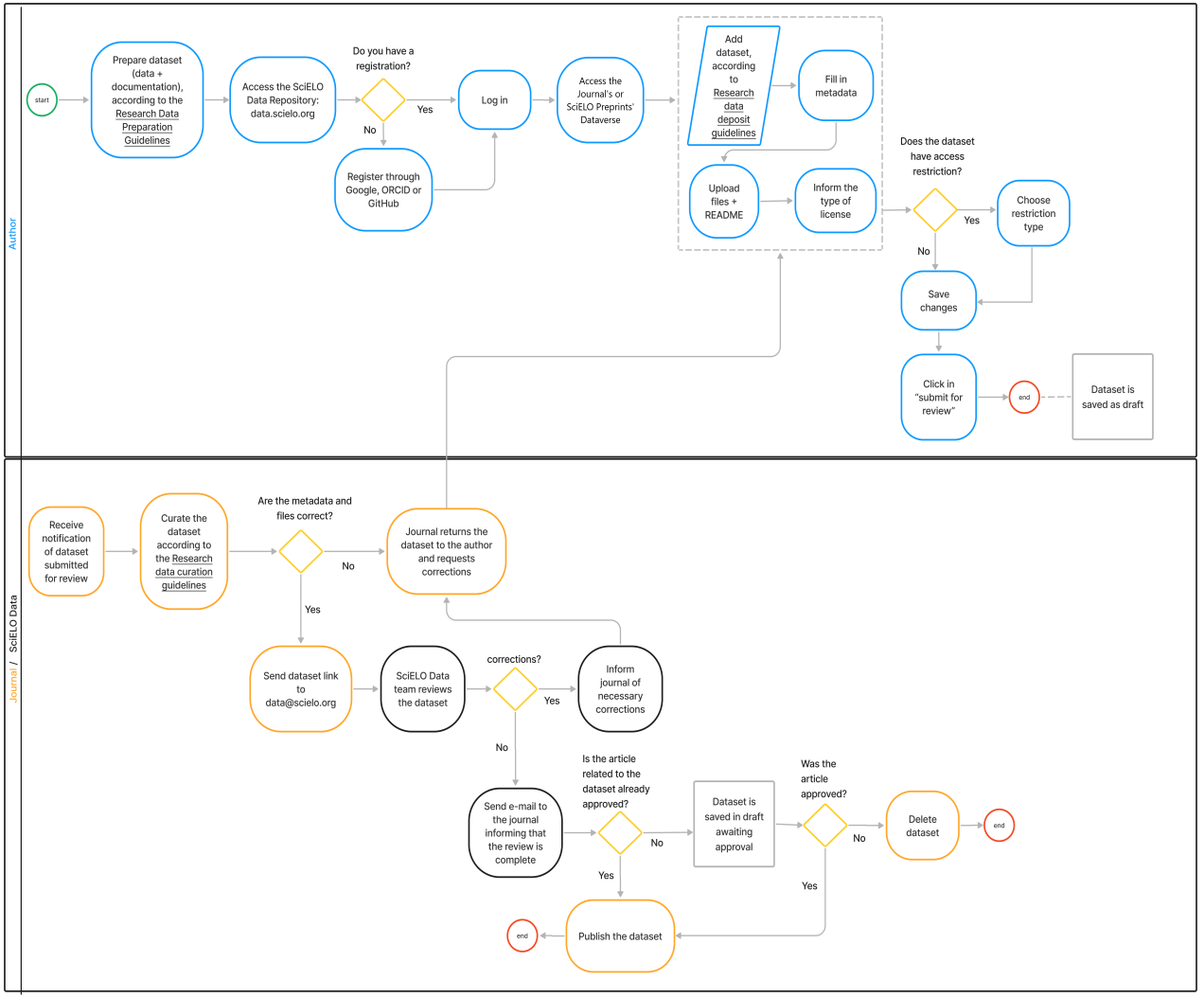
Lafferty-Hess, S., et al. Conceptualizing Data Curation Activities Within Two Academic Libraries. Journal of Librarianship and Scholarly Communication [online]. 2020, **8**, eP2347 [viewed 20 October 2021]. https://doi.org/10.7710/2162-3309.2347.

How to cite this document

SciELO. *Research data curation guide for editorial teams* [online]. SciELO, 2025 [cited DD Month YYYY]. Available from: _____.

This and other SciELO Data documents are available at:
https://www.scielo.org/pt/sobre-o-scielo/scielo-data-pt/

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. Supplementary Material

full-content materials must be deposited in SciELO Data.

Type	Description
Full Content	material **for understanding the work**, but it is allocated outside the article for technical reasons, such as: ● Voluminous material (such as a genome database) that supports the narrative's conclusions, but can never accompany a paper based on sheer mass; ● "Extra" tables that are not displayed with the paper, but that record the measurements on which the paper is based (e.g., tables that need to be available so that reviewers can check the content of the paper); ● Material added to the work for improvement purposes, such as a quiz, an instructional video, a form that can be filled out or copied, or similar material; ● A movie, MP3 file, or other binary material that is not directly part of the article content; and ● Figures that could not be included as part of the work due to stylistic considerations or space limitations.
Additional Content	Supplementary material that provides additional, relevant, and useful information to the article in the form of text, tables, figures, multimedia, or data, and that can help any reader achieve a deeper understanding of the current work through additional detail and context. Additional content **is not essential to understanding the work**.

Source: Adapted from: Recommended Practices for Online Supplemental Journal Article Materials - January 2013 and Journal Publishing Tag Library NISO JATS Version 1.3 (ANSI/NISO Z39.96-2021): <supplementary.material>.