DATA DOCUMENTATION AND CITATION CHECKLIST

Shelley Stall, Alison Specht, Margaret O'Brien, Jeaneth Machicao, Pedro-Luiz-Pizzigatti Corrêa, Romain David, Rorie Edmunds,, Nobuko Miyairi,, Yasuhiro Murayama,, Solange Santos,, et al.

Share your data –

Data Documentation and Citation Checklist

Welcome to "Your Open Science Journey"!

Plan, manage and share your data.
Use this checklist to improve your data management and sharing practices.

Target Audience

Primary: Team or project researcher

Secondary: Team or project lead

This checklist is generalized and will need to be adjusted based on your institution, lab, research team, and/or funder requirements.

A. Plan and Manage Your Data During Your Research Project

1. **Plan the Format and Organize Your Data Files:**

    a. **Use** *non-proprietary formats* (e.g., plain text tables) and compression file formats that do not lose information (e.g., TIFF) to ensure long-term access. Encrypted file formats can be problematic for future access. Open source alternatives for specialized software used in the analysis are also preferred.

    b. **Choose descriptive and useful file names** that both humans and machines can read. Examples of ideas to improve filename convention may include, 1) Using deliberate delimiters such as "-" to connect terms together and "_" to separate different information; 2) Avoiding special characters (e.g., $), spaces, and punctuation; 3) Creating meaningful names and metadata that explain the content; 4) Using methods that facilitate default ordering such as YYYY-MM-DD date format (an international standard). See Jenny Bryan's *How to name files.*

    c. **Organize your data using techniques that will result in high-quality, clean data**. For general information on data organization, see the EDI Repository's *Guidelines for cleaning data*.

    d. **Determine hierarchical folder and file structure for the project**. Be consistent, use one directory for one project, separate by function (e.g., data, code) and by output (e.g., figures, results). Ensure that original, raw data is separate, cannot be edited, and copies from the raw data are made for editing. Reference the compendium "file-structure" developed by the Turing Way.

    e. **Create one or more README files** to provide information about the data. Plan to capture the information as you are collecting it to help ensure that the data can be correctly interpreted at a later date. Include the basic elements that will also help you create metadata for a dataset.

        - Depending on your file structure and complexity you may want a README for each dataset as well as one to describe the entire file structure (also known as "file directory").

          Include these elements in your README file: Title, investigator(s) contact information, date(s) of collection, geographic information, descriptive

terms, language information, funding sources, sharing/access information, data and file overview, methodological information, and data-specific information. See *Guide to writing "readme" style metadata* from Cornell University which includes a downloadable template README file. Also, see the Smithsonian Libraries' *Describing Your Data: Data Dictionaries*. For general information on metadata, see the EDI Repository's *Creating metadata*.

Keep your README file updated.

2. **Manage Your Data More Efficiently with Versioning, Workflow, Notebook Tools:**
    a. **Use a version control system to keep track of file changes** or the history of a project (e.g., acquisition steps, reformatting). Platforms such as GitHub, GitLab, or Open Science Framework (OSF), provide additional collaborative and online backup capabilities.
    b. **Employ a workflow management tool** to create reproducible and scalable data analyses. Workflow languages such as the Common Workflow Language (CWL) offer a way to describe command-line tools and connect them together to create workflows. Similar options include Snakemake, Nextflow, Galaxy.
    c. For labs, **consider an electronic lab notebook** to document research, experiments, and procedures. Lab notebooks such as RSpace provide functionality to hook into research workflows and to document and share your research more efficiently.
    d. **Be Aware of and adhere to the usage license(s) and/or data request agreement(s) for data** previously created/collected by you or other persons.
3. **Store and Backup Your Data Files:**
    a. **Keep your files safe by ensuring they are automatically backed** up by a trusted system. Take advantage of your institution's backup services if offered. It is preferred that the backup system be in a different geographic location (e.g., cloud storage service).

B. SHARE YOUR DATA OPENLY

Use these steps when your research completes a milestone and before publication.

1. **Select Your Data Preservation Repository**
    a. **Planning** for where you will deposit your own created data will help guide you in structuring and managing it more efficiently for documenting and sharing later on. The Repository Guidance is helpful.
2. **Determine What Data to Preserve**
    a. **Data you collected or created (e.g., primary) and processed data** used for your research should be preserved and made available to allow other readers to assess your conclusions and build off your work.
    b. **For model or simulation data**, generally the model and configuration should be preserved. The EarthCube model community developed a framework to guide what to preserve from your research and is easy to adapt.
    c. **For very large data** (greater than 1 terabyte or TB) can be a challenge to preserve as there are often fees and additional resources required. One option to consider,

institutions often offer solutions for data preservation and compliance. Other options are provided in this blog post. [Internet Archive Link].

3. **Document Your Data Prior to depositing - README files, Metadata, and Licenses.**

    a. **Prepare your metadata. (recommended)**

    Many preservation repositories provide guidance for the type of documentation necessary to understand your data. (e.g. EDI Repository guidance) When this guidance is not available, follow the README file (above) guidance.

    b. **License your Data (required)**

    Select licensing permissions for the type of usage you envision for your own data. To encourage reuse by others, make your data as openly available as possible. Preferred permissible licenses are CC0 or CC-BY 4.0. See Creative Commons.

4. **Deposit Your Data with your Selected Repository**

    a. **Follow the guidance** from your selected repository.

    Include your README and/or Metadata files

    Identify your selected license. This is likely managed by the repository, but if not include information identifying your desired license.

    b. Ensure that the repository clearly displays your preferred citation.

C. Cite All Data.

1. **Include data citations** for the primary and processed research data you used (this includes your created data as well as any data created by others) in the References section of your paper. Doing so ensures proper credit is given for the data.
2. **Include a data availability statement** in your paper that describes where and how your data are available, including an online means to access your data. Check links and files before submitting your paper to the journal so as to ensure the data are accessible for paper peer review.
3. **Use the the availability statement and citation checklist** to ensure your citation and availability statements are complete. See Data and Software Sharing Guidance for Authors for more details.

Quick Links to related checklists

- Your Digital Presence
- Software Documentation and Citation Checklist

To recommend updates, please email datahelp@agu.org. Include the name and DOI for the checklist in your email.

To cite this checklist: Stall, Shelley, Specht, Alison, O'Brien, Margaret, Machicao, Jeaneth, Corrêa, Pedro Luiz Pizzigatti, David, Romain, Miyairi, Nobuko, Murayama, Yasuhiro, Santos, Solange, Wyborn, Lesley, Vellenich, Danton Ferreira, & Mabile, Laurence. (2022). Data Documentation and Citation Checklist. Zenodo. https://doi.org/10.5281/zenodo.7062402