Checklist: Research Data Management for Biology and Bioinformatics Projects

Overview: This document outlines some data management practices to consider during the lifecycle of the project - with microbiology and bioinformatics research projects in mind. Research data management is a continuous and cyclic process, and evolves throughout the project. Begin by reading through all the sections of the document to get a sense of things to consider before, during, and after your project. Then follow up with us to ask your questions and receive personalized feedback for each of these bullet points. Use this as a reminder to stay on track during your project, and/or as a checklist right before finalizing your project (e.g., thesis submission, publication).

How to make a copy of the document: Open the GoogleDoc, Go to File > Make a Copy. You will then be able to edit it and add your own notes. You can also save this document on your computer (File > Download). Go to https://bioinformatics.bact.wisc.edu/documentation/ for the most updated version.

Further readings and resources are available at the end of this document.

1. Planning:

General

- ☐ Understand the bigger context of the research and define your scope of contribution to the project program
- ☐ Review any grant data requirements (ask your PI if you don't know them), if applicable
- ☐ Define roles and responsibilities of different contributors as it relates to data or analyses

Project

- ☐ What is the research question?
- ☐ What is your sample design?
- ☐ What data will you be collecting?
- ☐ What are the data types (e.g. phenotypic, genomic, etc.)?
- ☐ Do you need to pull large amounts of publicly available data? (e.g., databases, or raw data for analysis)
- ☐ What bioinformatics methods or strategies will you be using to answer your question? How do they align with the data you are collecting?

2. Data Collection

Experimental design information

- ☐ Create data organization plan
    - ☐ Where will the data be saved (Box, Google Drive, OneDrive, ResearchDrive, etc.)?
    - ☐ How will the data be shared with others during the project?
    - ☐ How will you organize your project folders (subfolders, etc.)?

- [ ] Know your data:
    - [ ] Sample size, data file types (e.g. raw sequencing files, genome assemblies, tabular data, image files), data sizes
    - [ ] Save sequencing reports from sequencing providers (e.g. what machine was used to sequence, etc.)
- [ ] Review metadata collection standards in your field of study ahead of time
    - [ ] Review required vs. mandatory NCBI sample metadata template column names (e.g. GenBank, SRA, etc.)
    - [ ] Ensure you collect this data while you are doing your project
- [ ] Maintaining records: Use of an Electronic Lab Notebook (LabArchives; https://eln.wisc.edu/)? Physical lab notebook?
    - [ ] Have a backup plan - what happens if you lose your lab notebook?

Analyzing the data

- [ ] How will you keep track of code?
- [ ] Using Version Control (git)
- [ ] Learn how to use a private code repository (and how to make it public at the time of publication) (e.g., GitHub, GitLab)
    - [ ] Will you have it under your username or your lab organizational page?
- [ ] Using reproducible reports (e.g., Quarto Reports, Markdown files). Knowing when to use it for your own work and to communicate updates with collaborators during a project.
- [ ] NEVER enter research data into AI tools, especially not private information. If you used AI in your research, report transparently about it. Check UW-Madison AI academic integrity policy here: https://it.wisc.edu/ai/generative-ai-uw-madison-use-policies/
- [ ] If you are using a CHTC HTC's compute infrastructure, and use submit and .sh files for HTCondor, export/save/backup all your .sh, .sub, and .log, .err, and .out files.
- [ ] When using bioinformatics software and code, always use or keep track of
    - [ ] Portable software environments and versioning
    - [ ] Exact commands, parameters you used, and the order in which you run different commands
    - [ ] Use standardized file names and include a data dictionary
    - [ ] Using workflow managers (e.g., Snakemake, Nextflow)
- [ ] Use a "computer lab notebook" for digital work, in addition to your "wet lab" notebook
    - [ ] Keep all your notes in a specific location, and organize your notes
    - [ ] Timestamp your notes if possible (some software already includes that)
    - [ ] Note what you tried, your results, what worked, what didn't work
- [ ] Data Visualization:
    - [ ] Figures that are code-generated: All the R or Python code used to make the figures, including the raw data used; save images directly in the code vs. clicking "save as image".

- [ ] Beautifying the figures for presentations and publications: Esthetics such as font size are often improved in external programs like Illustrator. Make sure to clearly save what the figure looked like before and after for scientific transparency and integrity. Link to get Illustrator: https://kb.wisc.edu/helpdesk/62142

3. Data Storage

- [ ] Review data storage options: https://research.wisc.edu/data-storage/
- [ ] Use the tool Data Storage Finder storage.research.wisc.edu to find where to deposit your data depending on your role on campus, the classification of your data, and long- vs. short-term storage
- [ ] Is your data backed up? Where? How frequently? Who has access to it?
- [ ] If you need help moving large amounts of data, consider learning how to use Globus (https://it.wisc.edu/services/globus/)

4. Sharing your research

- [ ] Review this guide about data sharing essentials: https://researchdata.wisc.edu/data-access-2/
- [ ] Create and use your ORCID identifier in your publications and all your data products (https://orcid.org/)
    - [ ] Allows you to link your research products. Many publishers have a box to enter your ORCID identifier when you submit/publish a paper
    - [ ] Use your personal e-mail address because your UW-Madison email will eventually become inactive after you depart from the university.
- [ ] Begin early to upload your data to NCBI (if appropriate):
    - [ ] Tip: You can set the data to “private” or create a “reviewer-only link” to your data BEFORE your results are published.
    - [ ] Tip: You can set a publication date well into the future (years) or have it automatically public at the time of the publication
    - [ ] This will give you Accession IDs, which are often listed in papers and detailed in a Supplementary File of the paper.
    - [ ] If necessary, create a submission group (esp. if many people have data relevant to the project)
    - [ ] Bioproject (the umbrella term), Biosample (the physical samples), SRA (the sequencing reads), Genbank (the genetic sequence database), Genbank-Genome (assembled genomes)
- [ ] Make your code available on GitHub or GitLab, or as a zipped folder on Zenodo for long-term storage
    - [ ] Github vs. Gitlab comparison: https://kb.wisc.edu/shared-tools/109039 (Section: “How is GitLab different from Github”)

- ☐ Change from private repository to public, if applicable
- ☐ Option to transfer private ownership to lab organization ownership OR
- ☐ Make a fork of your personal repository to the lab repository, sync forks

☐ Other datasets can up uploaded to Dryad, and link it with your ORCID: https://researchdata.wisc.edu/data-repositories/

☐ Is your data **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable? Try this FAIR data assessment tool: https://fair-impact.eu/fair-assessment-tools or https://ardc.edu.au/resource/fair-data-self-assessment-tool/

☐ Ask yourself: if another scientist who had no relationship to your lab wanted to reproduce your experiment from start to finish, would they be able to with the information you made publicly available?

5. Get help

The checklist is written with a broad audience in mind, but slightly adapted to bioinformatics and biology needs. Reach out to us at (https://bioinformatics.bact.wisc.edu) to learn how this applies specifically to your projects, and how to do the things mentioned in the checklist.

6. Further reading and resources

Other checklists you might find useful:
https://researchdata.wisc.edu/resources/research-data-management-planning-checklist/