Data Management Plan

Purpose

It is important that you take care to organize and document the data you generate and any subsequent analysis. Organizing your data/analysis and providing documentation can seem boring and time-consuming, but it is critical. Without proper organization and documentation, your work is often useless, which is a far bigger waste of time and effort!

As a scientist or engineer, you never work in a vacuum, and this is certainly the case in our group. Others need to be able to understand and follow what you have done. Here are a few examples of important instances where you need to be able to interact with others:

- You are working in a team where people need to perform tasks informed by your data or do additional analysis directly on data you generated (e.g. fixing a code, making a figure for a poster or writing a paper).
- Scientists or engineers external to the University request to see your data (e.g. a collaborator wants to see it, or it is in a publication and is therefore public).
- You have left the University and someone is trying to understand what you have done and build on it.
- Your data has new uses that were unanticipated at the time it was taken.

Storing Data

In order to have consistency and uniformity in the way we store data, I ask that you store your data using the following standard practices. I describe these below based on situation.

Data Generation and Analysis – Daily Practices

- Graduate students
    - Your work machine (your desktop) should be backed up at least **daily** to the USB drive I have provided for you.
    - When using a personal computer for research, **you must regularly back up your data**. This should be done approximately **weekly**. You can back up your data on your work machine (desktop) or on the softmatter CAEDM group drive.
- Undergraduate students
    - When possible, you are encouraged to use a machine we have in the group (e.g. the windows laptop) that have backups.

    - When using a personal computer for research, **you must regularly back up your data**. This should be done approximately **monthly**. You can back up your data on your work machine (desktop) or on the softmatter CAEDM group drive.
- A few general comments about data generation and storage
    - You should backup and store all of the data, documents, talks, analysis that you generate on (i) a research desktop machine or (ii) on the softmatter CAEDM group directory. If you do not backup your data, it is not accessible to others, and it can easily be lost.
    - The exception to rule this is “garbage data”, i.e. simulations with bugs or errors that are not useful. Delete all garbage data. If you have a question about whether to keep the data or not, please ask.
    - A Key principle of data storage is that **all data should have at least two copies in two separate places.**

Data Storage Formats

- Data files should be named with descriptive names and dates.
    - For example, a good descriptive name of some simulations of a DPD block copolymer simulation might be: BCP_phase_diagram_sweep_fA.2022-01-19.
    - This example helps identify what is contained inside the folder! It gives a date, so you can find out what other simulations were done at the same time. It doesn’t have any spaces or weird characters (easy for scripts!).
- Create a “README.txt” file that goes with your data that gives some longer descriptions (1-3 sentences) of what is contained in the data. Text files are great because you can even read them from the command line.
- Avoid proprietary formats for file formats (e.g. Microsoft OneNote, Apple Keynote, etc.). MS Word (.doc, .docx) and MS Excel (.xls, .xlsx) are exceptions to this rule since they are ubiquitous. However, open source formats are even better (.ods, .odt), because we use Linux so much in our group. When appropriate, plain text files are preferred over all others.
- Separate raw data from analysis.
    - Raw data an analysis scripts are not often related one-to-one. In other words, some data has many different analysis scripts run on it. Some analysis scripts run on many bits of raw data. It helps if you store these in separate directories. Then use a README to help us understand how they are related!
- Please use the following naming conventions that follows the above rules:
    - For example, on [J-drive/groups/softmatter] create the following directories

```
XX.Lastname_Firstname.YYYY/
   proj_name/
      data/
```

```
        description_of_expt.YYYY-MM-DD
    analysis/
        description_of_analysis.YYYY-MM-DD
        data_paths.txt
    other_dirs_as_necessary/
proj2_name/
etc.
```

Key:

- XX = “Grad” if you are a graduate students, XX = “UG” if you are an undergrad student
- YYYY = year, e.g 2018; MM = month, e.g. 08 for August (with the zero); DD = day, e.g. 04 (for the fourth)
- proj_name = name of the project (e.g. 2D materials, polymer_nanoparticles, UGDT, training, etc)
- data = your raw data. Should contain a directory for every set of computations that you do, with a date for each one. *This directory should not contain your analysis.* For example, it should not contain lots of excel sheets and plots.
- analysis = all of your analysis should go here. Many different data sets may feed into the analysis. The paths to these data sets should be listed in a text file named data_paths.txt. All of these data sets must be on the softmatter group share.
- data_paths.txt = file that contains the location of the raw data contained in the analysis. These files must be located in data/ on the softmatter group share.

Writing a Paper or a Dissertation/Thesis

When you write a paper or a thesis, we need to store the document, the figures, the code, and the data. This helps other people use your work. You spend all this time doing the work! We need to keep it to make it useful! Most likely you will be gone when I get asked by another scientist for the data in the figure or when we decide we want to re-analyze some of the data that didn’t make it into the paper for use elsewhere. So, I need you to use the following format.

I want to create a single folder for all of the information from your paper, so I can easily back it up. All of the code, data, etc. should be contained within the paper directory. When you write a paper with me, I will create an git repository and an Overleaf document with the following file structure:

```
YYYY-Description_of_Paper/
    code/
    data/
    figures/
    manuscript/
    figures-old/
    notes/
```

revisions/

Key:

- code = a copy of the actual code you used to generate your data, e.g. pfpd or DPD-Rxn. The actual version you used for the paper!
- data = raw data for the paper. Organized in the same manner as your softmatter directory
- figures = directories containing the figures, figure scripts, and **the data for each figure**. This is not the raw data; this is the data points you plot in the figure. I should be able to reproduce each figure from what is in the file.
- manuscript = Latex and bibliography files for manuscript and supplemental info. Final versions of each figure. This is what is submitted to the journal
- figures-old = old versions of figures that don't end up in the paper.
- notes = other documents that go with the paper. Scanned notes of derivations, random stuff.
- revisions = copies of the manuscript at various important stages (submission, revision, final version, etc.) for easy of access. Also holds copies of reviews and correspondence with reviewers and editors.

- The git repository can't hold the raw data.
  - Git repositories are for things that are less than 100 MB. Your data is usually huge. Do not sync it to git. I have put data/ in the .gitignore for this reason.
- What do I do with the data for the paper while it is being written?
  - Put it in the [softmatter] CAEDM drive in a directory called: papers/YYYY-Description_of_Paper.
  - **You need to do this manually.**
  - You should do this while writing generating the data and writing the paper. Not at the end!
- When the paper is published we will back it up.
  - We will do this using a different CAEDM drive called [treearchive]
- How is this different for a thesis?
  - It is the same for a thesis. Follow the same process with me as if you are writing a paper. But instead, we won't submit it. Do the same thing with the text and the data.

Leaving the group

When you leave the group, you must make sure all of your data has been backed-up! Please follow the following practices to do so:

- Everyone

    - Make sure all of the data you have generated is **organized and documented** using the practices described here. Please document everything carefully!
    - Make sure that all data that is associated with a finished paper is backed up with the paper folder on the [treearchive] drive.
    - Move all data that is associated with an unfinished paper to the appropriate directory in [softmatter]/Papers/.
    - Delete all garbage data. Consult with other students and Dr. Tree when doing this if you have questions.
- Graduate students
    - Remove all of your data from ORC, and the [softmatter]/Grad.Your_Name/ directory. Move it all to your local work machine.
    - Organize the data your work machine and back up the home directory on your removable hard drive.
    - When you leave, I will save and label the hard drive, and reformat your computer for another graduate student.
- Undergraduate students
    - Remove all data from ORC and other sources. Put it all in your [softmatter]/UG.Your_Name/ directory. If you have done this right, there is probably not much here. Most likely your data should go with a paper, or it is garbage data. There may be exceptions (e.g. if you had a code development project).

Summary of Data Backup Locations

Drive	Contains	Timescale	Comments
Local USB Hard Drives	Complete backup of your local machine.	Daily	Every machine should have one.
CAEDM group: softmatter	G.*.YYYY/ UG.*.YYYY/ Papers/	Weekly/Monthly	(1) Primary data storage for undergrads (2) Secondary data storage for grad students. (3) Main place to share large data sets for the group. (4) Data repositories for papers in progress.
CAEDM group: treearchive, treearchive2	Long term archives: code, data, papers, podcast, proposals, talks, teaching	Several times per year (after papers, theses, etc. are published)	Primary copy of long term backups of all the important data, papers, and teaching documents. Copy important material from *softmatter* group.
CAEDM group: treedata	Tree backups		Backups of various Tree computers: PhD, Postdoc, BYU computers
Gaucho: /backup/	(1) Long term archives: code, data, papers, podcast, proposals, talks, teaching (2) Tree backups	At least yearly	Mirror of *treearchive* and *treedata* group. Secondary copy of long-term data.

Last updated: 01/19/22