**Lessons from Data Acquisition and Management during my PhD**

**Abstract**

Through a combination of mentorship and self-exploration, I have found a variety of ways to manage and store my data. However, having to find and learn many of these options on my own has resulted in my data being scattered and stored in many ways, sometimes making it difficult to know where a particular set of data is found. Passing my data onto the next generation of students will be more difficult because of how many resources I had to discover and learn about. Ideally a list of good data management tools and practices would be a welcome introduction to the graduate school curriculum. By creating a list of potential tools and practices to begin with, students will be empowered to find a data management style that works for them early in their careers.

**Importance of Data Management**

By creating a data management plan (DMP), it will be much easier to share, distribute, and organize data for a variety of different projects. However, determining the DMP that works for a student early in their career can be difficult without any guidance. The University of Wisconsin-Madison offers the use of DMPTool, an online template for creating a plan to organize and manage data1. In addition to creating a DMP, there are many tools for developing good data management practices including creating PowerPoint presentations, utilizing an online notebook, or data dictionaries. Working with good practices throughout the data collection process will make finding data for distribution or analysis easier in the long run. Whether it be running experiments or reading literature, there are many resources available to begin good data management.

With so many options at our disposal, a variety of research groups have taken to compiling potential good data practices. Goodman et al. compiled a variety of sources as well as rules for good data practices2. Borghi et al. wrote a paper to serve as a data management guide for researchers3. Inspire by these studies, I’ve decided to create a short resource repository composed of multiple methods to organize and manage data that I have learned throughout my research career. All these methods are not required but utilizing a few of these can aid in successfully organizing data for sharing and this in the future. Each of these resources can be helpful for both current and future researchers. This list of resources is something that I feel would have helped empower me to begin my data management journey early in graduate school. I believe a clear data management plan as well as a list of tools at a student’s disposal will increase the likelihood of good data management practices for future reproducibility.

**Developing better data management practices**

**Figure 1: Resource List**

Planning

* [Individual Development Plan](https://kb.wisc.edu/grad/116772): Way to annually assess skills and interests over time to determine progress towards academic and professional development goals
* [Data Management Plan](https://dmptool.org/plans): Tool to develop a plan for how data will be organized and distributed during the research process

Online Data Management

* [Labarchives](https://mynotebook.labarchives.com/): Lab notebook where you can create pages and folders to separate projects as well as attach data, figures, or other important files related to a project
* [GitHub](https://github.com/): Code and data repository that allows you to track the history of your data over time
* [SourceForge](https://sourceforge.net/): Open-source software management that allows you to host code and data for your projects

Citation Managers

The three citation managers found that are supported by [UW-Madison Libraries](https://www.library.wisc.edu/research-support/collecting-organizing-analyzing-information/citation-managers/). Each of these allows you to keep papers in an organized format, and to easily create citations for them in a variety of different formats.

* [Zotero](https://www.zotero.org/)
* [EndNote](https://endnote.com/)
* [Mendeley](https://www.mendeley.com/)

Along with lessons I’ve learned from data management below, this essay also focuses on describing a potential future method to normalize data management by focusing on three key sources: raw data, analyzed data, and interpreted data. This proposed method will focus on consolidation of input datasets and summarizing given information into simple figures that can be used in PowerPoint presentations or transitioned to a professional figure for a paper. By developing this method, anyone who needs to present data in any format will be able to better organize their data for the future.

Data lesson 1

Defining a naming scheme helps keep datasets organized.

*Management implication 1*

By defining a consistent and searchable naming scheme for datasets, we are better able to keep data organized and more easily find them. For example, naming a dataset of wild type fluorescence experiments as 2023-4-30\_wt\_fluorescence will allow you to search by three key words: date, experiment type, and experiment output. In following a naming scheme, we are more easily able to find the data that we’re interested in.

Data lesson 2

Separating data by project is intuitive and easy to follow.

*Management implication 2*

By separating the raw, analyzed, and interpreted data in these folders, we will easily be able to find data related to a particular project. This will mean we can find all the data for a particular experiment related to that project with a quick and simple search. Additionally, writing a README file for each project or each folder can be used to quickly describe what is found in the folder. This can be quickly referenced to determine how the data is related to the project, or how the data is analyzed4.

Data lesson 3

Coding is a powerful tool for data analysis of organized data.

*Management implication 3*

Using a coding language such as python, it is possible to create scripts that navigate to a folder full of data and run analysis on it. With well-organized folders, these scripts can find any files with similar names and analyze them. Ideally, this analysis would be able to create a new file that can be used in a variety of formats to be used in PowerPoint presentations, posters, or adapted for papers. This script will work similarly to the Designer function in PowerPoint, inserting a figure into a variety of different formats and allowing you to add any text or other figures you might need in another format (Figure 2).

**Figure 2**

A picture containing screenshot, text, multimedia software, software

Description automatically generated

Data lesson 4

Organization of computational data allows for strong reproducibility and ease of analysis.

*Management Implication 4*

Reproducibility of research in fields of research that deal with month long experiments and troubleshooting is difficult. Surprisingly, even in the field of computer science reproducibility of research has been found to be difficult as well. To combat this, there is a reproducibility review system where articles can be submitted to be reproduced, demonstrating that a reviewer was able to reproduce the results from the code and data5. Not everyone will have code or analyzed data by code, but by keeping methods organized and written well research can be more reproducible. Organizing data and code that is used for statistical analysis allows for increased reproducibility. Publishing this code and data in a supplemental format on repositories such as GitHub or SourceForge allows other studies to replicate findings, or to replicate certain figures for similar data, normalizing how pieces of data and specific statistics used within the scientific community.

**Perspective**

As a graduate student, I had to find and teach myself a variety of methods for good data management practices. It was difficult to find consistent ways to organize and manage my data early on in my graduate career. To combat this issue, here I have compiled a list of resources that I believe would have been helpful for myself as I began graduate school. Subsequently I’ve also proposed the development of a data focused pipeline to analyze a normalized set of organized data. When developed, it will be able to simplify data management for researchers. Additionally, the analysis pipeline will aid students in developing figures for presentations and papers that can be edited for personalization. With better guidance for students who are unaware of how to manage data, it will be easier to distribute data long term, increasing the reproducibility of research.

Citations:

1. How to Create a DMP. *Research Data Services* <https://researchdata.wisc.edu/how-to-create-a-dmp/>.
2. Goodman, A. *et al.* Ten Simple Rules for the Care and Feeding of Scientific Data. *PLOS Computational Biology* **10**, e1003542 (2014).
3. Borghi, J., Abrams, S., Lowenberg, D., Simms, S. & Chodacki, J. Support Your Data: A Research Data Management Guide for Researchers. *Research Ideas and Outcomes* **4**, e26439 (2018)
4. Guide to writing ‘readme’ style metadata | Research Data Management Service Group. <https://data.research.cornell.edu/content/readme>.
5. Peng, R. D. Reproducible Research in Computational Science. *Science* **334**, 1226–1227 (2011).