Beronda Montgomery's *From Deficits to Possibilities* essay: <https://career.grinnell.edu/blog/2022/08/26/from-deficits-to-possibilities-mentoring-lessons-from-plants-on-cultivating-individual-growth-through-environmental-assessment-and-optimization/>

Prepare an essay, based on your own area of study, that illuminates one (or more) of the responsible conduct of research topics we have studied over the semester, in the same vein as Beronda Montgomery's *From Deficits to Possibilities* essay, which was expanded to create her book, *Lessons from Plants*. The essay should be between about 1000-2000 words. Feel free to include figures or illustrations for interest and clarity. The thesis of the essay should represent an interesting approach to the responsible conduct of research, and should illustrate an effective way to communicate or teach about that approach based on some aspect of your area of study. You should give examples that support your thesis, and your writing should be clear and concise.

Essay Topic: Week 7: Data Acquisition and Management

Writing on this topic should in some way be about my failure to truly follow good data practices because we didn’t really learn them. So thinking of a way to make it easy to teach about good data practices, AND a lot of the fun for it could be put into ideas that I have about it in the future for a potential future lab, but more realistically and closer to now:

my basketball stats website:

* What are tools that will help me stay organized when making it?
* What are the things that I feel like I’ve made mistakes on in the past?
* How can I make these things easier in the future?
* What are some punny/nice sounding titles for this idea?
* Good figures for it?
  + I think I could make a simple flowchart for ways that scientists can save their data:
    - What if it was possible to make a processing software that could instantly transform that data into a nice figure? Kind of like the way powerpoint when you put an image in, it like gets creative in placing it and shit? Like what if you could have a powerpoint like system that will take any input figures or data, and descriptions of the data, and come up with a nice visual way of seeing that data? Like something that essentially creates paper like figures for you: If you have three pieces of data in figure form and their corresponding legends, it will automatically have templates setup for how those figures could look on a paper/powerpoint and essentially double as a way of keeping notes for your project?
    - The Methods about how code would work could be interesting if I can think of clever titles for how to do it/how it overall works? Or maybe if I just have clever titles for the above idea, then that’s probably good enough without going to in depth into how the methods work (make sure that the essay doesn’t have that in depth detail into the methodology)

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

**Abstract**

**Gathering Data**

During my graduate career, I have been exposed to a variety of methods for saving and managing 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. Although I have found a variety of ways to manage and store my data, passing it on to the next generation of students has proven difficult because of how many resources they will have to search through if they desire to find my data.

**Importance of Data storage and 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 data management plan that works for every student can be difficult. There are many tools for developing good data management practices including creating PowerPoint presentations, utilizing an online notebook, data dictionaries, or developing a personal organization scheme for one’s own data. 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.

Because there are so many options, I’ve created a short resource repository composed of multiple methods to organize and manage data that I have learned of 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. Ideally, having a list of resources that focuses on proper data collection and storage would aid in developing good data management practices.

Additionally, this essay focuses on describing a potential future method to centralize 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.

**Resources**

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/)

**Data Storage and management**

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 as a way 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 analyzed (cite 2).

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 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 (Figure). (say something about similarity to the how PowerPoint does with figures; describe how the pipeline works with a flowchart and maybe some supplementary figures)

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 data. Not everyone will have code and 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**

Students are exposed to a variety of different ways to organize schedules and lifestyles introduced to by teachers with different philosophies. It’s difficult to find consistent ways to organize our livelihoods and to find ideal ways that work for us. Here I have compiled a list of resources that I believe would have been helpful for students as they begin 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.

Citations:

1. Disciplinary Metadata | DCC. <https://www.dcc.ac.uk/guidance/standards/metadata>.

2. Guide to writing ‘readme’ style metadata | Research Data Management Service Group. <https://data.research.cornell.edu/content/readme>.

3. Hodson, S. Managing and sharing data: best practice for researchers.

4. Reproducible Research in Computational Science. <https://www.science.org/doi/10.1126/science.1213847> doi:[10.1126/science.1213847](https://doi.org/10.1126/science.1213847).

5. 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).

6. Goodman, A. *et al.* Ten Simple Rules for the Care and Feeding of Scientific Data. *PLOS Computational Biology* **10**, e1003542 (2014). (has a lot of rules for data; could be good to compare/cite; also has it’s own list of data sources generally for sharing, but not for personal data management)

Paper with examples: https://riojournal.com/articles.php?id=26439

README guide: <https://data.research.cornell.edu/content/readme>

Figure of reproducibility spectrum: https://www.science.org/doi/10.1126/science.1213847