A Graduate Student's Road Map for Data Management Training

Benjamin Roberts-Pierel[1], Eleanor Davis Pierel[2], Yuhan Rao[3]

1. Oregon State University; 2. University of South Carolina; 3. North Carolina State University

** All authors were Earth Science Information Partners (ESIP) Community Fellows for 2018-2020.*

1. Mapping Data Lifecycle into Graduate School Road Map

- Plan - Coursework/Proposal writing
- Collect - Data collection
- Assure - Data collection
- Describe - Data collection
- Preserve - Defense & Publication of manuscripts
- Discover - Coursework/Proposal writing
- Integrate - Data analysis & Dissertation/Thesis Writing
- Analyze - Data analysis & Dissertation/Thesis Writing

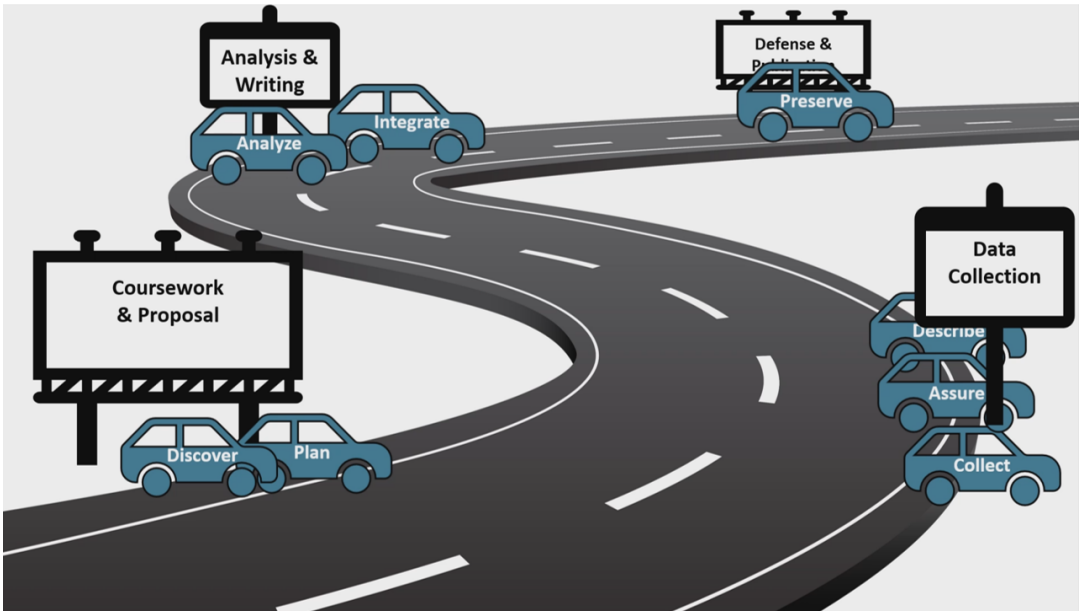

Figure 1. Mapping data lifecycle (DataONE) to different graduate school stages.

2. Your Graduate School Data Road Map

Congratulations on starting graduate school! It can often feel like a marathon and data

management might not seem like a priority but we promise (as current and past students ourselves) it is worth being familiar with the basics.

We have compiled a dynamic guide to help you wade through the ocean of data resources. This guide will help identify when and what you should be considering at different graduate school milestones. Just click on a relevant milestone and you'll find videos, tutorials, and other resources to help you and your research needs.

2.1 Program Orientation

Many programs offer an introduction to the requirements and stages of your specific degree. This is also a great time to become familiar with an introduction to data management.

Why is data management important?

Whether you are using existing data sources or creating your own primary data, having a plan for how you will collect, label, store, and maybe even publish your data is critical. Sure, you can wing it, we all have done this. However, future you, sorting through data at 2 am to find a specific file, will thank you deeply for thinking about data management ahead of time.

Watch this short video to get an Introduction to Data Management.

2.2 Coursework

As you are taking classes, writing papers, and living the life of a grad student, this is a great time to get familiar with the resources on campus.

It might sound basic but your university's librarians are there to help. They can provide direction on everything from information sources to data repositories.

By befriending your school's librarians (we have it on good authority they appreciate cookies), you will save yourself time, energy, and a smacked forehead when you finally find the perfect article for your proposal!

Many of your universities will have a person(s) devoted to data management and the data lifecycle (e.g. Oregon State University), reach out to them! They are very likely excited to talk to you about your work and answer questions about any stage of the data lifecycle.

2.3 Proposal writing & defense

At this stage in graduate school, we spend a lot of time thinking about the future. Perhaps you are asking yourself, is this a good research question? Is this even a

possible methodology? How can I do my analysis if I'm not even sure this will work?

We've been there. This can be a long phase that can feel like you are moving in circles but all of this planning really does pay off.

What does this phase have to do with data management?

Well, planning is a critical step because it is when you decide what data you need and how you will collect it. It is also when you might describe how you will store it and analyze it. These plans may change, but Future You will thank Past You when you start analysis maybe a year or more after you collected your data.

Why should you create a data management plan?

There are a few reasons such as saving you time, making Future You happy, and possibly helping you professionally, if other scientists can use your data and cite you. This is also the phase where you should consider whether you will work with human subjects and need to interact with the IRB. Not convinced? Watch this 10 minute persuasive video.

What are the steps to include in your data management plan?

Watch this 5 minute video on the elements to include and a few questions to ask yourself as you write.

Ready to start writing? Here are example templates to get you started.

2.4 Data collection

Wooooo! You finally get to go into the field or talk to people or download terabytes of data or all of the above! This is a really exciting (and maybe slightly terrifying) time in grad school. **You get to make research happen**.

This stage can also be expensive if you are collecting and storing lots of data. Often, you only have a handful of chances to collect the data that will make or break your research.

How can you save yourself heartache when you go to analyze your data?

1. Stick to your data management plan and backup your data
2. Collect metadata - Metadata is the "Data about data" and will help you (or someone using your data in the future) find, evaluate, and understand your data. It's all of the questions you would need to ask to be able to use your data in analysis. Watch this 10 Minute Video on Metadata.
3. Do quality control as you are collecting and inputting your data. Ask yourself, Does that number look right? Is something missing? Check out this handout on

data quality control.

4. Organize your data. This will make it more easily findable and usable in the future. Here is a one page explanation of the value of data organization and a worksheet to help you remember your naming strategy
5. Think about how you are going to structure your data for storage and analysis. It's tempting to append words to that file name until it's two lines long but come up with a convention and stick to it. You can check out these very simple best practices
6. Decide on a directory structure to accompany your naming convention (think folders!). You don't want to spend all of your time just looking for your data before you can start doing analysis
7. Can you employ the FAIR principles when storing (and eventually publishing) your data? Will your data be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable? Not all of these principles are possible for every project but they are goals that we can keep in mind for future projects.

You might be thinking, "Oh wow you just made my favorite part of research really, really stressful." Trust us. It's worth it.

2.5 Data analysis & Dissertation/Thesis Writing

We all know the analyzing part is the most exciting part of graduate school. Once you have collected data (correctly), it's time to have fun analyzing your data.

But there are things you need to pay attention to during data analysis. To name two on the top of the list:

- Ensure **reproducibility** - Can you or other people who are interested in your research reproduce the same results or figures in a month, a year, or a decade?
- Ensure **efficiency** - Are you analyzing the data in the most efficient way?

These two points are often overlooked by students since we typically want to get the results quickly and move forward. Establishing a habit of reproducibility and efficiency at the beginning of data analysis and dissertation writing stage.

There are many ways to establish the habit, such as:

- attending training workshops and conferences (in person or virtual);
- taking accredited courses from universities or online platforms;
- fostering open science culture within the research group.

Here are some existing resources that are a great starting point for establishing the habits.

- The Carpentries provides a rich collection of training for computer programing virtually with local certified instructors;
- Openscapes provides a learning community on open science in environmental sciences;
- rOpenSci for R users & pyOpenSci for python users;

2.5 Defense & Publication of manuscript(s)

Congratulations! You can see the light at the end of the tunnel! You might be close to finished but your data are still able to make a world of difference. There are just a few more steps on the data roadmap to maximize your thesis or dissertation's long-term potential.

To make sure your data live on even after you have moved on, consider these questions:

1. *Where are your data going to live for the long-term?* Many times data should or must be published with manuscripts. You can check here for possible places to archive.
2. *How your data might be used in the future?* Will your data set be essential/popular for other research? Or is your data set targeted for a specific user group? Your answer to these questions should impact where and how the data are stored.
3. *Are there any security or privacy issues with sharing your data?* If you collected sensitive data, such as human subject data, you may need to work with your Institutional Review Board or the data repository to ensure you are protecting the data properly.

2.7 Graduation!

Now it's time to go put your superb data management skills to work in the real world! Congratulations!