Data management – actually easier than herding cats

By Rosemary Hartman | September 22, 2014 |

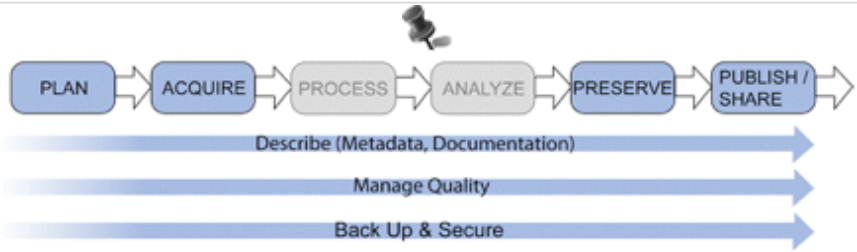
It's a new year, new grant cycle, time to think about everyone's favorite topic – Data management! OK, it's probably not the most fun thing you will ever do. If you really love databases, Endnote libraries, data dictionaries, and quality assurance plans, you will have great job security later in life. However, if you are like me you started grad school with a vague proficiency in Excel and some idea that you should take good lab notes. Managing data in any organized fashion was never as fun or interesting as collecting it. I was sure to remember what I did, why I did it, and where I put it, right?

You can get away with sloppy data management with a small project, but you should start good habits early. Even that pilot study needs a little help from the data organization cat!

No. Not really. It is amazing how difficult it is to remember why I didn't catch any mice at Big Marshy Lake, or what LITTER stood for when I came up with that acronym. If you want to produce useful data that allows you to write scientifically defensible publications and share your data with others, you need a data management plan.

Not convinced? Well, more to the point, many funding agencies now require a data management section in grant proposals where you describe your plans to efficiently collect, QC, analyze, store, and share your data.

So what is data management? I recently attended two sessions at ESA in Sacramento on tools and tips for managing ecological data. The organizers were from DataONE, (Data Observation Network for Earth) which is an organization that searches data stored across many different data repositories and organizes multiple tools and resources for managing and analyzing data. They broke down the process of data management and introduced tools to make it easier. I have attempted to compile their work into guidelines for your own data/cat management plans.



The data management life cycle, from USGS's data management website.

Step 1: Plan

Plan your data collection and management before you start. You will undoubtedly make changes as you go, but starting with a plan will prevent avoidable mistakes.

Developing your data management plan is hard work. It should be done in a very comfortable chair.

- Start by looking at the whole data life cycle, you want to plan every step before you collect your first data point. Check out https://dmptool.org/dashboard , it's like TurboTax for data management plans. They even include specific requirements that funding agencies require in management plans for their grant proposals.
- What protocols will you use? Your data should be comparable with others, so try to use established protocols and methods when you can instead of developing your own from scratch.
- What format will you use to record the data? Will you use paper data sheets or direct computer entry?
- Who will collect your data? Training interns/kittens before you start to make sure everyone is interpreting and recordings things the same way will help avoid errors. Training them in good data management skills is also critical (I

wish I had told my kittens to always write the date on their data sheets… I have no memory of when we collected those hairballs at the neighbors catnip patch).

- Start metadata early! Make a "data dictionary" of common terms and abbreviations that you can refer to in the field.
- It is important to think how others might use the data, so include anything that might seem redundant or unnecessary to record for a single study, but puts the data in context with other studies of its type.
- Where will the data be stored? Decide on a repository and whether you will have to put any restrictions on its use (see Preservation section below).

Step 2: Acquire

- Record any deviations from your protocols, no matter how small they seem at the time.
- Always include a "comment" field in your data tables to note any mistakes or changes to established protocols. Use the comment field instead of Microsoft comment bubbles or highlights to flag problems – the highlights will not easily convert between file types!
- Keep data and metadata linked, but only have one table per data sheet to allow easier analysis.
- You may be using data collected by others instead of collecting your own, in which case you have to find it.
    - You can search for data across the world here: https://cn.dataone.org/onemercury/
    - For federal and state data sets, try: http://catalog.data.gov/dataset
    - For California, try some of these:
    - http://ceden.waterboards.ca.gov/AdvancedQueryTool.php
    - http://www.dfg.ca.gov/biogeodata/bios/
    - http://www.ecoatlas.org/
    - http://cdec.water.ca.gov/

Along the way: Manage quality

(also known as Quality Assurance/Quality Control, or QA/QC)

This cat makes a fatal mistake in quality control. Never shred your mistakes. Fix them or omit them from analysis, but keep a record of the original version.

- QA/QC begins in the planning stage and follows you throughout the data life cycle.
- Do most of the work before data are collected:
    - Assign responsibility for quality control to one cat in the lab/field
    - Define and enforce data collection standards
    - Minimize repeat entries
    - Learn to use databases effectively (Most felines don't really know how relational databases work). Using relational databases minimizes repetition and allows you to query the information you need for analysis in the correct format quickly and easily.

- Use tools such as forms and data-checking techniques in most database programs
- If you are managing a large group of kittens, perform data or lab audits to make sure all felines are collecting data the same way
- Bring in an expert to check your work, or send a certain percentage of your samples to an alternate lab for confirmation.
- Search data for outliers before analysis. Some techniques for finding outliers include basic graphic techniques such as histograms, scatter plots, and quantitative checks such as comparing median with the mean.
- Deal with outliers on a case-by-case basis
- Always document changes made to the data.

Along the way: Describe your data (Metadata!)

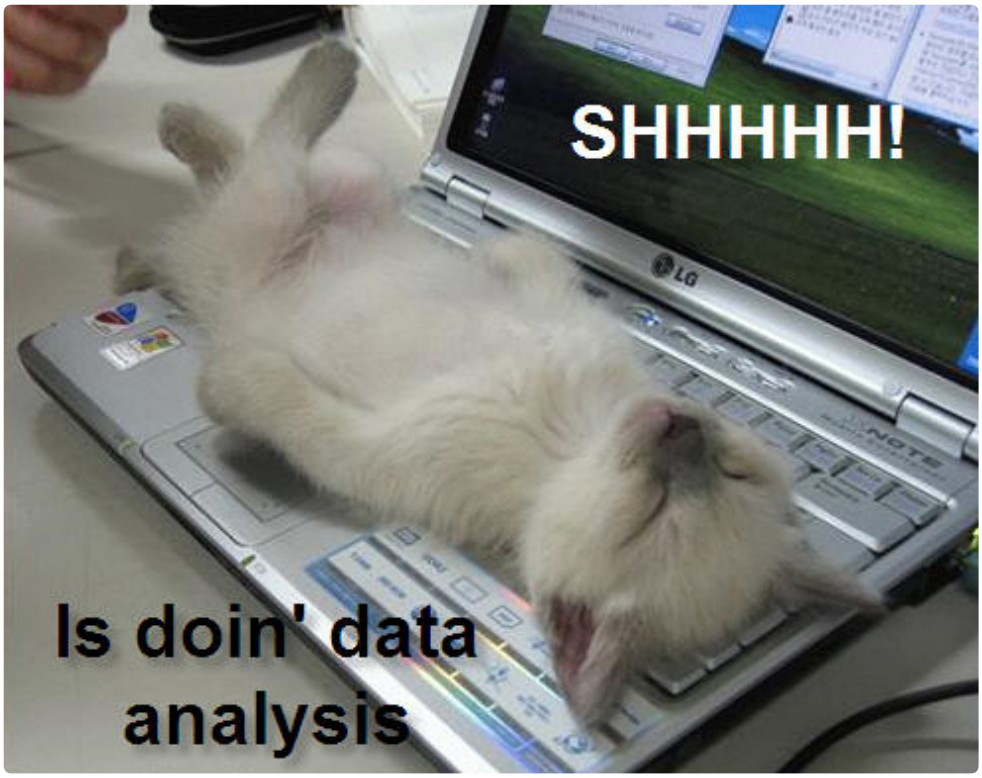
- Metadata is the who, what, where, when, why and how of your data. Like QA/QC, this should follow you throughout the data life cycle
- Record your planning process

- Record what protocols were used to collect all your data, who collected it, when it was collected, and where it was collected
- Record what quality control procedures were used on it, and any changes made after data was collected
- Record where the data will be stored and who will be allowed access to it.
- A document with all of these metadata should be included with your data whenever you do anything!
- This tool can help you put your metadata into a standardized format and stop you from forgetting anything: https://knb.ecoinformatics.org/#tools/morpho

Along the way: Back up and Secure

- Your data is precious. Do not leave your entire dissertation sitting on a thumb drive that gets lost in a hole in your pocket. Do not trust it to your hard drive. Keep it in multiple locations, one of which should be an on-line service that is accessible anywhere.
- Scan or make copies of paper data sheets
- Back up your files in non-proprietary formats such as .csv and .txt. This will make it easier for others to use and will add to its longevity.
- Store metadata with your data.

Step 3 and 4: Process and analyze your data



Don't forget to record exactly when and where cats fell asleep on your laptop during the data analysis stage.

- Hopefully you planned what transformations and statistical analyses you are going to do in the planning part of the data life cycle. Now is the time to get cracking!
- Do not neglect your metadata during this stage in the process. Record what analyses you tried that did not work out as well as the ones you plan on using so that you do not repeat your mistakes.
- I won't go further into how to do data analysis here, that's a whole 'nother bag of cats…

Step 5: Preserve

- If you have been backing up your data regularly, you should be half-way there. However, back-ups are designed to restore what you are working on in case you lose it, while archives and repositories are built for long-term storage and reuse by others.

Choose your data storage wisely.

- Re-evaluate your documentation. Would an outside researcher be able to recreate what you did? Is there sufficient information to place your data in context?
- Store your data in an on-line repository and include a data attribution file with full information on who produced the data set and who should be contacted for more information. These repositories also provide a DOI (Digital Object Identifier), to make your data easier to cite and discover. Repositories include:
    - Figshare (free, integrates with GitHub)
    - Zenodo (free, integrates with GitHub)
    - Dryad (for published research only, costs money, but good support, has integrated data submission with many journals)
    - The California Digital Library's Merritt archive (for UC researchers)
- Consider licensing and legal issues. For example, many federally funded projects require data to be publicly available. Some open-access journals (PLOS and others) also require data to be publicly available. However, data concerning human subjects or location data on species of special concern may be sensitive. (more on legal and privacy issues from DataOne's policy guide)

DOIs allow other people who use your data to cite it properly.

Step 6: Share your data

- Just because your data is up in a publicly available repository doesn't mean people will be able to find it or use it.
- Include information on where your data can be found in any publications you produce using it.
- Submit your dataset to Data Portals and Catalogs to them more visible and more likely to be employed by others. Data Catalogs and Portals (Like DataOne) provide searchable directories of data and usually include many data repositories.
- Tweet it! Facebook it! Social media is the way of the future!

Lots of scientists use twitter these days to publicize new publications, including data sets.

More resources on data management:

USGS data management website:

http://www.usgs.gov/datamanagement/index.php

Cool paper on data collection:

http://www.esajournals.org/doi/pdf/10.1890/0012-9623-90.2.205

Data Management Plan Tool:

https://dmptool.org/dashboard

Data management best practices:

https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

About Rosemary Hartman

Rosie is a just completed her PhD in Ecology studying the effect of introduced trout on amphibians in mountain lakes. Is now working for the California Department of Fish and Wildlife trying where she is planning monitoring of tidal restoration sites and writing a data management plan.

E-mail Other Posts