Lesson Name: Understanding data management and metadata with NEON aquatic macroinvertebrate data

Creator: Kaitlin Stack Whitney

Date: January 11, 2019

Data Management and Metadata

Why do we need to understand data management to work with environmental data?

In this class, we ‘skip’ the data collection process that happens in some of your environmental science classes (e.g. Field Skills class) and go straight to working with environmental datasets.

This is more and more common in environmental science fields, including ecoinformatics and bioinformatics. Even scientists who collect their own datasets often *couple* their field-collected data with datasets collected by others. For example, when I collect beetle samples from farms, I often use National Weather Service weather station data from nearby stations, rather than also collecting all my own temperature and rainfall data.

Also known as “other people’s data,” asking new questions with existing datasets can be very powerful. This approach has been used for decades in other disciplines, such as medicine, by mining health records and other information to find patterns, develop models, and design future studies. Understanding data management and metadata will help you efficiently and successfully work with others’ data, as well as make any data you create or work with more usable for other scholars in the future.

This week we’re working through a series of exercises, designed to be done in our two class periods, to understand why data management and metadata are important, how you can write your own, and how this will help you as an emerging environmental scientist.

What is data management?

Our first unit of the semester was an introduction to 'what is data' and different kinds of data (e.g. . Now we're moving on to *working* with those data.

Data management planning means thinking through the 'life cycle' of the dataset – whether or not you created it. The goal of data management is to produce self-describing data sets. If you give your data to a scientist or colleague who has not been involved with your project, will they be able to make sense of it? The goal is make a plan that lets you and others create, organize, manage, describe, preserve, and share data.

Thinking about data management is for the start, not the end, of projects. In our class, we start from data already in excel spreadsheet form – but data doesn't automatically start there. Often in environmental sciences, data management starts from the planning of a study about to be conducted in a lab or field setting.

Scientists may ask themselves:

> *What kinds of data do I need to collect to answer my study question?*
>
> *When and where will I collect the data?*
>
> *How will I record it?*
>
> *How will I transfer my observations into a form and format that can be analyzed for research?*

The process of entering and cleaning up spreadsheets can be problematic if not well planned. People may abbreviate species names or other information differently, for example, or enter dates using different formats. We're all human – someone may just make an accidental typo. A critical part of data management is giving a good look to the dataset before jumping into graphing and analysis.

The following exercise uses a seemingly long – yet not exhaustive! – list of considerations for you to think about when entering, reviewing, and cleaning up datasets, whether they are yours or someone else's data. You'll consider these questions using real sets of empirical environmental data from NEON.

What is NEON?

NEON is the acronym that stands for National Ecological Observatory Network. It is a US National Science Foundation observation facility with 81 field sites across the continent in a range of ecosystems. The goal of the NEON sites and data collection is to characterize and quantify how US ecosystems are changing. NEON began collecting data in 2012 and data collection is ongoing.

NEON collects lots of different kinds of environmental data and makes its datasets available to researchers and the public. We will be working with NEON datasets for these exercises. The first step is to ‘collect’ (acquire) data, then open and understand it, and then understand how we may be able to use it to answer new questions.

More background information on NEON is available here:
https://www.neonscience.org/observatory/about

What are aquatic macroinvertebrates?

Macroinvertebrates are invertebrates (insects, snails, crustaceans) that are large enough to be visible to human eyes without a microscope. Aquatic macroinvertebrates can live in freshwater or saltwater systems.

In our case, we’ll be examining freshwater aquatic macroinvertebrates. Of those, some communities live in lakes (often called ‘benthic’ as they live on the lake bottoms) and some live in rivers and streams. There are different species and different collection techniques depending on what kind of waterbody is being sampled.

Exercise 1.

1. Open the NEON Data Portal: http://data.neonscience.org/home

This is where researchers, like you!, can explore the data that NEON has collected and made available. There is a LOT of information, so how will we find the data that we want?

2. Open the Data Browsing feature: http://data.neonscience.org/static/browse.html

This is where we can explore datasets that NEON has created and made available. We don't SEE any macroinvertebrates here. So how will we find data about macroinverebrates?

3. Under filters, in the 'keywords' area, type/select "macroinvertebrates"

4. Click 'macroinvertebrate collection'

We would like to explore the data about the macroinvertebrates as whole organisms – there's another data product that's only DNA barcode information.

5. Select only the site 'CARI.'

Remember that I said there are 81 NEON sites! We don't want to download every single observation at every single

6. Select only the 2018 data at the CARI site.

Download the data.

You will likely have a zip folder – or several! – as a result. Open the files. This part will take some time – which files are the data? Which are supporting / supplemental information? How do you know?

Using **ONE** dataset of the set you downloaded (or the files provided in our course folder), consider the following questions and answer based on your review of the data. Be sure to indicate which file you're answering about in your response.

1a. Are values missing? How do you know?

1b. Are zeroes real zeros or are they filled in because data is missing? How do you know?

1c. Are rows or columns duplicated?

1d. Is spelling and naming order consistent?

1e. Are dates consistent?

1f. Are units specified and consistent?

1g. Are column names poorly chosen or ambiguous?

1h. Is the provenance of the data documented? (AKA does it say where all this data is from?)

1i. Do totals typed in differ from aggregates summed?

1j. Are there any suspicious values that look like they may be typos or other simple errors?

1k. Are data cells formatted properly (e.g. text as text, numbers as numbers) in the spreadsheet software?

1l. Is the formatting used for the tables conveying information or done to make the table look pretty?

1m. Are there comments written in cells that require cleaning up?

1n. Do any cells have multiple entries written into the same cell?

Exercise 2.

Still using the dataset you downloaded (or the one provided), your next step is to format the data from **two** datasets into one ‘compiled spreadsheet.’

Converting multiple tabs or tables into one larger table for analysis is a common problem when working with environmental data in real life. When you do so, for any dataset, you implicitly have to consider questions like those listed above and make on-the-spot decisions about how to clean up any errors or inconsistencies.

Save your resulting ‘compiled spreadsheet.’ You will turn it as part of your answers.

What is metadata?

Metadata is information *about* the data. It includes how the data was collected, what the units of measurement are, and descriptions of how to best use the data. Metadata makes data easier to use by all scientists; failure to develop metadata means data can be ‘lost’ (literally or just become unusable) due to the difficulty of understanding the data and its attributes. Undescribed data becomes less useful over time as information about the data is lost.

Imagine you make a spreadsheet and don’t explain what units the columns are in – how would your classmates know and use it appropriately if you weren’t around to ask?

What should good metadata include?

Good metadata includes many overlapping elements:

> *What, when, where, and how of data collection*
>
> *How to find and access the data*
>
> *What kinds of questions the data might be appropriate to apply to*
>
> *Warnings and caveats about known inconsistencies or problems with the data*
>
> *Descriptors about the extent of the data (e.g. number of rows and columns that should be there) to let others know if the file they have is complete*

This information can be displayed in several ways. It could be a separate written description of the dataset, images (like maps or photos), and be in several kinds of formats – a separate tab in a workbook, a text document stored with the data files, or special machine readable metadata file formats. Remember to plan from the user’s perspective. That user may be someone else – like your classmate – or you in the future!

Exercise 3.

Write a metadata document for the compiled spreadsheet you created (Exercise 2). Make it a separate tab in the same workbook and use the guiding questions above.

This exercise doesn't have one right answer. There are many, many different formats of metadata.

For example,

> https://www.ncddc.noaa.gov/metadata-standards/
>
> http://im.lternet.edu/node/1076 <- many examples of environmental metadata standards for different Long Term Ecological Research sites

For our class period today, you can make your own or use a template (in the course documents folder) from the CAP LTER site.

If you don't think you have the information to answer a prompt from the metadata files you have in hand, you should indicate that in your response. So you should color code your metadata document – use black text to indicate parts that you're answering directly from the provided data and use red text to indicate parts that you don't think you have the answer for based on the metadata provided.

Additional Resources for Exercise 1 – 3:

If you are having trouble understanding some of the data or metadata, there are supplemental files for reference that may be helpful. NEON has two documents on their website that go with the aquatic macroinvertebrate samples:

NEON.DOC.003046 AOS Protocol and Procedure: Aquatic Macroinvertebrate Sampling
NEON.DOC.001152: NEON Aquatic Sampling Strategy

http://data.neonscience.org/api/v0/documents/NEON_macroinvertebrate_userGuide_vA

What to Turn In

These exercises are due by turning in your responses to Exercises 1-3 (as a text document) and your compiled Excel workbook by the end of February 5 via the labeled Assignments folder on the course website.

Additional resources on data management, working with data, and metadata:

There are lots of places to do further reading if you're interested in learning more about data management, metadata, and best practices for working with "other people's data. Here's a short list of resources I find useful:

http://www.datacarpentry.org/spreadsheet-ecology-lesson/

http://github.com/Quartz/bad-data-guide

http://cbahlai.github.io/rqm-template/index.html

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