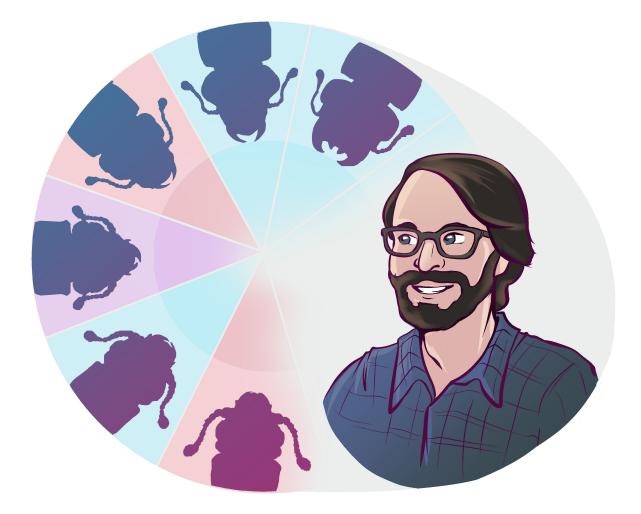


Sorting Through the Data with Thomas (Tommy) McElrath



Learning Objectives:

- -Identify 3 things we can learn about a species using biodiversity data
- -Discuss individual variations in sex
- -Discuss why more than two categories for sex is scientifically accurate and valuable
- -Understand the value of natural history collections in biodiversity data.
- -Identify at least 2 ways errors can be introduced to data
- -Discuss how data errors can hinder scientific research.
- -Explain 3 ways to look for and address errors in data
- -Consider the value of collaboration in science and data.



Meet the scientist:

I'm Thomas McElrath. I'm the insect collection manager here at the Illinois Natural History Survey, which means that I manage a collection of specimens of about 7 million specimens, and then all the associated data with those specimens.

I'm also a beetle researcher. I study the taxonomy and systematics of beetles called monotomids. Those are little tiny brown beetles that mostly live under bark or feeding on minute fungi or sometimes in other insects.

In my spare time I'm usually hanging out with my twin girls who are 18 months old. I also like disc golf, and I like video games quite a lot. Minecraft is one of my favorites.



"What surprises you about biodiversity data?"

One of the things that surprises me most about biodiversity data is how bad it is most of the time. Just how little people talk to each other about how it should be recorded or use the standards. And then when you try to analyze datasets, it becomes a cluster of nonsense in how much of the data they did record or didn't record.

For example, there's the field for "sex" in Darwin Core, which is a standard for recording biodiversity data. Seems pretty straightforward. How many different entries or unique values would there be on a large data aggregator like GBIF for sex?

If you think about how you would record that sex insects, you can think of at least three discrete categories; male, female, and then gynandromorph or intersex. If you want to get into genetics you could throw in a few more categories for sex. Then there's a certain fuzziness with how things get recorded, like if something is marked "M" or "F" or "male" or "female" and whether it's capitalized or not. All of that considered you'd expect okay, maybe there's 20 or 30 unique entries.

But no. There's over 20,000 unique entries in the field for "sex".

Most of those are because people are trying to shove other concepts besides sex into that field. So they will do things like "two males" or "two males, three females". They're adding in an individual count, and they're also not separating out two separate collection objects.

So when you're trying to analyze that data, it's a real challenge. Let's say I want to look at a species and I want to see whether there's a bias towards one sex ratio or not. Unless you figure out how to discretize data properly (which means to separate the information out into distinct useful categories), that's going to be impossible. So that's my biggest surprise with biodiversity data, is just how bad we are at standard-izing it.



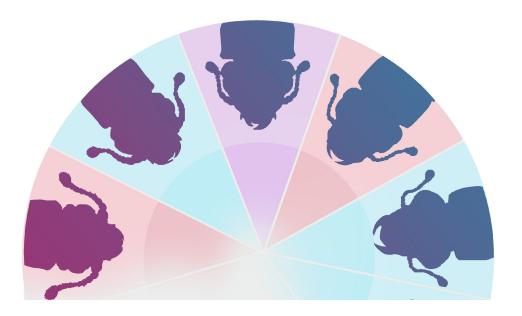
In terms of fun things that have happened in my work with biodiversity data... When I was first coming up in taxonomy, I was re-using the identification keys for this group of beetles. The key noted that one subfamily had a certain characteristic that was different between males and females.

This is a pretty distinct subfamily, so I didn't need that one characteristic to identify them. And after looking at over 150 specimens, I realized that there just weren't any male specimens. All of the specimens we had in our collections worldwide were female. Most likely an assumption was made when they were first collected that if they examined a random number of them, like 10, they'd get some males in there. But no. So it's one of those things where if you make assumptions about sample sizes and ratios you'll just get something completely wrong.

I think natural history collections are great because you get to see that even only looking at male and female (as in, they have male genitalia or they have female genitalia), how different the range of what you could call gender expression is in insects. There's major males, there's minor males. I work on a group of beetles where the males typically have huge heads and giant eyes and it's pretty obvious they're male. And then there are males that don't have those massive heads. There's a huge range between them. There are females with larger heads that will actually exceed the size of some of the smaller males' heads. Then of course there are gynandromorphs, where they literally have both male and female parts. It's a total spectrum. Sex is a spectrum.

As for the head size difference, it's probably a sexual selection thing. The females generally prefer larger males and males that have larger sexual characteristics. Beetles are a really diverse group, and there's really great examples of species in the same genus where one species has massive males with crazy exaggerated sex characteristics. And then a couple species over they don't have those exaggerated characteristics. Because eventually what will happen is whichever sex is having the ornamentation or the thing that is detrimental to its fitness, eventually something that has a smaller version, or something that is not exerting as much resources to invest in that one trait will find a way to reproduce without it.

So for example, in stag beetles often the male's head ornamentations get so exaggerated that they can't move very quickly. When that happens, smaller males will just dart in and mate with females before they get there. So you'll get these arms races where for a while as species will have these giant ornaments until evolution happens and they don't need it anymore.





"How should we approach a data set that we hope to use for research?"

Basically, I'll try to look across the whole dataset, usually using Open Refine or Excel, and clean it up.

So my first step is, does the data align with a standard. And a lot of the datasets I get from my research don't, so I might start by renaming the fields to be Darwin Core compliant since that's what we use here in my organization.

Then we'll check to make sure what's actually listed in the field is what should be there. So you know, checking things like: in the individual account column, or in the numbers of specimens column do they also put in stuff like male female, things like that? We make sure those are pulled out into distinct entries.

I'll also go through and check for duplicates of the same concept. So if, in the data set, there's "Champaign", "Champaign County", "Champaign CO", all representing Champaign County, the concept, I'll go through and change all those to be the same thing.

This is more of a philosophical thing, but I think it's wrong to think that you have to do everything by yourself with data. A lot of people have thought about data management and cleaning data in the past, so getting help with doing it is always a fantastic idea. Never think that you're alone.

Thought Exercise 1:

Discuss the interview above and the role of sexual selection and sexual expression as a spectrum in evolution. Identify at least two ways a species with a spectrum of sexually-selected secondary sex characteristics (such as head or mandible size) might be beneficial to the species' overall resiliency and evolution.

1.



Thought Exercise 2:

You want to investigate a particular species of beetle to understand their sex ratios. Are there more males than females? More females than males? The same amount of both?

You decide to download data from different natural history collections around the world to compare how many specimens are male and how many are female. After you've downloaded data from different natural history collections, what are the next steps? **Name 3 ways you can start looking for and addressing errors in the data.**

1.		
2.		
3.		

Development of this module was supported by the National Science Foundation under DBI-1730526 RCN-UBE: Biodiversity Literacy in Undergraduate Education - Data Initiative (BLUE Data Network) and DBI-1746715 RCN-UBE: Engaging Underrepresented Populations in Biodiversity Sciences: A Series of Workshops for Undergraduate Students. Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

