

Rescuing Data with Diane Srivastava



Learning Objectives:

- -Formulate research questions that could be addressed using data
- -Understand factors that contribute to data's long term stability and usefulness
- -Apply vocabulary and concepts related to biodiversity datasets
- -Discuss how data skills and standards are relevant to the scientific process



Meet the scientist:

I'm Diane Srivastava. I'm an ecologist. I identify as a queer mixed-race woman. I love hiking, skiing and canoe camping.

As an ecologist, I am particularly interested in food webs and how they're structured, and what climate change can be doing to them. My research has taken me around the world, from the arctic to the tropics.My favorite systems are in the tropics and involve insects.

I also do a lot of synthesis of previously published data. And, in fact, I am the director of the Canadian Institute for Ecology and Evolution, which is a synthesis center, which just means that we run programs across Canada where researchers synthesize and save biodiversity data.



Data is incredibly fragile. The rate at which it is lost is astounding. We think you just put it away somewhere and it's done. But data is lost through human error, like humans recycling their hard drives without thinking about it. It's also lost because the underlying technology changes very quickly, and if any proprietary software is involved it can be lost. It's just amazing the rate that we lose what we think is locked down computer data. So if you don't actively look after it it's usually lost within years.

How do you look after data?

Well, keeping it in open formats. Very basic formats like text files or CSV files.

It also means archiving it, but not only the data, but what we call the metadata, which is all the information about the data. So a regular human being who hasn't been part of that project could come in and understand what the data means, or even a future you could understand what the data means. We have little codes for things, because we don't want to, you know, spend a lot of time typing in the full name of something. But then can you actually remember what treatment SC_2 was? I mean, really? That's where you need the metadata.

Living Data Project:

One of the programs that I help run through the Canadian Institute for Ecology and Evolution is called the Living Data Project. It came about as a part of a load of new initiatives started to get scientists to preserve their data.

A lot of those programs kind of assume you were just starting out collecting your data. But in ecology and environmental science the data that was collected in past decades is actually incredibly important. In the time of climate change we need to know what things were like decades ago to know how much things have changed.

So we call it data rescue. Because these are data that have a high value, but they're at risk of extinction. These data are not just in universities, they're in government agencies, they're in nonprofits. We try to



make sure that researchers who are about to retire know about the program because that's a really good point to be thinking about your career data and what's important to save for the next generation.

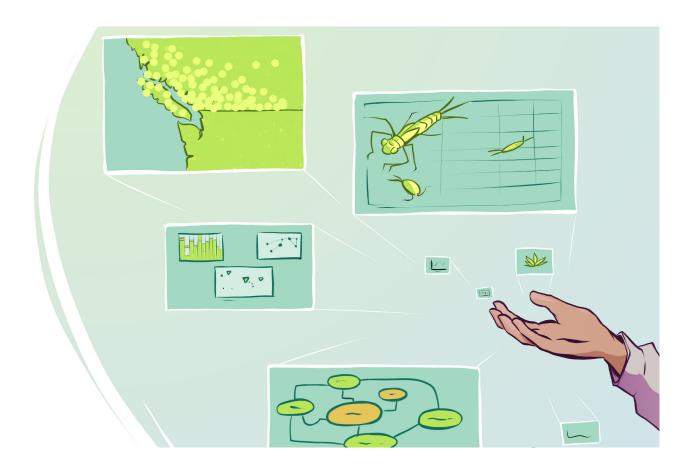
We have to first think about priorities because you can't save everything, there is so much data in the world, and we only have so many resources So I think the hardest thing is to think, okay, which, which is the most important data to save?

For example, if it's a time series... there's no way you could go back and replace a time series. Yet it's so critical for knowing about how things have changed. Or a really extensive spatial survey, that we couldn't replicate.

So you also should think about how we would use this data. Are there compelling use scenarios for this data? You really have to narrow down because you're not going to save all this. You just want to target some data that you can save really well, that has high value.

Exercise:

You have been tasked with saving datasets from nonprofits, government agencies, and individual researchers and universities and incorporating them into a publicly accessible database online. You have limited resources with which to save data, so you want to prioritize the data that you think will be the most important to archive. Review the fictional example datasets on the following page. Discuss what your priorities might be and rank the example datasets in order of which to archive first.





Dorine Dalrymple's lab data.

50 years worth of observations of local prairie pollinators and host plants that one researcher has gathered over the course of her career. The data follows a clear standard that's compatible with your existing databases, but may still require some minor cleaning due to variation in date format and several species names that have changed over the decades.

Jet Jameson's field notebooks data

A 150 year old set of field notebooks and specimens from a naturalist who spent his time studying wetlands areas in the southern United States. There are no geotags to indicate the exact coordinates of each site, which could be a problem. However, the observations are detailed, with walking directions on how he arrived at each site and notes on weather and temperature of each day he was in the field. The field notebooks have been digitized by handwriting recognition software, but the data needs extensive cleaning, error checking, and entry into a usable format before it will be ready to incorporate into your aggregator.

The Migratory Avian Species Project data

A large international collaboration between government agencies, nonprofits, and tribal groups resulted in a massive dataset of where and when 25 migratory bird species were spotted across North America in 2015 and 2016. However, the dataset is a proprietary format and the software to read it is no longer available. One of the collaborators likely has a spreadsheet or text file of the data that they used for their own research, but it will take some time to track it down.

Mike Mayoleth's pesticide research data

A dataset of the effects of a dozen common chemical pesticides on captive bred fruit flies in controlled lab conditions. The data was collected as part of a two year study on the potential impacts of pesticides in 2004. Some of the metadata is missing, such as explanations for the abbreviations of certain chemical names.

Pacific Northwest Weather data

Data on 30 years of weather patterns and wind speed trends in the Canadian Pacific Northwest from a detailed study between 1983-2003 by a government agency. The data would pair well with some datasets that are already in your archive, such as species spatial occurrence data from the same geographic area and time period. The data has been exported as a text file from a proprietary file format, which has caused some errors to appear that will need to be combed through carefully to sort out.

The Meadow Protection League's field survey data

A 20 year old set of field observations of a meadow butterfly that is now critically endangered. The data was collected by volunteers at a nonprofit, and has been recorded inconsistently with varying fields for date, time, location, and whether or not host plants were mentioned. It's also in several different formats and will take time to clean and compile.

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.

