# Downloading and Cleaning Data from iDigBio

Created by: Tiffany Doan, New College of Florida; Tanya Dewey, Colorado State University; John Carroll, Georgia Southern University; Liz Alter, California State University Monterey Bay

## INTRODUCTION

The purpose of your research this semester is to determine if a particular species has changed its geographic range over time and to examine reasons for the change. We have used GBIF and iNaturalist to get comfortable with digital natural history data, but for the remainder of the project, we will shift into working with iDigBio. This exercise provides practice downloading and cleaning digitized natural history collections data from iDigBio. At the end of this exercise, you should be able to:

- Access the iDigBio data portal and download digital specimen records
- Clean the downloaded specimen data

Before you begin, you will watch two short videos explaining the mission of iDigBio and how to use the iDigBio data portal. Watch both before proceeding.

- Digitizing Biological Collections- iDigBio Project Overview
- iDigBio basic search tutorial

## PART 1: DOWNLOADING DATA FROM IDIGBIO

Work together to download the data on your species from iDigBio. Use the tutorial that you watched to help you search for your species. Discuss with your group which settings in the data portal you should use. You will need to input all synonyms of the species as well. Download all the data of each synonym using the download function. The download will give you a zip file with multiple file types. Unzip the files. Set up a Google Drive folder for your group (where all the group members have access) and upload folders of all the files (one folder for each synonym). Although you are keeping all the files, the ones you will use today are called occurrence_raw.

Doan, T., Dewey, T., Carroll, J., & Alter, L. (2022). Distribution CURE: Exploring species distribution changes and their drivers. BCEENET- Biological Collections in Ecology & Evolution Network, QUBES Educational Resources. https://qubeshub.org/publications/2349/1

## PART 2: DATA CLEANING PRACTICE AND CLEANING OF SPECIES FILES

A critically important part of data cleaning is that you save your files with different names at every stage. For our purposes, rename all your occurrence_raw.csv files with the name of the species you searched and the download date instead of occurrence. For example, if you were doing your project on Eastern Grey Squirrels you would call your file Sciurus_carolinensis_20201003_raw.csv. You will notice that that there are no spaces in the file name (underscores are used instead of spaces). Follow these guidelines for your file names. Select the occurrence_raw file that has the most records. Make a copy of that file and name it for example Sciurus_carolinensis_Combined_raw.csv. Copy and paste all the data from the other files with the other synonyms into this single file. Once it has all the data from all the synonyms, save it and make sure to use this file for the rest of the exercise. The activity below uses the example of occurrence data for the plant species *Shortia galacifolia*, the Oconee Bell, but you should use the files of your species.

1. Visit the resource Cleaning biodiversity data (Gaynor, 2020)
2. Open the file called: AdaptationDataCleaning.docx
3. Follow the instructions provided in the document but stop at step 3b and complete the “Additional Location Cleaning” below.

### *Additional Location Cleaning*

Remove specimen records with unlikely localities.

Delete any rows with decimal longitude or decimal latitude coordinates of 0.00

**Save another file with your species name, date, and “Cleaned.csv”** (for example Shortia_galacifolia_2020-09-01_Cleaned.csv). You will use this cleaned dataset for the georeferencing activity.

### References

Gaynor, M. (2020). Cleaning biodiversity data: A botanical example using Excel or RStudio. Biodiversity Literacy in Undergraduate Education, QUBES Educational Resources. doi:10.25334/DRGD-F069