**DATA CLEANING**

This was an adaptation created by Shelly Gaynor (University of Florida) for a BCEENET (<https://bceenetwork.org/>) workshop in 2020 using the species *Shortia galacifolia*, the Oconee bells or acony bell, which is a rare North American plant in the family Diapensiaceae found in the southern Appalachian Mountains.

Reference the original resource for more background information: <https://qubeshub.org/publications/1899/>

**Background:**

* Learning objective: Understanding data structure, standards, and reproducibility.
  + Data organization best practices (Reading: Broman and Woo, 2018)
  + Data standards (see https://www.tdwg.org/)
    - [Darwin Core Standards](https://www.tdwg.org/standards/dwc/)
      * What columns do we want to retain?
        + This will depend on your research projects.
  + Data reproducibility
    - Saving data along the way – you should have files for:
      * Raw data
      * Cleaned data
    - Archiving and attributions.

**Activity**

* Manual in Microsoft Excel or Google Doc.

**Basic Steps:**

1. Download data and save raw “.csv”
2. Resolve taxon names
3. Remove duplicates
4. Location cleaning
5. Save Cleaned “.csv”
6. **Manual (Microsoft Excel or Google Doc)**
7. **Raw data is located in “raw/Shortia\_galacifolia\_062620.csv”**
8. **Resolves taxon names**
9. Remove rows that do not belong to the focal taxon and its synonyms

A screenshot of a computer

Description automatically generated

1. Copy the filtered to a new sheet
   * Select all: ‘Ctrl’/’Command’ ‘A’
   * Copy: ‘Ctrl’/’Command’ ‘C’
   * Add Sheet
   * Paste: ‘Ctrl’/’Command’ ‘V’
   * Delete the original sheet

A screenshot of a cell phone

Description automatically generated

1. Insert column named “name”
   * Fill in all rows with your accepted taxon name: “Shortia galacifolia”

A screenshot of a cell phone

Description automatically generated

1. Save copy in cleaned folder

A screenshot of a cell phone

Description automatically generated

1. **Removes duplicates**
   1. Select columns to retain, copy+paste into new sheet, and rename:
      * + coreid = ID
        + name
        + dwc.basisOfRecord = basis
        + dwc.catalogNumber = catalogNumber
        + dwc.collectionCode = collectionCode
        + dwc.collectionID = collectionID
        + dwc.coordinateUncertaintyInMeters = coordinateUncertaintyInMeters
        + dwc.decimalLatitude = lat
        + dwc.decimalLongitude = long
        + dwc.eventDate = date

**A screenshot of a computer screen

Description automatically generated**

**Hint:** Search for the columns by selecting the first row and using the search feature. Select the whole row by clicking the column name. Copy: ‘Ctrl’/’Command’ ‘C’, Paste: ‘Ctrl’/’Command’ ‘V’. Then delete the old sheet.

b. Remove identical rows

* + - ‘Data’ -> ‘Remove Duplicates’
      * Select columns lat, long, and date.
        + If a specimen shares lat, long, and event date we are assuming that it is identical. Many specimen lack date and lat/long, so this may be getting rid of information you would want to keep.

A screen shot of a monitor

Description automatically generated

1. **Location cleaning (OPTIONAL)**
   1. Remove specimen with missing latitude/longitude
      * + Filter ‘(Blanks)’ and Copy/Paste into new sheet.

A screenshot of a video game

Description automatically generated

* 1. Rounds up the latitude/longitude to our desired coarseness and removes points that are not precise enough
     + - Select ‘lat’ and ‘long’
       - Change format from ‘General’ to ‘Number’
       - Round using the two buttons circled below.

A screenshot of a cell phone

Description automatically generated

**Plants Specific Steps:**

* 1. Remove unlikely points:
     + - Removes coordinates at 0.00
         * Delete any rows where lat/long is 0.00/0.00
       - Removes coordinates in cultivated zones, botanical gardens, or outside our desired range
         * There aren’t easy ways to do this in Microsoft Excel or Google docs.

1. **Save Cleaned “.csv”**