**DATA CLEANING**

This tutorial is adapted slightly from a module presented to BCEENET in 2020, developed by Shelly Gaynor, with input from Molly Phillips.

**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 are located in “Shortia\_galacifolia\_062620.xlsx”**
8. **Resolve taxon names**
9. Remove rows that do not belong to the focal taxon and its synonyms
   * Sort on column dwc:scientificName (column EV); values; A to Z
   * 2 options:
     + Delete all rows that do not have desired scientific name (*Shortia galacifolia*) [if you are not sure they are the desired taxon]
     + Keep all rows and change the scientific name of those that differ from *Shortia galacifolia* because you know that they are synonyms

A screenshot of a computer

Description automatically generated

1. Copy the filtered data 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” with all rows having accepted name
   * Change “name” heading to “Shortia galacifolia”
   * Fill in all rows with your accepted taxon name (“Shortia galacifolia”) by
     + Click on the column heading above “Shortia galacifolia”
     + Click on Editing 🡪 Fill 🡪 Down
   * Change heading back to “name”

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. **Remove duplicates**
   1. You may want to remove some columns from the csv because they are not needed for your purposes. To do so, select columns to retain, copy+paste into new sheet, and rename, for example, as follows:
      * + 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 column by clicking the letter above the column name. Copy: ‘Ctrl’/’Command’ ‘C’, Paste: ‘Ctrl’/’Command’ ‘V’. Then delete the old sheet.

b. Remove identical rows

* + - ‘Data’ -> ‘Remove Duplicates’
      * Records that have the same unique identifier
      * Records that have the same collectionID, collectionCode, and catalogNumber
      * Records that have the same lat, long, and date
        + If two records share lat, long, and event date, they may be identical. Many specimens lack date and lat/long, so this could remove information you would want to keep.
    - Select all; then ‘Data’ -> ‘Remove Duplicates’
    - Check ‘My list has headers’
    - Select columns to search
    - Click OK

A screen shot of a monitor

Description automatically generated

1. **Location cleaning (OPTIONAL)**
   1. Remove specimens with missing latitude/longitude
      * + Filter ‘(Blanks)’ and Copy/Paste into new sheet.
          - Highlight Lat and/or Long columns
          - Under ‘Data 🡪 Filter’, click on small arrow by column name
          - In the menu that appears, scroll through the list of values and uncheck ‘Blanks’

A screenshot of a video game

Description automatically generated

* 1. Round up the latitude/longitude to our desired coarseness and remove points that are not precise enough
     + - Select ‘lat’ and ‘long’
       - Under ‘Home’, change format from ‘General’ to ‘Number’
       - Round to the desired number of decimal places using the two buttons circled below.

A screenshot of a cell phone

Description automatically generated

* 1. Remove unlikely points:
     + - Remove coordinates at 0.00
         * Delete any rows where lat/long is 0.00/0.00 (find them using Find)
       - Remove 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, but this is easily done in R

1. **Save Cleaned “.csv”**