DATA CLEANING

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
            - 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

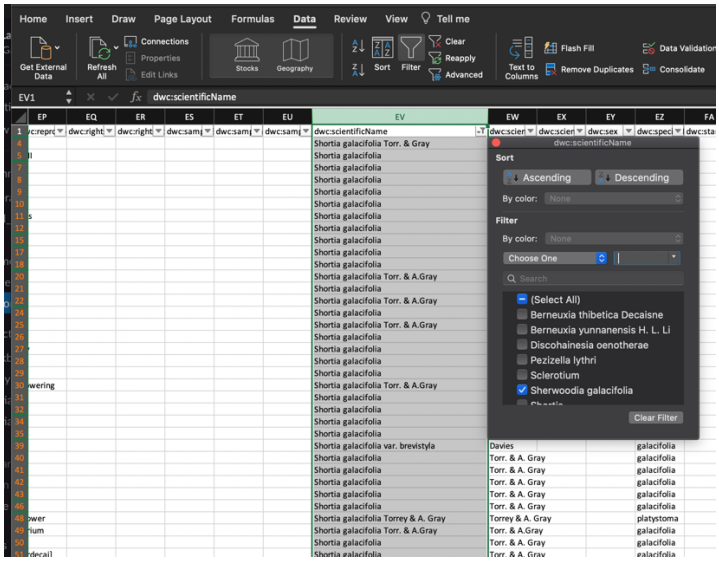
- Manual in Microsoft Excel or Google Doc.
- Depending on the size of your dataset and your comfort with R and RStudio, you may or may not want to use the R script that we provide.
    - R studio can be accessed via QUBES hub

Basic Steps:

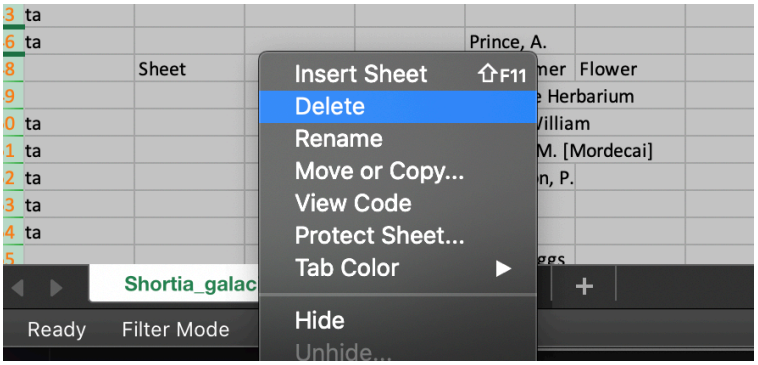
0. Download data and save raw ".csv"
1. Resolve taxon names
2. Remove duplicates
3. Location cleaning
4. Save Cleaned ".csv"

(A) Manual (Microsoft Excel or Google Doc)

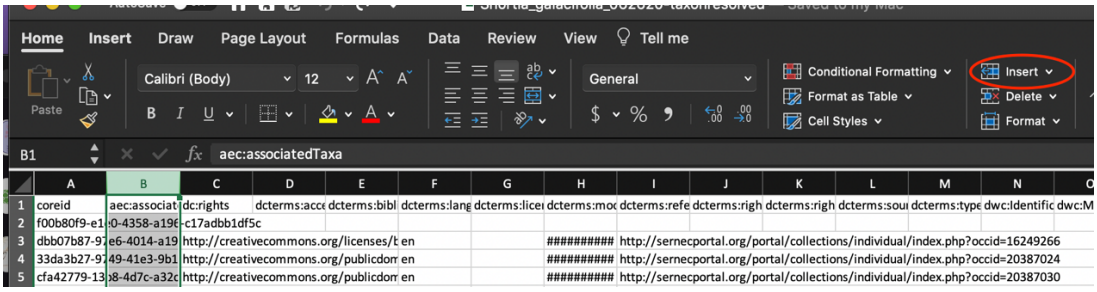
0. **Raw data is located in "Manual/raw/Shortia_galacifolia_062620.csv"**
1. **Resolves taxon names**
    a. Remove rows that do not belong to the focal taxon and its synonyms



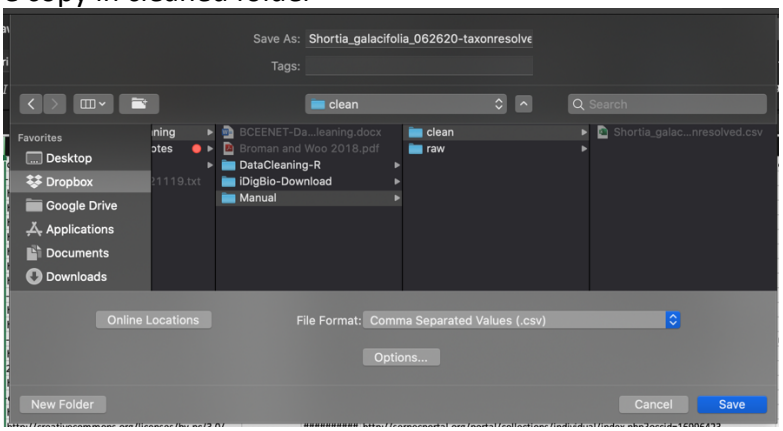
    b. Copy the filtered to a new sheet
        - Select all: 'Ctrl'/'Command' 'A'
        - Copy: 'Ctrl'/'Command' 'C'
        - Add Sheet
        - Paste: 'Ctrl'/'Command' 'V'
        - Delete the original sheet



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



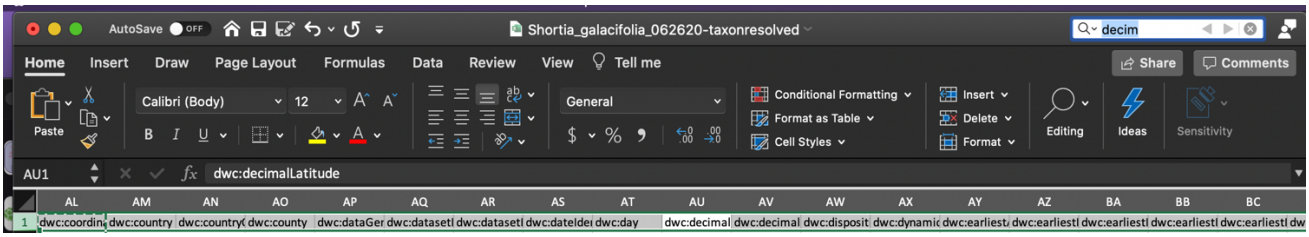
d. Save copy in cleaned folder



2. Removes duplicates

a. Select columns to retain, copy+paste into new sheet, and rename:

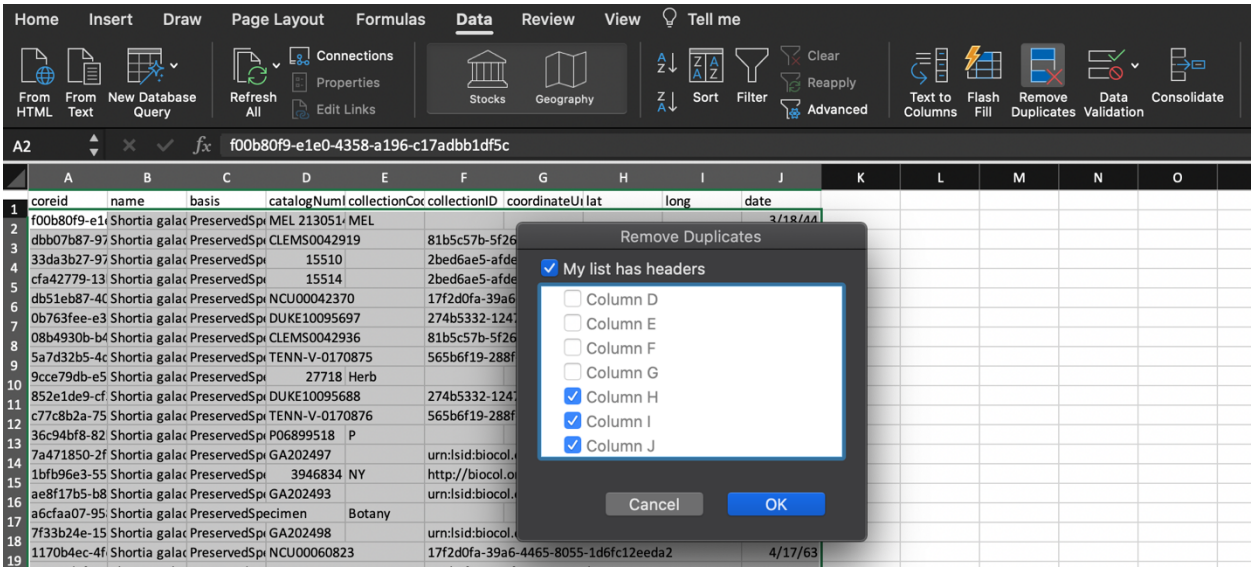
- coreid = ID
- name
- dwc.basisOfRecord = basis
- dwc.catalogNumber = catalogNumber
- dwc.collectionCode = collectionCode
- dwc.collectionID = collectionID
- dwc.coordinateUncertaintyInMeters = coordinateUncertaintyInMeters
- dwc.decimalLatitude = lat
- dwc.decimalLongitude = long
- dwc.eventDate = date



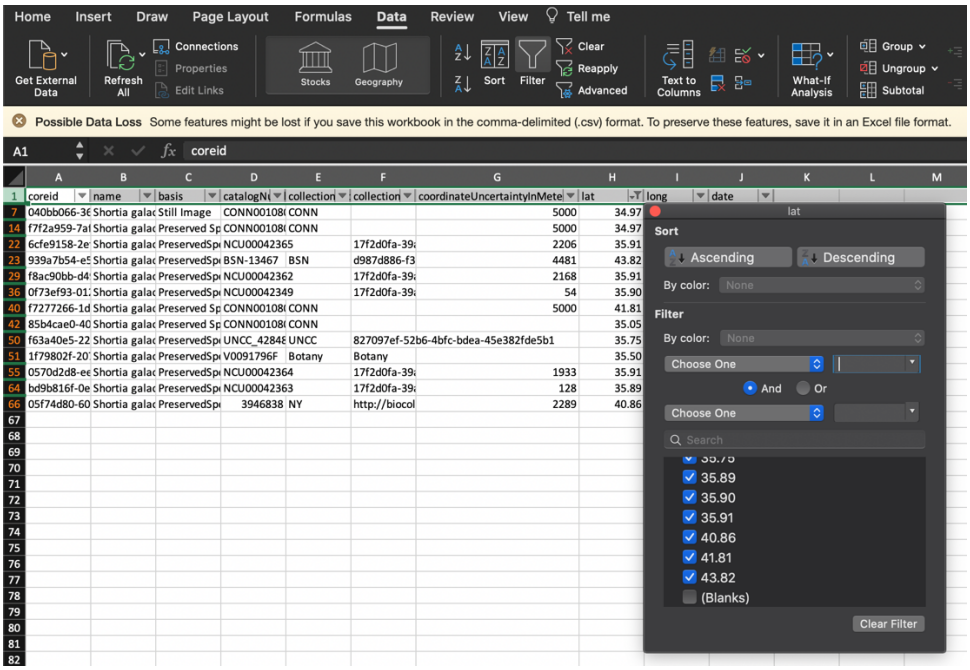
Hint: Search for the columns by selecting the first row and using the search feature. Select the whole row by clicking the column name. Copy: 'Ctrl'/'Command' 'C', Paste: 'Ctrl'/'Command' 'V'. Then delete the old sheet.

b. Remove identical rows

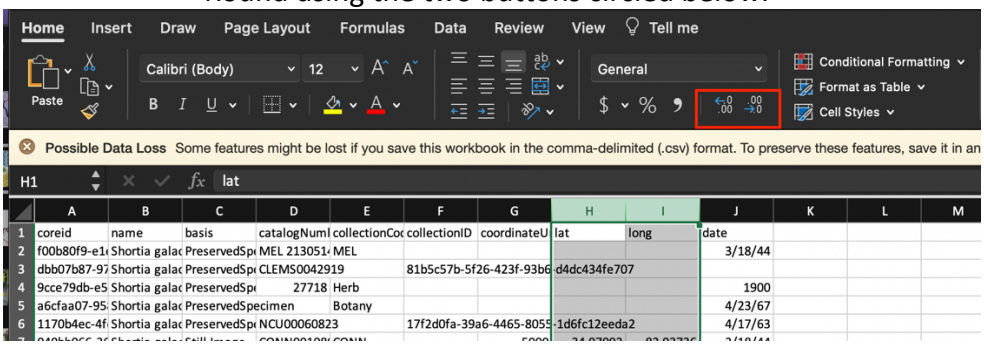
- 'Data' -> 'Remove Duplicates'
  - Select columns lat, long, and date.
    - If a specimen shares lat, long, and event date we are assuming that it is identical. Many specimen lack date and lat/long, so this may be getting rid of information you would want to keep.



3. **Location cleaning (OPTIONAL)**
    a. Remove specimen with missing latitude/longitude
        - Filter '(Blanks)' and Copy/Paste into new sheet.



    b. 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.



Plants Specific Steps:

    c. 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.

4. **Save Cleaned ".csv"**

(B) R based

Files are located in "DataCleaning-R" folder.

- Open the R project by double clicking the DataCleaning-R.Rproj file.
    - An R script is available "DataCleaning.R", as well as a PDF "DataCleaning.pdf"