

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

Home Insert Draw Page Layout Formulas Data Review View Tell me

Get External Data Refresh All Properties Edit Links Stocks Geography Sort Filter Advanced Text to Columns Remove Duplicates Consolidate

EV1 dwc:scientificName

EP	EQ	ER	ES	ET	EU	EV	EW	EX	EY	EZ
1	dwc:left	dwc:right	dwc:right	dwc:samj	dwc:samj	dwc:scientificName	dwc:scientificName	dwc:sciz	dwc:sciz	dwc:sciz
2						Shortia galacifolia Torr. & Gray	Sort			
3						Shortia galacifolia	Ascending Descending			
4						Shortia galacifolia	By color: None			
5						Shortia galacifolia	Filter			
6						Shortia galacifolia	By color: None			
7						Shortia galacifolia	Choose One			
8						Shortia galacifolia Torr. & A. Gray	Q Search			
9						Shortia galacifolia	(Select All)			
10						Shortia galacifolia	<input type="checkbox"/> Berberis thibetica Decaisne			
11						Shortia galacifolia Torr. & A. Gray	<input type="checkbox"/> Yunnanis H. L. Li			
12						Shortia galacifolia	<input type="checkbox"/> Dischidiales oenotherae			
13						Shortia galacifolia	<input type="checkbox"/> Pezizella lythri			
14						Shortia galacifolia	<input type="checkbox"/> Sclerotium			
15						Shortia galacifolia Torr. & A. Gray	<input checked="" type="checkbox"/> Sherwoodia galacifolia			
16						Shortia galacifolia	Photo			
17						Shortia galacifolia	Clear Filter			
18						Shortia galacifolia var. brevistyla	Davis			galacifolia
19						Shortia galacifolia	Torr. & A. Gray			galacifolia
20						Shortia galacifolia	Torr. & A. Gray			galacifolia
21						Shortia galacifolia	Torr. & A. Gray			galacifolia
22						Shortia galacifolia	Torr. & A. Gray			galacifolia
23						Shortia galacifolia	Torr. & A. Gray			galacifolia
24						Shortia galacifolia Torrey & A. Gray	platyoma			
25						Shortia galacifolia Torr. & A. Gray	Torr. & A. Gray			galacifolia
26						Shortia galacifolia	Torr. & A. Gray			galacifolia
27						Shortia galacifolia	Torr. & A. Gray			galacifolia
28						Shortia galacifolia	Torr. & A. Gray			galacifolia
29						Shortia galacifolia	Torr. & A. Gray			galacifolia
30						Shortia galacifolia	Torr. & A. Gray			galacifolia
31						Shortia galacifolia	Torr. & A. Gray			galacifolia
32						Shortia galacifolia	Torr. & A. Gray			galacifolia
33						Shortia galacifolia	Torr. & A. Gray			galacifolia
34						Shortia galacifolia	Torr. & A. Gray			galacifolia
35						Shortia galacifolia	Torr. & A. Gray			galacifolia
36						Shortia galacifolia	Torr. & A. Gray			galacifolia
37						Shortia galacifolia	Torr. & A. Gray			galacifolia
38						Shortia galacifolia	Torr. & A. Gray			galacifolia
39						Shortia galacifolia	Torr. & A. Gray			galacifolia
40						Shortia galacifolia	Torr. & A. Gray			galacifolia
41						Shortia galacifolia	Torr. & A. Gray			galacifolia
42						Shortia galacifolia	Torr. & A. Gray			galacifolia
43						Shortia galacifolia	Torr. & A. Gray			galacifolia
44						Shortia galacifolia	Torr. & A. Gray			galacifolia
45						Shortia galacifolia	Torr. & A. Gray			galacifolia
46						Shortia galacifolia	Torr. & A. Gray			galacifolia
4										

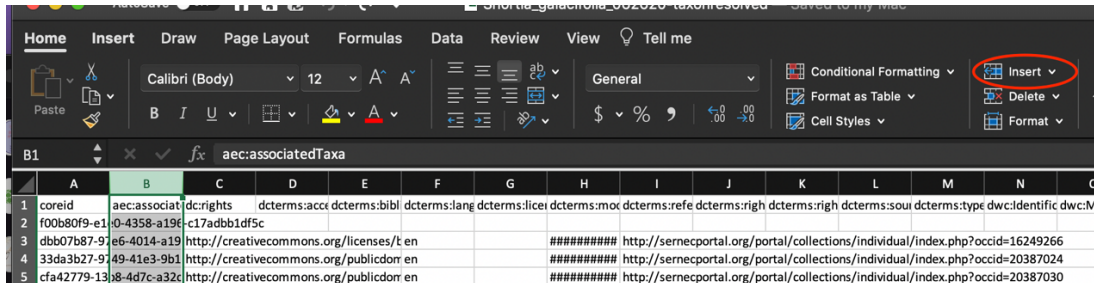
- b. Copy the filtered to a new sheet

- Select all: 'Ctrl'/'Command' 'A'
- Copy: 'Ctrl'/'Command' 'C'
- Add Sheet
- Paste: 'Ctrl'/'Command' 'V'
- Delete the original sheet

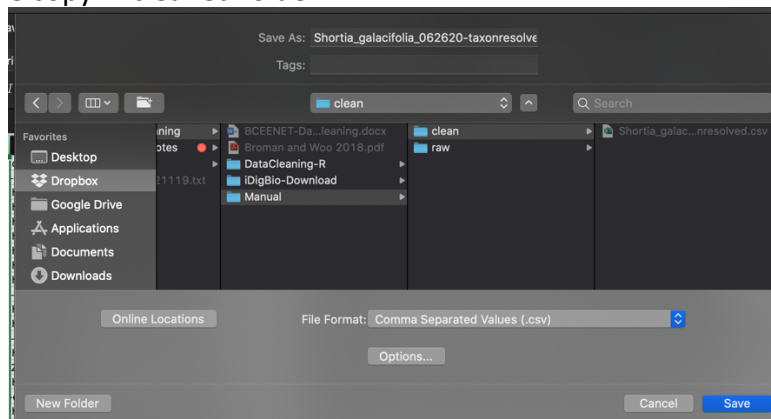
A screenshot of the Google Sheets interface. The 'Insert Sheet' menu is open, showing options: Insert Sheet (with a plus icon and 'F11'), Delete (highlighted in blue), Rename, Move or Copy..., View Code, Protect Sheet..., and Tab Color (with a right arrow icon). The background shows a spreadsheet with a column of tabs labeled 'ta' and a row labeled 'Sheet'. Other visible text includes 'Prince, A.', 'Flower', 'Herbarium', 'William', 'M. [Mordecai]', 'n, P.', 'pps', 'Shortia_galac', 'Ready', and 'Filter Mode'.

c. Insert column named “name”

- Fill in all rows with your accepted taxon name: “Shortia galacifolia”



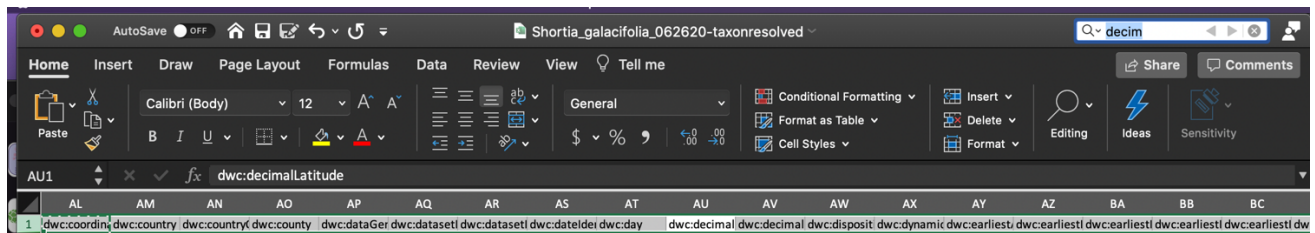
d. Save copy in cleaned folder



2. Removes duplicates

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

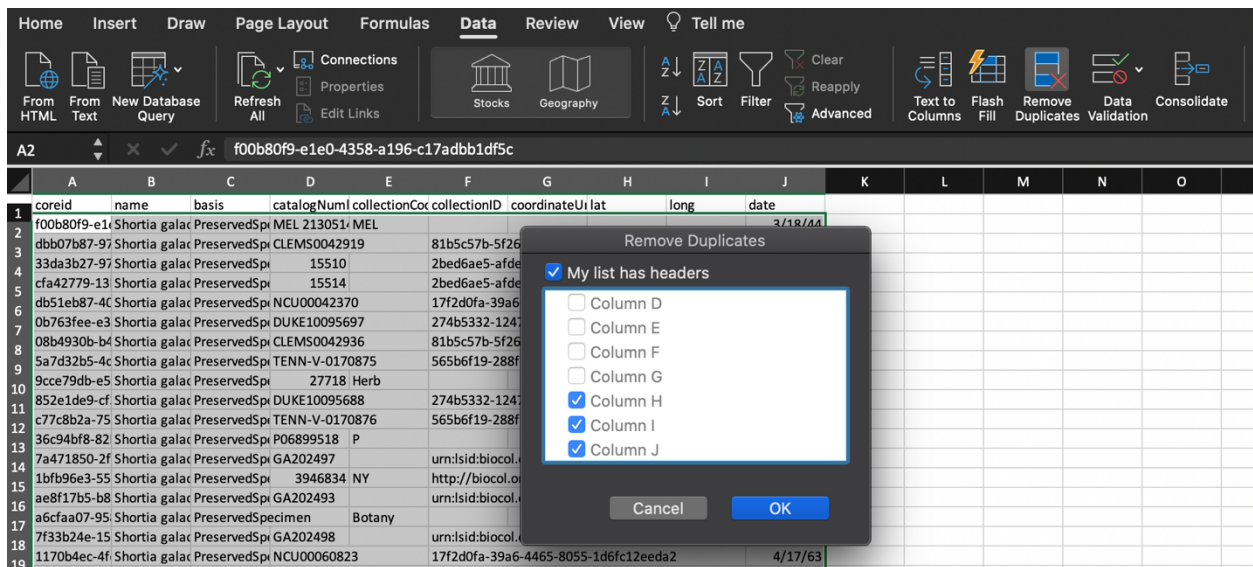
- coreid = ID
- name
- dwc.basisOfRecord = basis
- dwc.catalogNumber = catalogNumber
- dwc.collectionCode = collectionCode
- dwc.collectionID = collectionID
- dwc.coordinateUncertaintyInMeters = coordinateUncertaintyInMeters
- dwc.decimalLatitude = lat
- dwc.decimalLongitude = long
- dwc.eventDate =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

Excel interface showing a data table with columns: coreid, name, basis, catalogNum, collection, collectionCoc, collectionID, coordinateUncertaintyMeters, lat, long, date. A 'Sort' dialog box is open, showing the 'lat' column selected for filtering. The filter results show values like 35.89, 35.90, 35.91, 40.86, 41.81, 43.82, and '(Blanks)'.

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.

Excel interface showing the 'Number' format tab in the ribbon. The 'lat' and 'long' columns are selected. The 'Number' format options are visible, including 'General', 'Number', 'Text', 'Percentage', 'Currency', 'Accounting', 'Date', 'Time', 'Fraction', 'Scientific', and 'Custom'. The 'Number' format is selected, and the 'Round up' and 'Round down' buttons are circled.

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"