How to extract duplicate references using R script and data cleaning in Excel

VERSION 2.0

KATIE PEARSON – MARCH 25, 2021

DEVELOPED FOR THE CALIFORNIA PHENOLOGY NETWORK

This code was developed to rapidly georeference herbarium specimens by importing georeference data that already exist for duplicate specimens. The code looks through each record in a provided dataset and determines whether that record is found in the *omoccurduplicatelink* table (i.e., a duplicate has been linked to that record in your Symbiota portal). If a georeferenced duplicate is found, the code will add these data into a new output file (newMycoll) such that the unique identifier (occid), catalog number, other catalog number, collector, and collector number corresponds to those fields from the **input** dataset, and the georeference data is copied from the duplicate record. The user must then clean the output file so it results in one row per specimen record.

INPUT:

1. Dataset (CSV file) of non-georeferenced specimen records from the target collection. The dataset must contain, at minimum, the column **id** (the Symbiota unique identifier).
2. The *omoccurduplicatelink* table (as a CSV file) from your Symbiota portal (You will need backend access to your portal to download this. Contact your portal administrator for help accessing this file). Make sure to download this file AFTER you have completed Step 1 for your target collection.
3. A simplified download of the entire Symbiota *omoccurrences* table for your portal, ideally only containing specimens that have been georeferenced, containing the following fields: **occid**, **decimalLatitude**, **decimalLongitude**, **geodeticDatum**, **coordinateUncertaintyInMeters**, **footprintWKT**, **coordinatePrecision**, **georeferencedBy**, **georeferenceSources**, and **georeferenceRemarks**.
4. Optional: CSV file of collection ID numbers (collids) and their corresponding collection acronyms. The file should contain a column called **CollidEquals** in which each collid has "collid=" appended to each number (e.g., "collid=123"). The column **Acronym** should contain the standard acronym of the collection (e.g., "OBI"). This file can be created from the *omcollections* table in your Symbiota portal. (You will need backend access to your portal to download this. Contact your portal administrator for help accessing this file).

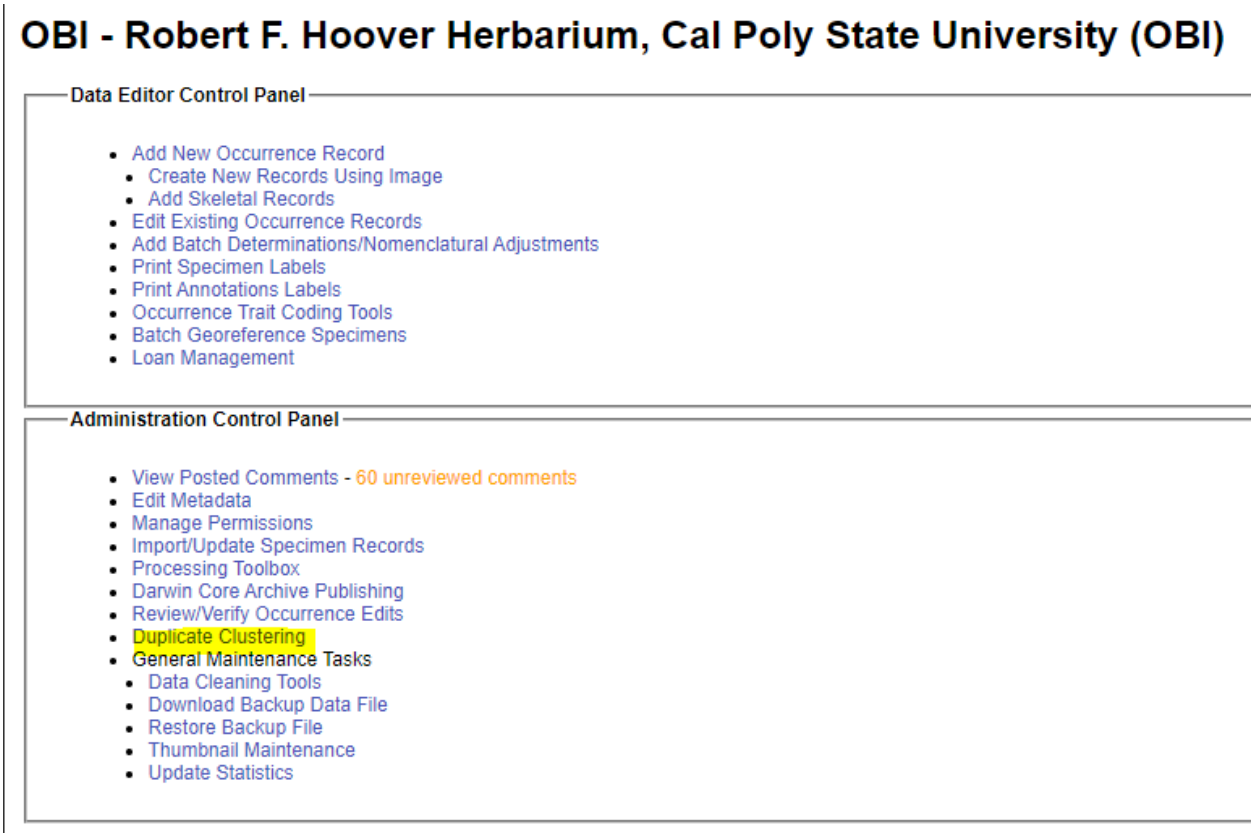
OUTPUT:

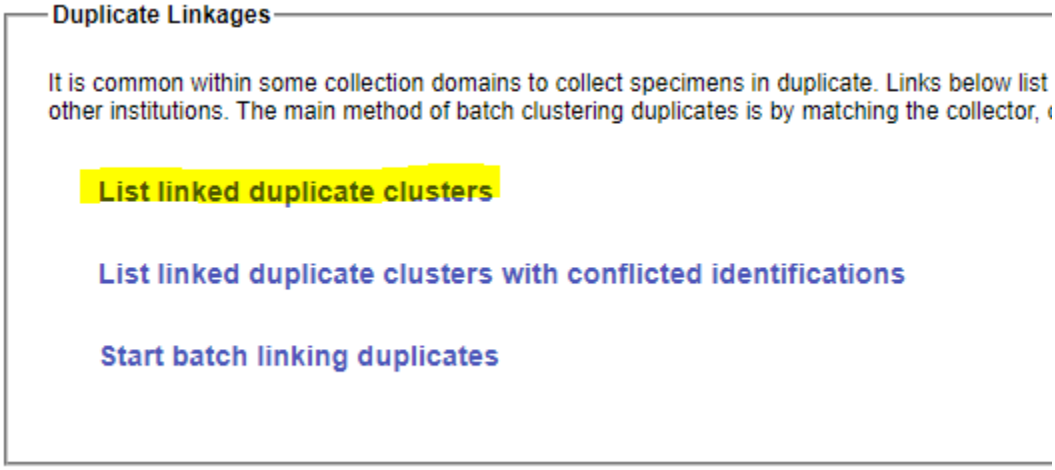
- A dataset containing georeference data (see columns of input #3) for all specimens from your target collection (input #1) that have georeferenced duplicates.
    - Note that, in the case of multiple duplicates, one record in the target dataset may result in several records/rows in the output dataset. For this reason, cleaning steps 3-6 are provided.

These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

STEP 1: ENSURE DUPLICATES ARE LINKED IN YOUR SYMBIOTA PORTAL

Run the duplicate clustering tool for your target collection to ensure that duplicate records are identified and stored in the *omoccurduplicatelink* table.





These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Step 2: Run the code

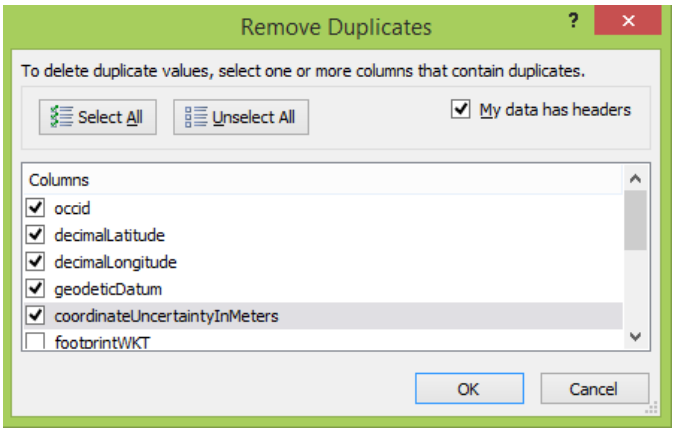
Run the ExtractDuplicateGeoreferences R code, making sure to update the path to each input file as described in the comments of the code. The dataset of the collection you are seeking to update (input #1) should be loaded as the "mycoll" data frame. Input #2 will be called *duplinks*, and input #3 will be called *others.* Other notes and configurations are listed below and in the comments of the code.

Notes

- Your input file (input #1) should ideally only contain specimen records that have **not** already been georeferenced. If your input file contains specimens that ARE georeferenced, uncomment lines 45-47 and line 103 to include a step that skips already georeferenced records.
- Your reference table (input #3) should ideally only contain specimen records that **have** been georeferenced. If your reference table file contains specimens that are NOT georeferenced, uncomment lines 77-79 and line 98 to include a step that skips reference records that are not georeferenced.
- In this version of the code, the georeferenceRemarks field will be populated with the statement "copied from duplicate," followed by the collection id number (collid, according to the Symbiota portal) of the duplicate specimen, the catalog number of the duplicate specimen, and the georeference remarks from that specimen. The code below line 108 replaces the "collid=" values with the acronyms of their respective institutions. This relies on the optional input file (input #4) named in the **Input** section of this documentation.

(Optional) Step 3: Remove duplicate duplicates

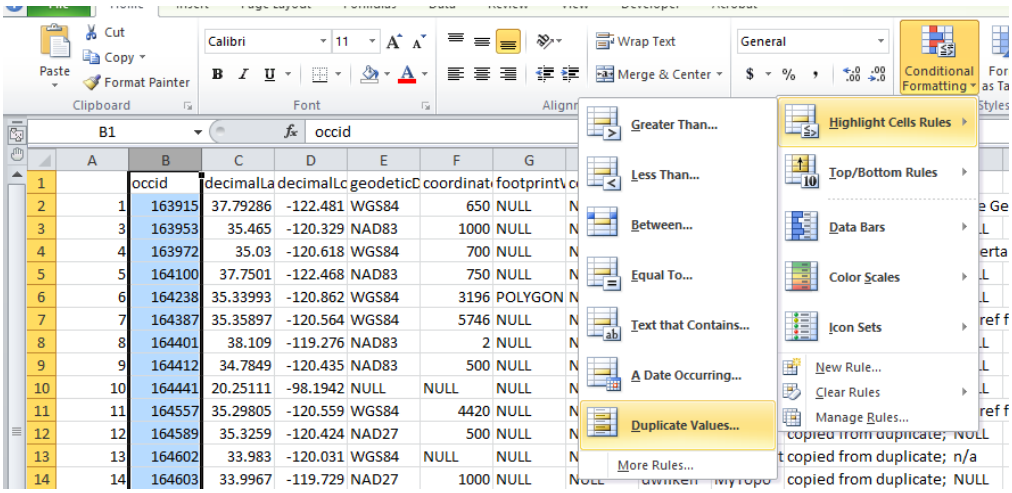
Remove exact duplicate lat/longs+uncertainty in Excel:



These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Step 4: View duplicates

Visualize duplicates that still exist:



Step 5: Select best duplicate georeferences

Ultimately, your collection can decide which of the duplicate georeferences you will want to keep. The following rules were used by the CAP TCN to select georeferences, with some exceptions. We prioritized georeferences with:

1. purported coordinates from label/collector
2. non-NULL georeference remarks, datum, coordinate uncertainty, etc. fields
3. named georeferencers (not just UCBerkeley, etc.)
4. smaller error radii (unless they are suspicious)

Examples

813	185036	37.221	-118.609	NAD83	1000	NULL	NULL	NULL	MyTopo	copied from duplicate; NULL
814	185036	37.22111	-118.609	NULL	NULL	NULL	NULL	NULL	GoogleEa	copied from duplicate; NULL

Keep 813 because it has datum and error radius (1000)

819	185152	34.4	-119.714	NAD83	1000	NULL	NULL	dwilken	MyTopo	copied from duplicate; NULL
820	185152	34.44332	-119.72	WGS84	NULL	NULL	NULL	n/a	Coordinat	copied from duplicate; n/a
821	185152	34.4414	-119.714	WGS84	161	NULL	NULL	UCDavis I	RSA: 2007	copied from duplicate; NULL
822	185152	34.44333	-119.72	NULL	NULL	NULL	NULL	NULL	Berkeley	copied from duplicate; NULL

Keep 821 because it has the smallest error radius

824	185153	34.437	-119.667	NAD27	1000	NULL	NULL	dwilken	MyTopo	copied from duplicate; NULL
825	185153	34.43838	-119.668	WGS84	NULL	NULL	NULL	n/a	Coordinat	copied from duplicate; n/a
826	185153	34.44056	-119.667	NULL	NULL	NULL	NULL	NULL	Berkeley	copied from duplicate; NULL

Keep 824 because it has a georeferencer (dwilken)

These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

1332	191747	39.7664	-122.523	NAD83	400	NULL	NULL	NULL	MyTopo	copied fro
1333	191747	39.76639	-122.523	NAD83	805	NULL	NULL	Bill Carlso	NULL	copied fro

Keep 1333 because it has a georeferencer (Bill Carlson)

1390	192428	39.2903	-122.749	NAD 83	NULL	NULL	NULL	Collector	Label	copied fro
1391	192428	39.29028	-122.749	NAD 1983	2	NULL	NULL	NULL	NULL	copied fro

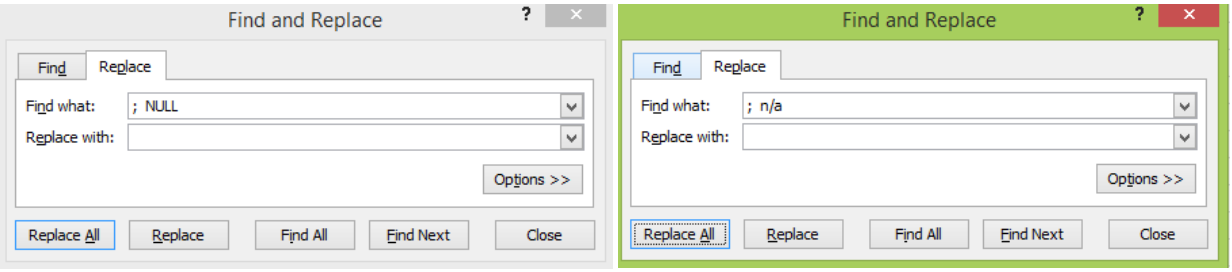
Keep 1390 because the source is the collector/specimen label

1476	194414	35.47843	-120.543	NAD83	NULL	NULL	NULL	NULL	NULL	copied from duplicate; 1/4-1 section (appx. 400-1600 M); 1/4-1 section (appx. 400-1600 M)
1478	194414	35.47842	-120.543	NAD83	NULL	NULL	NULL	NULL	NULL	copied from duplicate; (copied from RSA778854)

Keep 1476 because georeference remarks are more specific

Step 6: Final cleaning

Remove "; NULL" and "; n/a" from the end of georeference remarks field (if present)



Step 7: Pre-integration check

Before you can re-integrate the georeferences into your Symbiota database, you will need to ensure that none of the fields into which you plan to import data already have data in your database. Check your input file (input #1) for any pre-existing values in the **decimalLatitude**, **decimalLongitude**, **geodeticDatum**, **coordinateUncertaintyInMeters**, **footprintWKT**, **coordinatePrecision**, **georeferencedBy**, **georeferenceSources**, and **georeferenceRemarks** fields.

If any data exist in the above-named fields in your input dataset, you will need to edit these records in your Symbiota portal (if you want to replace pre-existing data with the output of this code) or remove the line corresponding to that record in your output file (if you want to retain the pre-existing data rather than replace them with the output of this code).

Step 8: Re-integrate data into Symbiota database

Upload the data into your database using a Skeletal Text File Import tool. This will ensure that only the data from the fields in your output data file are changed and that any pre-existing data (e.g., georeferences that were added during that time it took you to run the code) are not overwritten.

These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

OBI - Robert F. Hoover Herbarium, Cal Poly State University (OBI)

Data Editor Control Panel

- Add New Occurrence Record
  - Create New Records Using Image
  - Add Skeletal Records
- Edit Existing Occurrence Records
- Add Batch Determinations/Nomenclatural Adjustments
- Print Specimen Labels
- Print Annotations Labels
- Occurrence Trait Coding Tools
- Batch Georeference Specimens
- Loan Management

Administration Control Panel

- View Posted Comments - 60 unreviewed comments
- Edit Metadata
- Manage Permissions
- Import/Update Specimen Records
  - Skeletal Text File Import
  - Full Text File Import
  - DwC-Archive Import
  - IPT Import
  - Notes from Nature Import
  - Saved Import Profiles
  - Create a new Import Profile
- Processing Toolbox
- Darwin Core Archive Publishing
- Review/Verify Occurrence Edits
- Duplicate Clustering
- General Maintenance Tasks
  - Data Cleaning Tools
  - Download Backup Data File
  - Restore Backup File
  - Thumbnail Maintenance
  - Update Statistics

These materials were made possible by National Science Foundation Award 1802312. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.