1 Introduction

This document describes our approach to georeferencing—i.e., the process of estimating latitude and longitude (and error) for localities—which follows best practices established by the Global Biodiversity Information Facility [1] and iDigBio [2].

2 Working in refined spreadsheets

If you’re working with spreadsheets generated using the protocols outlined in Frost SOP 12: Digitizing pinned specimens with spreadsheets this section describes steps to georeference those localities.

2.1 Understand your headers

We’ve populated the cells under the purple- and yellow-shaded headers in the Google Sheets spreadsheet: [https://goo.gl/ea6UNw](https://goo.gl/ea6UNw). We’re now going to add data under the green-shaded headers, which are more Darwin Core terms: [3].

- decimalLatitude = The estimated latitude in decimal degrees (e.g., 40.793); positive degrees represent North, while negative values are South
- decimalLongitude = The estimated longitude in decimal degrees (e.g., -77.860); positive degrees represent East, while negative values are West
- coordinateUncertaintyInMeters = Horizontal distance from a georeferenced point; it represents the radius for the smallest circle that contains all of a location
- georeferencedBy = Who obtained the last georeference for a locality
- georeferenceProtocol = Description of methods for obtaining a georeference; in our case we would cite this document as [https://scholarsphere.psu.edu/files/cj82k747v](https://scholarsphere.psu.edu/files/cj82k747v)
- georeferenceRemarks = Comments about specifics deviating from method or assumptions that were made in order to georeference
- geodeticDatum = Datum used to generate georeference; ours will most likely be “WGS84”
2.2 Using GEOLocate

In this process we will use the Web application GEOLocate: [http://www.museum.tulane.edu/geolocate/web/WebGeoref.aspx](http://www.museum.tulane.edu/geolocate/web/WebGeoref.aspx). Open this URL in a browser window. **Keep in mind:** If any elements of the locality are dubious or demonstrably inaccurate the locality should not be georeferenced! Likewise, if the locality cannot be determined with any certainty, or if the specimen is from a lab colony, the locality should be skipped.

**With one spreadsheet directly**  In another browser window open your spreadsheet. Find the “Sort range” option under “Data” in the menu. Sort on countryCode ⇒ stateProvince ⇒ county ⇒ locality (see Figure 1).

![Figure 1: Screen shot of sorting (“Sort range”) function in Google Sheets. Find under “Data” in the main menu](image)

**With OpenRefine**  In some cases we might need to georeference specimens in multiple spreadsheets. For this we will use OpenRefine.

1. Start the application, add your spreadsheets by clicking “choose files”, selecting the relevant files, and clicking “open” then “next >>” (Figure 2).

2. Next click “configure parsing options” and create a project with an informative project name (top right); something like “PSUCarabidae2018” rather than “ARD123” (Figure 3).

3. After clicking “create project” you should see something like in Figure 3.

4. To be efficient about georeferencing we should use OpenRefine’s “facet” method. This process will find all the unique strings in a column, in this case “verbatimLocality”. See Figures 4 and 5.

5. Now run a clustering method to get similar/same localities together; we usually use the key collision/metaphone3 approach. Hovering over a particular cluster in the results will reveal a link to “browse this cluster”. Click that to georeference the selected rows as a unit.
Figure 2: Starting OpenRefine and selecting your files

Figure 3: This project is ready to go!

Figure 4: Creating a text facet for verbatimLocality

Note: Before you begin find these localities in your spreadsheet. We’ve already estimated the coordinates for these common localities in our collection:

- Archbold Biological Station: 27.181438, -81.352017, 3036
Figure 5: The facet appears on the left, and now we can cluster same or similar localities

Figure 6: The metaphone clustering approach usually does well to get like localities together

- **State College**: 40.793395, -77.860001, 3403
- **Black Moshannon State Park**: 40.898391, -78.056396, 2110
- **Scotia Barrens**: 40.801747, -77.943623, 3036
- **Stone Valley**: 40.659788, -77.91639, 1506
- **The Rock, State College**: 40.83994, -77.8276, 1733 (For this locality add a note to georeferenceRemarks: “Locality determined based on conversations with Penn State professor emeritus, Ralph Mumma, who collected with Stuart Frost”
Now georeference your first locality. Go to the GEOLocate window. The Locality String should be the most specific locality information listed on the record; separate the city or recognized area (e.g., a state park) with a comma.

Input country, state, and county (if available it can greatly improve accuracy of the georeference estimate).

Click “Georeference” and retrieve the latitude, longitude, and coordinate uncertainty. See Examples 1 and 2 below for explanations.

Be sure to fill out cells for **georeferencedBy**, **georeferenceProtocol**, **georeferenceRemarks** (if you have anything to add), and **geodeticDatum**.

**Example 1: Well known municipality**  Georeferencing for the locality “Sta. Coll.”—which we’ve interpreted to be US (**countryCode**), Pennsylvania (**stateProvince**), Centre (**county**), State College (**municipality**)—should be pretty straightforward; see Figure 7. Notice that the resultant polygon (in orange) is the entirety of the recognized town of State College. The coordinates of the polygon bounds are in the string in the white text box in the bottom right corner of Figure 7. The latitude, longitude, and error (“40.793395 -77.860001 3403”) are also in this box, above the polygon coordinates. It’s these three numbers that we’ll add to our spreadsheet as **decimalLatitude**, **decimalLongitude**, **coordinateUncertaintyInMeters**. For fun, try georeferencing this locality with nothing entered in “county”. Did you notice a change in accuracy?

**Example 2: Municipality or state without polygon**  São Paulo is both a state and a municipality in Brazil. If the locality simply reads “Brasil, Sao Paulo” how should it be georeferenced? Try searching for each of these strings in GEOLocate: Sao Paulo, São Paulo. Diacritics, when included, can make a huge difference in what is returned. The two red dots in Figure 8 are likely the most appropriate choices. One is the centroid for the state, and the other is the centroid for the city (if it’s known that the specimen was collected in the city). Because no polygons are provided for these estimates you should select the most appropriate location and edit the uncertainty so that the circle includes the entire named place (Figure 9). Be sure to document what you did and why in **georeferenceRemarks**. See [4] and pages 23–32 in [1] for more guidance on how to estimate coordinate uncertainty.
Figure 7: Screen shot of GEOLocate, with Example 1 data added
**Figure 8:** Screen shot of GEOLocate, with Example 2 data added
Figure 9: Screen shot of GEOLocate, with point selected and uncertainty edited
References


