Protocol for georeferencing based on locality data

Authors

Deborah Paul, Erica Krimmel, Nelson Rios, Trevor Dalton, Katelin Pearson, Aja Sherman, Austin Mast

Date last edited

2021-05-07

Goal

Use the GEOLocate Collaborative Georeferencing (CoGe) platform to assign coordinates and uncertainty to locality descriptions using best practices, established efficiencies (e.g., seeding with existing coordinates for collecting localities that can be matched), and technicians familiar with each region's geography.

Relevant fields in the dataset

Data evaluated from

Fields mapped to a "coge" namespace were imported into GEOLocate. Fields not mapped to a "coge" namespace were evaluated during the georeferencing process but viewed in the record view interface of BIOSPEX as needed, rather than within the GEOLocate platform. Due to the structure of the GEOLocate database, some fields present in BIOSPEX were concatenated upon export and others separated upon import. These are indicated by naming fields within brackets in the order of the individual field components, e.g., "[municipality + locality]."

- BIOSPEXid → *coge:CatalogNumber*
- institutionCode → *coge:InstitutionCode*
- scientificName → *coge:ScientificName*
- continent → *coge:ContinentOcean*
- country_rapid → *coge:Country*
- stateProvince → *coge:StateProvince*
- county → *coge:County*
- [islandGroup + island + municipality + locality] → *coge:Locality*
- maximumElevationInMeters → *coge:MaximumElevation*
- minimumElevationInMeters → *coge:MinimumElevation*
- decimalLatitude → *coge:Latitude*
- decimalLongitude → *coge:Longitude*
- coordinatePrecision → *coge:CoordinatePrecision*
- eventDate → *coge:YearCollected*
- locationRemarks
- recordedBy
- recordNumber
- fieldNumber
- georeferencedBy

This protocol was created as part of NSF DBI 2033973, RAPID Grant: Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses. Documents associated with this grant are archived at https://doi.org/10.5281/zenodo.3974999.

- georeferencedDate
- coordinateUncertaintyInMeters
- geodeticDatum
- georeferenceVerificationStatus
- georeferenceRemarks
- georeferenceProtocol
- georeferenceSources
- habitat
- dataGeneralizations
- hasGeospatialIssues
- issue
- idigbio_flags
- verbatimCoordinates
- verbatimCoordinateSystem
- verbatimElevation
- verbatimLatitude
- verbatimLongitude
- verbatimLocality

Enhanced data recorded in

- BIOSPEXid ← *coge:CatalogNumber*
- decimalLatitude_rapid ← *coge:Corrected latitude*
- decimalLongitude_rapid ← *coge:Corrected longitude*
- coordinateUncertaintyInMeters_rapid ← *coge:Corrected uncertainty radius*
- coordinateUncertaintyWKT_rapid ← *coge:Corrected uncertainty radius circular polygon_WKT*
- footprintWKT_rapid ← *coge:Corrected uncertainty polygon_WKT*
- coordinatePrecision_rapid
- geodeticDatum_rapid
- georeferencedBy_rapid ← *coge:Verified by*
- georeferenedByID_rapid
- georeferencedDate_rapid ← *coge:Date verified*
- georeferenceProtocol_rapid
- [georeferenceRemarks_rapid + georeferenceSources_rapid] ← *coge:Verification remarks*
- georeferenceVerificationStatus_rapid ← *coge:Verification type*
- flagGeoreference_rapid

Process & Parties Responsible

The **first stage** of this process is completed by the System Administrator in BIOSPEX.

1. Prepare export of data from BIOSPEX containing fields as determined by “Data evaluated from” (above, those fields indicated with a mapping to “coge” namespace field). The export contains records that have already been georeferenced because these can provide context for georeferencing similar localities.

2. Exclude rows where the record has been determined inappropriate for georeferencing, based on the following criteria:
    a. Information in *dataGeneralizations* or *informationWithheld* indicates that existing coordinate data or other specific locality information is available upon request but is not present in the current dataset.
    b. A data provider has informed this project that their published locality information does not reflect the full amount of information available.
3. Resolve any other known issues in the data that are specific to a data provider.
4. Provide CSV file for upload to GEOLocate (see GEOLocate documentation at http://www.geo-locate.org/community/coge_import.html).

The **second stage** of this process is completed by key collaborators associated with GEOLocate.

5. Create a community for this project in CoGe at https://coge.geo-locate.org.
6. Upload the CSV file provided above (Step #4) to the new CoGe community.
7. Pre-process records using GEOLocate's capacity for automatic georeferencing.

The **third stage** of this process is completed by the Data Curators in GEOLocate.

8. Log into the CoGe data management portal (https://coge.geo-locate.org) and claim a subset of records (i.e., one or several countries) by clicking *Members*, then *Define users working dataset*, and then selecting the desired search criteria and the user to which the criteria will be applied. Exit the CoGe platform.
9. Log into the CoGe georeferencing web app (http://www.geo-locate.org/web/WebComGeoref.aspx) and click "Continue" after confirming that the information listed under *Available communities* and *Data sources in selected community* matches the criteria that you claimed in the CoGe data management portal.
10. Begin georeferencing by clicking *Next Record(s)*.
11. For a given locality record, review anything listed under *Similar Records* and *Identical Records*, and check the box next to each that should be considered the same functional locality as what is highlighted in yellow at the top of the workbench. The same coordinates and metadata will be assigned to every specimen record that is checked in the workbench.
12. Spend up to 15 minutes attempting to assign coordinates to the locality according to the Georeferencing Quick Reference Guide (Zermoglio et al., 2020) and other guidelines provided in documents in the Resources section below. Make sure to evaluate specimen record information (see "Data evaluated from" above) available in BIOSPEX, and use external gazetteers and other online resources where additional research is required. Where feasible, localities with a spatial resolution more specific than the level of country should be georeferenced even when the uncertainty radius will be very large (e.g., >10000 meters). If you are able to determine coordinates, ensure that the following information is captured:
    a. Latitude, longitude, and uncertainty in the *Calculated Coordinates* box.

This protocol was created as part of NSF DBI 2033973, RAPID Grant: Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses. Documents associated with this grant are archived at https://doi.org/10.5281/zenodo.3974999.

        i. Note that the smallest acceptable uncertainty for this project is 30 meters. If the record has existing coordinates and indicates that they were acquired from a GPS unit, use 30 meters as the uncertainty radius.
    b. If appropriate, an uncertainty polygon (see FAQ).
    c. Information for *georeferenceRemarks_rapid* and *georeferenceSources_rapid*, in that order and separated by a semicolon, in the *Add Comments* box.
        i. *georeferenceRemarks_rapid* should include any notes or comments about the spatial description determination, explaining assumptions made in addition or opposition to those formalized in the Georeferencing Quick Reference Guide.
        ii. *georeferenceSources_rapid* should include a concatenated list of maps, gazetteers, or other resources used to georeference the locality, described specifically enough to allow anyone in the future to use the same resources. Separate distinct values with a vertical bar (|). If only the GEOLocate platform was used, no information is necessary to record here.

13. For any locality requiring more than approximately 15 minutes of research, skip the record by clicking *Skip selected* and entering one of the following reasons, per the Georeferencing Quick Reference Guide:
    a. “dubious” (see 2.4.1 Dubious Locations)
    b. “cant-find” (see 2.4.2 Cannot be Located)
    c. “multi-near” (see 2.4.3.1. Multiple Related Nearby Features)
    d. “multi-far” (see 2.4.3.2. Multiple Unrelated Features)
    e. “contradiction” (see 2.4.4 Demonstrably Inconsistent)
    f. “captive” (see 2.4.5 Cultivated or Captive)
    g. “other” (please explain)
14. Periodically review records that were previously skipped via the *Review* tab, and always feel free to return to and refine past georeferences if new knowledge is gained.

The **fourth stage** of this process is completed in GEOLocate concurrently with stage three.

15. At the point where a Data Curator finishes georeferencing all of the localities within a country, they confirm that any skipped localities do indeed need to be skipped. They also return to the first 10 locality records completed for the country and review these for quality control.
16. Data Curators periodically review 10 locality records georeferenced by each of the other Data Curators to confirm that the protocol is being applied consistently.
17. Senior Personnel and the Digitization Specialist periodically review random subsamples of new georeferences via the *Review* tab to check for metadata consistency and completeness.
18. Project team track georeferencing progress using statistics available in the CoGe data management portal.

The **fifth stage** of this process is completed by the Digitization Specialist in OpenRefine and System Administrator in BIOSPEX, with input from others as needed.

19. Export data from GEOLocate via the CoGe data management portal and the following settings:

a. Select fields to download: InstitutionCode (mandatory), CatalogNumber (mandatory), ScientificName (mandatory).
b. Delimited text: CSV.
c. Include polygons, all sizes, in WKT format.
d. Download by specimen records.
e. Include skips, include corrections, include unprocessed: Check all boxes.
f. Restrict to latest work.

20. Import the data into OpenRefine, rename all fields per mappings described in “Enhanced data recorded in” above, and add or edit project-specific data based on the following guidelines:

a. Populate the field *georeferenceProtocol_rapid* with the value “Zermoglio PF, Chapman AD, Wieczorek JR, Luna MC & Bloom DA (2020) Georeferencing Quick Reference Guide [Community review draft]. Copenhagen: GBIF Secretariat. https://doi.org/10.35035/e09p-h128, as modified by research project NSF DBI 2033973 (https://doi.org/10.5281/zenodo.3974999)” for any row that was either georeferenced or skipped as part of this protocol.
b. Reassign values in the field *georeferenceVerificationStatus_rapid* by replacing “corrected” or “skipped” with “verified by research project NSF DBI 2033973 (https://doi.org/10.5281/zenodo.3974999)”.
c. Reassign values in the field *georeferencedBy_rapid* by replacing GEOLocate usernames with full names, e.g., replace “kdpearso” with “Katelin D. Pearson”.
d. Populate the field *georeferencedByID_rapid* with an ORCID ID determined by the value of *georeferencedBy_rapid*.
e. Translate dates in the field *georeferencedDate_rapid* into YYYY-MM-DD format per the ISO 8601, e.g., replace “10/1/2020 18:30:28” with “2020-10-01”.
f. Delete value “N/A” from the fields *decimalLatitude_rapid* and *decimalLongitude_rapid*.
g. Populate the field *geodeticDatum_rapid* with the value “WGS84” for any row for which coordinates were assigned as part of this protocol.
h. Populate the field *coordinatePrecision_rapid* based on the coordinate values present in *decimalLatitude_rapid* and *decimalLongitude_rapid*. Where the precision of the latitude and longitude differ, use the more precise value to assign a value to *coordinatePrecision_rapid*. Use the following rules to assign this value, and do not truncate or round the coordinate values themselves in any circumstance:
    i. If the most precise coordinate value has a precision more specific than “0.0001” (e.g., “42.89775”) assign the value “0.0001” to *coordinatePrecision_rapid*. This project does not assert confidence in precision any more specific than this.
    ii. If the most precise coordinate value has a precision equal to or less than “0.0001” (e.g., “42.8977” or “38.6”) assign a value to *coordinatePrecision_rapid* based on this level of precision (e.g., “0.0001” or “0.1”, respectively).
    iii. If the values for both *decimalLatitude_rapid* and *decimalLongitude_rapid* are integers, assign a value of “1” to *coordinatePrecision_rapid*.
i. Delete value “N/A” from the field *coordinateUncertaintyInMeters_rapid*. Ensure that there are no zero values present. If zero values are present, an uncertainty must be determined and added.
j. Delete value “N/A” from the field *footprintWKT_rapid*.

   k. Translate and clean up the values in *georeferenceRemarks_rapid*. For records where the locality was skipped (i.e., no coordinates were assigned), replace shorthand keywords based on the guidelines below. Always retain any specific details recorded by the georeferencer in addition to the reason for skipping.
      i. Replace "dubious" with "Unable to georeference, see Georeferencing Quick Reference Guide (2020) 2.4.1 Dubious Locations for rationale."
      ii. Replace "cant-find" with "Unable to georeference, see Georeferencing Quick Reference Guide (2020) 2.4.2 Cannot be Located for rationale."
      iii. Replace "multi-near" with "Unable to georeference, see Georeferencing Quick Reference Guide (2020) 2.4.3.1 Multiple Related Nearby Features for rationale."
      iv. Replace "multi-far" with "Unable to georeference, see Georeferencing Quick Reference Guide (2020) 2.4.3.2 Multiple Unrelated Features for rationale."
      v. Replace "contradiction" with "Unable to georeference, see Georeferencing Quick Reference Guide (2020) 2.4.4 Demonstrably Inconsistent for rationale."
      vi. Replace "captive" with "Unable to georeference, see Georeferencing Quick Reference Guide (2020) 2.4.5 Cultivated or Captive for rationale."
      vii. Replace "other" with "Unable to georeference."
   l. Clean up values in *georeferenceSources_rapid* as appropriate. Confirm that sources are cited explicitly wherever possible; see FAQ below for examples. Add GEOLocate as a source to every record using the citation, "Rios, N. 2020. GEOLocate software for georeferencing natural history data. Computer program and documentation distributed by the author, website: http://www.geo-locate.org".
   m. Delete the fields *InstitutionCode* and *ScientificName*. You should be left with only the *BIOSPEXid* field and those whose names end in "_rapid."

21. Export data from OpenRefine as a CSV file and import this into BIOSPEX.

The **sixth stage** of this process is completed by the Digitization Specialist in OpenRefine and System Administrator in BIOSPEX, with input from others as needed.

22. Review the entire dataset from BIOSPEX and identify rows where the project was unable to georeference the collecting locality. Where legacy data exists in the fields *decimalLatitude* and *decimalLongitude*, do the following:
    a. Migrate information from the following fields into their *_rapid* versions: *decimalLatitude, decimalLongitude, coordinateUncertaintyInMeters, footprintWKT, coordinatePrecision, geodeticDatum, georeferencedBy, georeferencedDate, georeferenceProtocol, georeferenceRemarks,* and *georeferenceSources.* For example, if a value exists in *georeferencedDate_gbifP*, copy this value into *georeferencedDate_rapid*. Prefer data sources in the order of *_gbifP, _gbifR, _idbR, _idbP*.
    b. Add the standard value, "unverified by research project NSF DBI 2033973 (https://doi.org/10.5281/zenodo.3974999)" to the field *georeferenceVerificationStatus_rapid*.
23. Where legacy data do not exist in the fields *decimalLatitude* and *decimalLongitude*, add the standard value, "not enough locality information to be assessed by research project NSF DBI 2033973 (https://doi.org/10.5281/zenodo.3974999)" to the field *georeferenceVerificationStatus_rapid*.
24. Assign values to the field *flagGeoreference_rapid* using the following guidelines:

a. NEW = no previous coordinates existed and coordinates were assigned as part of this protocol
b. REVIEWED_ALTERED = previous coordinates existed, but the coordinates or error radius were modified as part of this protocol
c. REVIEWED_RETAINED = previous coordinates existed, were reviewed as part of this protocol, and were left unchanged
d. REVIEWED_DISCARDED = previous coordinates existed, but these were discarded upon review as part of this protocol
e. LEFT_BLANK = no previous coordinates existed and no coordinates could be assigned as part of this protocol
f. UNREVIEWED_RETAINED = previous coordinates existed, but could not be reviewed as part of this protocol due to lack of supporting locality information, and so coordinates were left unchanged

25. Export data from OpenRefine as a CSV file and import this into BIOSPEX.

Communication

Questions and discussion about this protocol or work related to it can be posed in the FSU iDigBio Slack #georeferencing channel.

Resources

The following external resources are essential for Data Curators to review:

- Chapman AD & Wieczorek JR (2020) Georeferencing Best Practices [Community review draft]. Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc-gg7h-s853
- Zermoglio PF, Chapman AD, Wieczorek JR, Luna MC & Bloom DA (2020) Georeferencing Quick Reference Guide [Community review draft]. Copenhagen: GBIF Secretariat. https://doi.org/10.35035/e09p-h128
- Darwin Core locality class terms and definitions: https://dwc.tdwg.org/terms/#location

The following external resources may also be of interest to Data Curators:

- Georeferencing resources (including nice visuals) from the California Phenology Thematic Collections Network: https://www.capturingcaliforniasflowers.org/georeferencing-protocols-and-guides.html
- Florida State University collaborative georeferencing protocol: https://www.idigbio.org/wiki/images/8/8c/FSU_Collaborative_Georeferencing_Protocol.pdf
- Bionomia, a tool to link natural history specimens to the world's collectors: https://bionomia.net/
- Wieczorek J, Guo Q & Hijmans RJ (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18(8):745-767.
- Links to additional tools on the Georeferencing Online Mapping Resource Hub: http://georeferencing.org/
- Georeferencing workflows assembled by the paleo collections community: https://tdwg.github.io/esp/georeferencing/workflows.html
- Falling Rain Directory of Cities and Towns in the World: http://www.fallingrain.com/world/
- Presentation of additional internet resources for georeferencing: https://www.idigbio.org/sites/default/files/workshop-presentations/georef-research-use/InternetResources2016.pptx

This protocol was created as part of NSF DBI 2033973, RAPID Grant: Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses. Documents associated with this grant are archived at https://doi.org/10.5281/zenodo.3974999.

FAQ

This section serves as a space in which to document questions that arose as part of the georeferencing process, as well as our collective answers for them. In contrast to the Slack communication channel, this FAQ will be archived as a part of this georeferencing protocol. Conversations that start in Slack may be appropriate to synthesize here for the purpose of documentation.

What is the best way to deal with localities that are geographic features with a very large uncertainty radius?

Place your point at the visual center of the feature and draw your uncertainty radius to encompass the full extent of the feature.

Where should the point be placed when georeferencing a cave?

In most cases, the point should be placed at the mouth of the main cave entrance, with an uncertainty radius drawn to include the extent of the cave, if known. If a more specific location within the cave is known, the point should be placed as accurately as this information allows (e.g., “entrance” or “mouth of cave”). In some cases, drawing a polygon that reflects the extent of the cave is appropriate. See *2.1.3.5. Feature – Cave* of the Georeferencing Quick Reference Guide for more.

In a situation where a city and a nearby cave share the same base name, should we assume that the collecting locality meant the cave even if “cave” is not specifically mentioned in it? Should we incorporate the cave in the uncertainty radius?

Not all bats live in caves, so do not assume that a locality refers to a cave unless it specifically mentions “cave.” If your point is centered on a city, you do not need to include a nearby cave in your uncertainty just because the cave shares a base name with the city.

What, if any, special treatment should we give to localities with a legacy georeference, i.e., that are coming into this project having already been georeferenced?

For the most part, treat these as you would any other locality. Ideally the legacy georeference is accurate and its existence will save you time, but if it is not accurate you should correct the georeference as you would with any locality in GEOLocate. At the very least, you will need to redetermine the uncertainty radius for a legacy georeference. You are not overwriting any of the original record’s data, and you do not need to add a remark that you accepted a legacy georeference. However, if you find that a legacy georeference needs to be significantly corrected, this would be useful to add a remark about.

What sources are appropriate to cite for georeferenceSources?

You never need to cite GEOLocate itself, because this will be added in bulk when we export data and import it back into the main dataset. You also do not need to cite specific base maps within GEOLocate unless the information on one base map was unique and particularly useful. In general, you should cite any external resources that you used to determine coordinates for the locality, using descriptive text in addition to or in place of web URLs where that is reasonable. The following are examples of sources you might cite:

- A particular map, e.g., “USGS 1:24000 Florence Montana Quad 1967”
- A particular base map accessed as a GEOLocate layer, e.g., “GEOLocate Mapnik (OSM) layer”

- An external gazetteer, e.g., “PhilAtlas (www.philatlas.com), Tuyan”
- PhilAtlas. Available: https://www.philatlas.com/visayas/r07/negros-oriental/manjuyod/candabong.html
- A publication, e.g., “Strinati, Pierre. 1953. Une grotte chaude près d' Alhama de Murcia. Revista de Ciencias III(1), 91-100. doi: http://hdl.handle.net/10651/30586”
- A website, e.g., “https://www.wikiloc.com/hiking-trails/sierra-de-la-camorra-mollina-malaga-21589346”
- Falling Rain Global Gazetteer Version 2.3. Available: http://fallingrain.com/world/EK/00/Aman.html

Many of the place names in our localities are misspelled or outdated. Should we update these?

You do not need to worry about updating misspellings or outdated place names; however, if you reference the place name in your remarks, you should use the correct spelling. It may also be helpful to comment on outdated place names in the remarks if this affected your process for determining where to place the point, e.g., if you had to look up the modern equivalent for an outdated name.

If two distinct features (that are located significantly apart from one another) are listed in a locality description, where should you place the point?

If you are certain that one of the towns is not a broader political boundary (e.g., province, county), place the point in the center between the two named features and extend the error radius to the centers of both the named features. An exception to this is when the first named place is a large city and the second place is a village or otherwise small location. In this case, assume the village or smaller place is the more specific location and georeference accordingly.

What kinds of localities are appropriate to draw a polygon for?

Polygons are useful for recording the extent of features that will have an otherwise disproportionately large uncertainty radius. For example, a river or road has a clear extent and its linear shape means that the uncertainty, as determined by a radius, will include a significant area that is clearly outside of the river’s extent.

I came across a locality that I have already georeferenced, how can I assign the same coordinates, uncertainty, and (if appropriate) polygon to it?

Occasionally, out of caution for being accurate, GEOLocate won’t group together localities that are actually the same, which means you might come across a locality that you have already georeferenced. In this scenario, find the locality record that you previously georeferenced use it to assign the exact same coordinates, uncertainty, and polygon (if appropriate) to the new locality record. Find the previously georeferenced locality in your *History* tab and click on the text description. A yellow dot will appear on the map to show you the location of this point. Click on the yellow dot, which should then turn green. Click on the *Workbench* tab and you will see the green dot here as well. Clicking the *Correct* button will assign the coordinates, uncertainty, and polygon from the locality in your history to the locality on your workbench. The comments do not get automatically assigned, so if they are important make sure to copy them from the history locality (a button *view comment* will appear on the right for localities that have comments) into the workbench locality.

What is the best way to record uncertainty for a locality that is described by measurements from a reference point, e.g., “4.8 km west and 5.2 km south of Mt. Snow.”

This type of locality is described in the Quick Reference Guide under 2.2.4. Offset – Distance along Orthogonal Directions and further directions on calculating the uncertainty for this type of locality are included in the

Georeferencing Best Practices, 3.4.6. Uncertainty Related to Offset Precision. You may wish to use the georeferencing calculator for this type of locality in order to save yourself from having to do calculations yourself. That said, you might have some localities where the description allows you to freehand draw an uncertainty that is possibly more accurate than what can be calculated. For instance, if the locality was something e.g., “4.8 km west and 5.2 km south of Mt. Snow, near a small stream" then it probably makes more sense for you to base the uncertainty more on the added feature of the small stream than the recommendations of the calculator.

Results

The first stage of this task was completed by System Administrator Robert Bruhn on 2020-09-25 with assistance from Digitization Specialist Erica Krimmel. Concatenation for fields needing this was done in R. As instructed in Step #2 (above), values in 68 rows for *informationWithheld* and 121 rows for *dataGeneralizations* were classified as “inappropriate for georeferencing.” However, given that this amount was inconsequential to the total amount of localities being georeferenced (~0.2%), these rows were not removed from the dataset. Only one provider-specific issue was identified for Step #3; locality information for specimens from the Australian Museum were provided in the *localityRemarks* field instead of the expected *locality* field. For the 1,866 rows affected by this situation, the values for *localityRemarks* were translocated into *locality*.

An initial CSV file was given to key collaborator Nelson Rios for upload to GEOLocate (Step #4) on 2020-09-25. This dataset consisted of 919 records in which the locality was in Spain, to be used as an introductory dataset for Data Curators to become comfortable with the CoGe platform. Of these localities, 706 had values in the *locality* field and were georeferenced by Data Curators Trevor Dalton, Katelin Pearson, and Aja Sherman. Data for the remaining 56,879 records were given to Nelson Rios on 2020-09-29 (Step #4) and the second stage was completed the same day.

The third stage of this protocol was in progress from 2020-09-28 to 2021-01-06 and took approximately 530 hours of work. Data Curators reviewed a total of 59,112 records (65.8% of the project’s total 89,837 records) that had locality information, and of these, they were able to verify or assign coordinates to 56,203 records (95.1%). Data curators varied in rate of georeferencing from 100-132 records/hr. Data Curators worked country-by-country, as described above (Step #10). Specimens for which the country was undetermined were tackled last. This proved to be a useful strategy because Data Curators were able to call up previously-georeferenced records from their history and apply those georeferences to some of the undetermined records. In Step #12a, three records were assigned an uncertainty of less than 30 meters based on the recommendation of the Georeferencing Calculator. Otherwise, records were georeferenced to the following levels of uncertainty:

- >1,000,000 m = 42 records
- 100,000 - 1,000,000 m = 4,737 records
- 10,000 - 999,999 m = 10,383 records
- 1,000 - 9,999 m = 29,521 records
- 100 - 999 m = 8,235 records
- 31 - 99 m = 1,027 records
- 30 m = 2,258 records

This protocol was created as part of NSF DBI 2033973, RAPID Grant: Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses. Documents associated with this grant are archived at https://doi.org/10.5281/zenodo.3974999.

2,909 records (4.9% of the records with locality information) could not be georeferenced for the following reasons (Step #13). Note that a single record may reference more than one of these reasons.

- 357 records per the Georeferencing Quick Reference Guide (2020) 2.4.1 Dubious Locations
- 1,262 records per the Georeferencing Quick Reference Guide (2020) 2.4.2 Cannot be Located
- 93 records per the Georeferencing Quick Reference Guide (2020) 2.4.3.1 Multiple Related Nearby Features
- 211 records per the Georeferencing Quick Reference Guide (2020) 2.4.3.2 Multiple Unrelated Features
- 371 records per the Georeferencing Quick Reference Guide (2020) 2.4.4 Demonstrably Inconsistent
- 0 records per the Georeferencing Quick Reference Guide (2020) 2.4.5 Cultivated or Captive
- 200 records because the locality was too big to assign any sort of meaningful geographic centroid to.
- 420 records did not include a reason for not georeferencing.

The *flagGeoreference_rapid* values were applied to the following numbers of specimens (Step #24).

- NEW: 21,089 records
- REVIEWED_ALTERED: 25,586 records
- REVIEWED_RETAINED: 9,516 records
- REVIEWED_DISCARDED: 949 records
- LEFT_BLANK: 31,738 records
- UNREVIEWED_RETAINED: 942 records

The following are conventions, assumptions, or helpful notes collated by the Data Curators during the third stage of this protocol:

- When georeferencing localities in Chinese-speaking countries (China, Hong Kong, Taiwan), I searched for the locality as is and then converted into an alternative romanized version (i.e., Wade-Giles vs. pinyin) of the verbatim locality. I used the converter available at https://libraries.indiana.edu/chinese-studies-pinyin-wade-giles-conversion-table. For example, “Yibin” in pinyin is equivalent to “Ipin” in Wade-Giles.
- If verbatim coordinates fell out in an ocean/major body of water, I adjusted the coordinates to be on land closest to the provided point. This mirrors the recommendation in the Georeferencing Quick Reference Guide for assigning coordinates de novo.
- When a locality was comprised of only a single value (no state/province or county indicated), and the locality could be applied to either a major city or a province/state/county, I placed the point at the center of the city and set the error radius to encompass the whole province/state/county. This is indicated in the *georeferenceRemarks_rapid* field.
- I often estimated the extent of a location from Google Satellite, e.g., by visually estimating the extent of a heavily-populated area of a city or a forested region in the case of a forest reserve with unclear boundaries on all other map layers. This is indicated in the *georeferenceRemarks_rapid* field.
- When the extent of a feature was unclear, I set the error radius as halfway to another major feature/city/landmark so as to create a relatively large error radius. For example, many South African localities were farms, which have unclear extents (and satellite view shows farmland all around). I would place the point at the center of the farm (or often, the point provided by the 2018 African Chiroptera Report) and set an error radius to the next major town or the nearest town that would require the largest error radius. For example, if town A is 5 mi N, town B is 10 mi S, town C is 15 mi W, and town D is 16 mi W, I would set the error radius halfway to town D, since the farm could extend in the

direction of town D until that town. Specific details are recorded in the *georeferenceRemarks_rapid* field.

- Specimens belonging to the South Africa National Biodiversity Institute (SANBI) often did not have "country" populated. These specimens were assumed to be from South Africa unless the locality was clearly in another country. For example, "Zimbabwe Ruins" were located in Zimbabwe.
- SANBI specimens also appeared to be missing some information from the locality description, so I interpreted them as if the information were there. For example, "32 km ENE, De Hoop Nat. Res. Guano Cave" would ordinarily be interpreted as 32 km ENE of the Guano Cave in De Hoop Nature Reserve. However, I noticed that the African Chiroptera Report listed this locality as "De Hoop Nature Reserve, 32 km ENE Bredasdorp". It seems that the "Bredasdorp" was somehow dropped from the locality description. This may seem like a large assumption, but this happened consistently with SANBI specimens. A specific place (like a farm or nature reserve) would be mentioned along with a distance that, according to the ACR, should be accompanied by another location name (e.g., "Farm 230 (Wagondrift), 16 km SSE" corresponds to "Farm 230 (Wagondrift), 16 km SSE Lamberts Bay").
- I used Google Translate to assist with localities that appeared to not be locations or appeared to be in another language. For example, to determine that "air pana" means hot springs in Malay.

The fifth stage of this protocol was completed by Digitization Specialist Erica Krimmel on 2021-02-26 and took approximately 25 hours of work, the majority of which was spent standardizing and performing quality control on the values in *georeferencingRemarks_rapid* and *georeferencingSources_rapid* (Step #20k–l, above).

The sixth stage of this protocol was completed by Digitization Specialist Erica Krimmel on 2021-03-26 and took approximately 1 hour of work. Legacy georeference data for 951 rows was present and migrated into *_rapid* fields and assigned the *flagGeoreference_rapid* status of UNREVIEWED_RETAINED.

This protocol was created as part of NSF DBI 2033973, RAPID Grant: Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses. Documents associated with this grant are archived at https://doi.org/10.5281/zenodo.3974999.