

# Protocol for compiling raw data from multiple sources and deduplicating records

## Authors

Erica Krimmel, Austin Mast

## Date last edited

2020-09-23

## Goal

Compile raw data from multiple sources and deduplicate records to create an initial dataset which will be enhanced throughout the course of this project.

## Relevant fields in the dataset

### Enhanced data recorded in

- gbif
- idigbio

## Process & Parties Responsible

This process will be completed entirely by Erica Krimmel on the iDigBio and GBIF websites, and in R, following these steps:

1. Acquire specimen occurrence data from iDigBio by searching in the Portal (<https://www.idigbio.org/portal/search>) for *Basis of record* = “preservedspecimen” and *Order* = “chiroptera” and downloading the results.
2. Acquire specimen occurrence data from GBIF by searching in the user interface (<https://www.gbif.org/occurrence/search>) for *Basis of record* = “Preserved specimen” and *Family* = “Hipposideridae” or *Family* = “Rhinolophidae” or *Family* = “Rhinonycteridae” and downloading the results.
3. Prepare iDigBio data in R:
  - a. Import both the raw occurrence data and the occurrence data processed by iDigBio during ingestion.
  - b. Add a suffix to field names to indicate data source (“\_idbR” for raw and “\_idbP” for processed).
  - c. Subset data to include only records representing a specimen identified as being in one of our families of interest (Hipposideridae, Rhinolophidae, or Rhinonycteridae).
  - d. Split the field *idigbio\_geoPoint\_idbP* into two new fields, *decimalLatitude\_idbP* and *decimalLongitude\_idbP* to match the format of the iDigBio raw and GBIF raw and processed data.

This protocol was created as part of [NSF DBI 2033973](https://www.nsf.gov/awardsearch/showAward?AWDNO=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>.

4. Prepare GBIF data in R:
  - a. Import both the raw occurrence data and the occurrence data processed by GBIF during ingestion.
  - b. Add a suffix to field names to indicate data source (“\_gbifR” for raw and “\_gbifP” for processed).
  - c. Exclude records published as part of a “checklist” dataset, because these are typically based on literature referring to specimens, not directly linked to specimens.
5. Compile data from iDigBio and from GBIF into a single dataset, deduplicating specimen records based on matches where the combination of *institutionCode*, *collectionCode*, *catalogNumber*, and *occurrenceID* is the same in both the GBIF raw and iDigBio raw data.
6. Record the provenance of each record in the fields *idigbio* (“1” = record was present in iDigBio data; “0” = record was absent in iDigBio data) and *gbif* (“1” = record was present in GBIF data; “0” = record was absent in GBIF data).
7. Save this primary records dataset as a CSV.

## Communication

Communication for this task will be via weekly team meetings.

## Results

This task was completed by Erica Krimmel on 2020-09-23 and took a total of approximately 31 hours, including the initial data exploration required to write effective code. Step #1 (above) iDigBio search retrieved 829,063 records; the decision to search based on order versus the specific families of interest to this project was because of a concern that iDigBio taxonomic data may be slightly out of date. Step #2 GBIF searches resulted in 40,101 (Hipposideridae) and 40,404 (Rhinolophidae) records. Although Rhinonycteridae is also a family of interest there are no records in GBIF under this taxonomy so no data were downloaded. The raw data from Steps #1 and #2 were archived on Zenodo in the form of three Darwin Core Archives (one from iDigBio and two from GBIF). Step #3c reduced the number of iDigBio records to 59,122. Step #4c reduced the number of GBIF records from 80,505 to 79,855 by excluding 650 records from 25 GBIF “checklist” datasets. Step #5 resulted in a dataset of 89,837 records, of which (Step #6) 9,982 were only present in iDigBio, 30,715 were only present in GBIF, and 49,140 were present in both iDigBio and GBIF. There are almost certainly instances where our deduplication criteria of *institutionCode* + *collectionCode* + *catalogNumber* + *occurrenceID* was too strict and two records representing the same physical specimen were not merged. Given the potential complexity of solving this problem, we determined to just be conservative in asserting that two records were duplicates.

Code associated with this protocol can be found in ‘RAPID-code\_compile-deduplicate.R’ (archived at <https://doi.org/10.5281/zenodo.3974999>).

This protocol was created as part of [NSF DBI 2033973](https://doi.org/10.17863/nsf.2023973), 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>.