Protocol for standardizing and flagging taxonomic name data

Authors

Erica Krimmel, Aja Sherman, Nancy Simmons, Austin Mast

Date last edited

2021-04-23

Goal

Flag currently unaccepted names and suggest a currently accepted synonym where possible. Flag specimen records that would represent geographic outliers to a species' distribution were the taxonomic identification to be correct. This project will not attempt to verify the taxonomic identification of any specimens because that is best performed onsite at the collections curating the specimens.

Relevant fields in the dataset

Where fields do not have a suffix (e.g., *collectionCode*), data from all sources were coalesced into a single column. Where fields have a suffix (e.g. *scientificName_gbifR*), the suffix indicates the source of the data. Sources have been kept distinct in situations in which the values of the fields differ, primarily as a result of different processing methods for standardizing taxonomic information. Data sources are GBIF raw (*_gbifR*), GBIF processed (*_gbifP*), iDigBio raw (*_idbR*), and iDigBio processed (*_idbP*); see document 'RAPID-protocol_compile-deduplicate.pdf' for an explanation of these data sources.

Data evaluated from

- BIOSPEXid
- institutionCode
- collectionCode
- catalogNumber
- family_gbifR / family_gbifP / family_idbR / family_idbP
- genus_gbifR / genus_gbifP / genus_idbR / genus_idbP
- subgenus_gbifR / subgenus_gbifP / subgenus_idbP
- specificEpithet_gbifR / specificEpithet_gbifP / specificEpithet_idbR / specificEpithet_idbP
- infraspecificEpithet_gbifR / infraspecificEpithet_gbifP / infraspecificEpithet_idbR / infraspecificEpithet_idbP
- scientificNameAuthorship_gbifR / scientificNameAuthorship_idbR
- taxonRank_gbifR / taxonRank_gbifP / taxonRank_idbR / taxonRank_idbP
- scientificName_gbifR / scientificName_gbifP / scientificName_idbR / scientificName_idbP

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.

- typeStatus_gbifR / typeStatus_gbifP / typeStatus_idbR
- taxonomicStatus_gbifP / taxonomicStatus_idbP
- acceptedScientificName_gbifP
- taxonConceptID
- nameAccordingTo
- identificationQualifier
- identifiedBy
- identifiedByID
- identificationVerificationStatus
- identificationRemarks
- previousIdentifications
- country_rapid
- dataGeneralizations
- issue
- idigbio_flags

Enhanced data recorded in

- kingdom_rapid
- phylum_rapid
- class_rapid
- order_rapid
- family_rapid
- genus_rapid
- specificEpithet_rapid
- infraspecificEpithet_rapid
- scientificName_rapid
- scientificNameAuthorship_rapid
- flagTaxonomy_rapid
- taxonRank_rapid
- nameAccordingTo_rapid
- identificationQualifier_rapid
- identificationRemarks_rapid
- identificationVerificationStatus_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).

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.

The second stage of this process involves manipulating the data to improve the efficiency of the third stage, and is completed by the Digitization Specialist in OpenRefine.

2. Standardize format between GBIF and iDigBio data by uppercasing values in the following fields: *taxonRank_gbifR, taxonRank_gbifP, taxonRank_idbP, taxonRank_idbR, taxonomicStatus_gbifP, taxonomicStatus_idbP*.
3. Remove fields in which no data are present in any row.
4. Concatenate fields as follows:
    a. *concatenatedName_gbifR* = *genus_gbifR* + *subgenus_gbifR* + *specificEpithet_gbifR* + *infraspecificEpithet_gbifR* + *scientificNameAuthorship_gbifR*
    b. *concatenatedName_gbifP* = *genus_gbifP* + *subgenus_gbifP* + *specificEpithet_gbifP* + *infraspecificEpithet_gbifP*
    c. *concatenatedName_idbR* = *genus_idbR* + *specificEpithet_idbR* + *infraspecificEpithet_idbR* + *scientificNameAuthorship_idbR*
    d. *concatenatedName_idbR* = *genus_idbP* + *subgenus_idbP* + *specificEpithet_idbP* + *infraspecificEpithet_idbP*
5. Rename the following fields to indicate that values will be enhanced as part of this protocol:
    a. *nameAccordingTo* → *nameAccordingTo_rapid*
    b. *identificationQualifier* → *identificationQualifier_rapid*
6. Create a new *family_rapid* field populated with values from *family_gbifP* or, for rows in which that field is blank, *family_idbR*.
7. Create a new *scientificName_rapid* field populated with values from *acceptedScientificName_gbif* or, where rows in that field are blank, *scientificName_idbR*.
8. Create a new *identificationRemarks_rapid* field.
9. Create a new *identificationVerificationStatus_rapid* field.
10. Create a new *BIOSPEXlink* field by concatenating the BIOSPEX base URL to the value in *BIOSPEXid* so that a user can easily link to the specimen record in BIOSPEX.
11. Sort data by *scientificName_rapid*, then *country_rapid*, then *institutionCode*, then *catalogNumber*.
12. Export data from OpenRefine as a CSV file and import into Google Sheets.

The third stage of this process involves evaluating the validity of taxonomic names present in the data, and is completed by the Data Curators in Google Sheets with input from key collaborators as needed.

13. Filter the Google sheet by *scientificName_rapid* to reduce the amount of records in view, recognizing that variations of the same taxa can be selected simultaneously (e.g. "Hipposideros abae," "Hipposideros abae J.A.Allen, 1917," and "Hipposideros aba").
14. For each row, evaluate the taxon concepts presented by the data as described below. Where taxon concepts are not congruent, evaluate the row using information for the

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.

lowest rank, e.g., a scientific name at the rank of subspecies, and preferring information directly from the data provider (i.e., in the raw data fields) vs. that processed by the aggregators. Taxon concepts to be evaluated follow:

   a. Raw data from GBIF (i.e., *family_gbifR* + *concatenatedName_gbifR* + *scientificName_gbifR* + *taxonRank_gbifR* + *typeStatus_gbifR*).
   b. Processed data from GBIF (i.e., *family_gbifP* + *concatenatedName_gbifP* + *scientificName_gbifP* + *acceptedScientificName_gbifP* + *taxonRank_gbifP* + *taxonomicStatus_gbifP* + *typeStatus_gbifP*).
   c. Raw data from iDigBio (i.e., *family_idbR* + *concatenatedName_idbR* + *scientificName_idbR* + *taxonRank_idbR* + *typeStatus_idbR*).
   d. Processed data from iDigBio (i.e., *family_idbP* + *concatenatedName_idbP* + *scientificName_idbP* + *taxonRank_idbP* + *taxonomicStatus_idbP*).

15. For each row, holistically review information in the following fields: *country_rapid, identificationQualifier_rapid, identifiedBy, identificationRemarks, identificationVerificationStatus, taxonConceptID, previousIdentification, dataGeneralizations, issue_gbifP, idigbio_flags_idbP*.
16. For each row, use the synthesis of Steps #15 and #16 to evaluate the value for *scientificName_rapid* (representing a taxonomic name inclusive of author, at any rank, hereafter referred to as “name”) against the project’s preferred taxonomic sources (listed below). In circumstances in which the project’s preferred sources disagree, conduct a review of more recent literature to determine the current consensus. Preferred sources include:
    a. Wilson DE, Reeder DM, eds. 2005. *Mammal Species of the World: A Taxonomic and Geographic Reference.* Johns Hopkins University Press.
    b. Wilson DE, Mittermeier RA, eds. 2009. *Handbook of the Mammals of the World. Vol. 9. Bats.* Lynx Edicions.
    c. Simmons NB, Cirranello AL. 2020. *Bat Species of the World: A Taxonomic and Geographic Database.* Available at https://batnames.org.
17. Record evaluation of the name from *scientificName_rapid* using the guidelines below. Note that multiple options may need to be recorded to capture the evaluation. For example, if the original value of *scientificName_rapid* was “Hipposideros galeritus celebensis,” after evaluation, the value of *scientificName_rapid* would be “Hipposideros cervinus Gould, 1854” and the *flagTaxonomy_rapid* field would be “synonym (from Hipposideros galeritus celebensis), elevated to species (from Hipposideros galeritus cervinus).” Guidelines for recording evaluation follow:
    a. Name matches preferred taxonomic sources exactly, including authority → Leave as is.
    b. Name matches preferred taxonomic sources exactly, not including authority → Leave name as is and update authority.
    c. Generic, specific epithet, or infraspecific epithet part of name is misspelled → Provide correct spelling in *scientificName_rapid* and note “misspelled (from [name])” in *flagTaxonomy_rapid*.

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.

   d. Name is currently unaccepted because subspecies has been elevated to species → Provide the current, elevated name in *scientificName_rapid* and note "elevated subspecies (from [name])" in *flagTaxonomy_rapid*.
   e. Name is currently unaccepted because genus has been reassigned → Provide the current name in *scientificName_rapid* and note "reassigned genus (from [name])" in *flagTaxonomy_rapid*.
   f. Name is currently unaccepted because species has been reassigned → Provide the current name in *scientificName_rapid* and note "reassigned species (from [name])" in *flagTaxonomy_rapid*.
   g. Name is currently unaccepted for a reason other than above but recognized as formerly accepted → Provide currently accepted name in *scientificName_rapid*, and note "synonym (from [name])" in *flagTaxonomy_rapid*.
   h. Name is doubtful based on specific, non-taxonomic details of the specimen record, e.g., the collecting locality, and this doubt is grounded in published literature → If possible, provide correctly identified name in *scientificName_rapid*, and note "misidentified (from [name])" in *flagTaxonomy_rapid*. Otherwise, use misidentified name in *scientificName_rapid*, note "misidentified ([name])" in *flagTaxonomy_rapid*, and record more detailed comments in *identificationRemarks_rapid*.
   i. Name is incorrect because of data transcription or translation error → Provide currently accepted name in *scientificName_rapid*, and note "nonexistent (from [name])" in *flagTaxonomy_rapid*.
   j. Only species in genera/genus?

18. Confirm that the value in *family_rapid* is consistent with the family of the taxon in *scientificName_rapid*. Update if needed.
19. Record the taxonomic source (most frequently one of the preferred sources listed above) in *nameAccordingTo_rapid*, using "|" to separate multiple sources.
20. Populate *identificationRemarks_rapid* with any relevant notes about the evaluation process, including uncertainty that can only be resolved by physical review of the specimen (e.g., specimens for which the collecting locality is out of the known range of the taxon). Example descriptions are provided below:
    a. "Recommend visual assessment to verify name correction."
    b. "Taxon name currently unresolved."
    c. "Taxon range currently unresolved."
    d. "Collection locality ([location]) out of range."
21. Populate *identificationVerificationStatus_rapid* with, "Identification reviewed 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."
22. Repeat Steps #15–23 until all rows in the data have been reviewed.

The fourth stage of this process is completed by the Digitization Specialist and Data Curators in OpenRefine and Google Sheets.

23. Import Google Sheet data back into OpenRefine.
24. Review data, checking that formatting is consistent with Steps #15–23.
25. Populate the following higher taxonomy fields based on value in *scientificName_rapid*: *kingdom_rapid, phylum_rapid, class_rapid, order_rapid, family_rapid, genus_rapid, subgenus_rapid, specificEpithet_rapid, infraspecificEpithet_rapid*.
26. Populate *taxonRank_rapid* based on value in *scientificName_rapid*.
27. Export data from OpenRefine as a CSV and import data into BIOSPEX.

Communication

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

Results

The first stage of this task was completed by System Administrator Robert Bruhn on 2020-10-26 and resulted in a file named “TAXONOMIC_26_2020-10-19_biospex.csv”.

The second stage of this task was completed by Digitization Specialist Erica Krimmel on 2020-11-05 with assistance from Aja Sherman and resulted in a file that was imported into Google Sheets. The second stage required approximately 6 hours of work, including some exploration of the data in preparation for writing this protocol. During the second stage, the following fields were removed from the data due to a lack of values (Step #3): *identifiedByID_gbifR*, *identifiedByID_gbifP*. No previous values existed in the field *nameAccordingTo*. Existing values in *identificationQualifier* were standardized as appropriate (e.g., “CF” to “cf.”). Existing values in *issue_gbifP* were standardized to eliminate those unrelated to taxonomy. The remaining issue flags were “TAXON_MATCH_FUZZY” (n = 2216), “TAXON_MATCH_HIGHERRANK” (n = 2031), and “TYPE_STATUS_INVALID” (n = 25). Existing values in *idigbio_flags_idbP* were standardized to eliminate flags that were unrelated to taxonomy, leaving ~30 remaining data quality flags.

The third stage of this task was accomplished in three rounds, with the first completed by Data Curator Aja Sherman on 2020-12-18 after 140 hours of work, the second completed by Aja Sherman on 2021-01-15 after 15 hours of work, and the third completed by Aja Sherman on 2021-01-20 after 43 hours of work. During the first round, names were assessed following the alphabetic (by genus, species) order presented in Wilson & Reeder (2005). This allowed for an efficient workflow as Sherman could holistically understand the nomenclatural history of a currently accepted name and then work backward to update historic synonyms of that name present in our data. If evaluating a name took or was expected to take more than one hour, this name was skipped and evaluated during Sherman’s third round. The second round consisted of reviewing work completed by the first round to ensure internal consistency in how names were evaluated. The third round involved more extensive literature review and consultation with taxonomic experts for names that were complex or ambiguous to evaluate. The third round resulted in a separate document listing all unique names found in the data and the result of their evaluation.

The fourth stage of this task was completed by Digitization Specialist Erica Krimmel on 2021-02-14.

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.