Protocol for disambiguating people name data

Authors

Deborah Paul, David Shorthouse, Katelin Pearson, Erica Krimmel, Austin Mast

Date last edited

2021-04-30

Goal

Add globally unique identifiers (i.e., ORCID IDs, Wikidata QIDs) for the people who collected (or identified) the bat specimens referenced in the dataset we are enhancing. Create Wikidata pages for deceased collectors when necessary. Link specimen records of Hipposideridae, Rhinolophidae, and Rhinonycteridae aggregated by GBIF with their disambiguated collectors in Bionomia which can foster future collaborations and data enhancing and refining efforts.

Relevant fields in the dataset

Data evaluated from

- BIOSPEXid
- gbifID_gbifR
- datasetKey_gbifP
- occurrenceID_gbifR
- institutionCode_gbifR
- collectionCode_gbifR
- catalogNumber_gbifR
- recordedBy_gbifR
- eventDate_gbifP
- country_rapid, (if null then) country_gbifR
- countryCode_rapid, (if null then) countryCode_gbifP
- dateIdentified_gbifP
- identifiedBy_gbifR
- scientificName_gbifR
- family_gbifP

Enhanced data recorded in

- recordedBy_rapid
- recordedByID_rapid
- identifiedBy_rapid
- identifiedByID_rapid

Definitions

The following definitions will aid in interpretation of this workflow documentation:

Agent namestring (or agent): a string of characters purported to represent a person but not yet linked to an externally referenced, shared identifier.

Person: an externally referenced entity meant to represent an individual human that has a resolvable, unique identifier and other defining properties (metadata) such as demographics, affiliations, lists of publications, among other items.

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).

The **second stage** of the process is completed by a key collaborator and Data Curator in Google Sheets.

2. Generate a working list of unique agent namestrings by parsing *recordedBy_gbifR* and *identifiedBy_gbifR*. Import this list into Google Sheets. Hereafter, this sheet will be referred to as the "Agent Namestrings Sheets Document".
3. Using the Bionomia Google Sheets add-on, create a new column in the "Working List" Sheet of the Agent Namestrings Sheets Document that contains, for each agent namestring, a link to the Bionomia person profile that potentially matches the provided agent namestring (e.g., https://bionomia.net/0000-0003-4118-8575 for a person with an ORCID ID or https://bionomia.net/Q335022 for a person with a Wikidata QID).
4. In the "Working List" Sheet of the Agent Namestrings Sheets Document, classify each entry to indicate the certainty of the match between the unique agent namestring from the project data and the person profile result from Bionomia, using the following controlled vocabulary:
    a. "Yes" indicates that the agent namestring is confirmed to correspond to the correct person in Bionomia as represented by data in this project.
    b. "No" indicates that the agent namestring does not appear to correspond to a person in Bionomia as represented by data in this project.
    c. "Maybe" indicates that it is unclear whether the agent namestring represents the same person as the project data represents.
5. Confirm a definitive identifier (ORCID ID or Wikidata QID) for each entry where the certainty of the results from Bionomia (see Step #4, above) is recorded as "yes." Gather the "yes" rows to populate a "Known People" sheet within a new Google Sheets document, hereafter referred to as the Strings and People Sheets Document. This sheet

should contain, at minimum, the collector/determiner name (as listed in ORCID or Wikidata), the person's identifier (ORCID ID or Wikidata QID), and a column used to track whether that person's specimens have been reviewed in Bionomia. This sheet will be used in the sixth stage. The results flagged as "maybe" or "no" will be evaluated in the third and fourth stages and will be found in the GBIF Agent Strings sheet of the Strings and People Sheets Document Document.

The **third stage** of the process is completed by a key collaborator and Data Curators in Google Sheets.

6. Begin to review the Agent Namestrings Sheets Document, starting with the most unambiguous (i.e., unique) and complete (e.g., full first, middle, and last) names.
7. Determine whether each entry refers to a person who already exists in either the ORCID or Wikidata system. Use sources such as Google and ResearchGate (and other references listed below). Evidence of a person's identity as the collector of bat specimen(s) should be inferred from a combination of evidence such as:
    a. The person is described in one or more obituaries, biographies, appropriately supported Wikipedia articles, natural history archives, or other reputable sources as a collector of bats, mammals, or associated taxa; zoologist; mammalogist; speleologist; naturalist; zooparasitologist; ornithologist; or other professional who might reasonably be expected to have traveled to or lived in the collecting locations. Note that these people can be varied and include, for example, entomologists, scientific illustrators, artists, spies, government officials, clergy, etc.
    b. The person is described in one or more reputable sources (see 7a above) to have collected, studied, or otherwise been active in the country/countries in which the specimens associated with the agent string were collected.
        i. To view the specimens associated with an agent string, enter the agent string into the following URL formula (for the name F. M. Last): https://bionomia.net/agents?q=F.%20M.%20Last. Each initial or part of the name that is separated by a space is represented by %20 in the URL.
        ii. The collection location and other information about the specimen can be accessed by clicking on the scientific name of the entry of interest.
    c. The person is an author of one or more scientific articles about chiropterology, mammalogy, etc., particularly if the specimens associated with the agent string were collected in the same region(s) described by the scientific article(s) that may be found by searching resources such as Google or ResearchGate.
8. If research in step 7 confirms that an entry refers to a person who already exists in either the ORCID or Wikidata system, add the appropriate definitive identifier (ORCID ID or Wikidata QID) to the Agent Namestrings Sheets Document in the column *Definitive ORCID or Wikidata*. Note that if the existing Wikidata entry does not include all of the

information listed in Step 9, below, it must be added according to the same procedure described in Step 9.

9. If research confirms that no entry for this person exists in either the ORCID or Wikidata system, and the person is deceased, create a Wikidata page for the person. Record the new Wikidata QID in the *Definitive ORCID or Wikidata* column of the Agent Namestrings Sheets Document. Use and refer to the video tutorial and document on making a Wikidata page by Siobhan Leachman. When creating a Wikidata page, the information described below is required to subsequently create a profile in Bionomia, and each piece of information must be supported by, at minimum, one citation from an obituary, published biography, genealogical record, museum archive, or other reputable source. Conduct a thorough search such that each can be supported by two citations whenever possible. The information required is:
    i. Person name, including aliases if known.
    ii. Birth date, at least as specific as a year.
    iii. Death date, at least as specific as a year.
    iv. At least one external identifier. This identifier can be the Bionomia ID (which corresponds to the Wikidata QID) or ORCID ID.
10. **Optional, and not performed in our RAPID project:** When research confirms that no unambiguous entry for this person exists in the ORCID system, and the person is alive, contact the person and request that they either confirm which ORCID ID belongs to them or create an ORCID ID if they do not already have one. Once they have done so, record their ORCID ID in the Known People sheet of the Strings and People Sheets Document and request that they make their profile on Bionomia public.

The **fourth stage** of the process is completed in the Strings and People Sheets Document by Data Curators, as well as external community volunteers, in a workshop hosted by the Co-Principal Investigator and a key collaborator.

11. Workshop Process
    a. Develop an invitation list with email addresses for bat collectors and collection managers familiar with those working with or curating Rhinolophidae, Hipposideridae, Rhinonycteridae, and related bat groups. Track RVSPs via a Google Form.
    b. Develop the one-day workshop agenda.
    c. Develop workshop materials (see Bionomia Link to Workshop materials).
    d. In the one day workshop, train attendees (see video) and work together both online and offline to:
        i. attribute specimens in Bionomia, and
        ii. add extra information about the collectors where we have little to no information in the GBIF Agent Strings sheet in Strings and People Sheets Document. Contributors that add details are encouraged to cite sources

        and provide their initials for proper attribution. This information will be used by Data Curators to disambiguate collectors in the fifth stage.
        iii. (Optional) conduct steps from the sixth stage (attribute specimens to collectors in Bionomia) using the ORCID IDs and Wikidata QIDs produced during the third stage.
    e. Share workshop materials via Zenodo and Bionomia.
    f. Maintain communication with participants via email during the course of the workshop.
    g. Archive a copy of the Strings and People Sheets Document on Zenodo (https://doi.org/10.5281/zenodo.3974999).
    h. Leave open the original Google Sheets Document for further community work.

12. Encourage community members to continue to contribute to the Strings and People Sheets Document for two weeks following the workshop, at which time, the document is closed to editing.

The **fifth stage** of the process is completed by a key collaborator and the Data Curators in Google Sheets.

13. Sort the Strings and People Sheets Document according to the number of specimens associated with each agent string. (Note this specimen count column was generated via a one-off, command-line ruby query in interactive ruby (irb) to produce a csv file against a local MySQL store of pre-processed data created through bi-monthly data refreshes executed in support of Bionomia). For the first 100 entries collecting or identifying the most specimens, conduct steps 7–9 unless a Wikidata QID or ORCID ID has already been assigned to them.
14. Conduct Steps 7–9 for the remaining agent strings in the Strings and People Sheets Document for which a community member added a note during the fourth stage.

The **sixth stage** of the process is completed by a key collaborator, workshop participants, and the Data Curators in Google Sheets.

15. Log in to Bionomia at https://bionomia.net.
16. Select a person from the Known People sheet of the Strings and People Sheets Document. Navigate to the “Help Attribute” page in Bionomia for this agent (see Bionomia Helping URL column) and filter the discovered specimens by those in the family Hipposideridae, Rhinolophidae, or Rhinonycteridae. See the following resources for instructions on using Bionomia to attribute specimens to collectors/determiners:
    a. Instructions for Bionomia from the Workshop recording (see fourth stage).
    b. “Getting Started with ORCID” available at https://bionomia.net/help.
    c. “Claiming Specimen Records” available at https://bionomia.net/help.
    d. “Helping Others” available at https://bionomia.net/help.

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.

17. Evaluate each specimen record and assign it as being collected by the person if either of the following scenarios are true:
    a. Value in the Bionomia field *Collected By* matches the preferred name or an unambiguous (i.e., highly unique and/or containing first, middle and last name of collector) alias of the person (e.g., “Harold Elmer Anthony” vs. “Harold Smith”). Value in the Bionomia field *Date Collected* is within the lifespan of the person.
    b. Value in the Bionomia field *Collected By* appears to be a minor misspelling of the person’s name or is an otherwise ambiguous alias of the person (e.g., consists of only initials and a last name, or the name is common, such as “John Thompson” or “Jose Garcia”). Value in the Bionomia field *Date Collected* is within the lifespan of the person, and at least one other item of evidence (institution, collecting locality, collector number, known co-collector(s), etc.) supports the assumption that this specimen was collected by this person. If necessary, click the link under *Scientific Name* to review the full specimen record in GBIF for additional evidence.
18. Evaluate each specimen record and assign it as being identified by the person if either of the scenarios described in Step 17 (above) are true.
19. Where a specimen record cannot be assigned with confidence, leave it alone.
20. Where a specimen record definitely does not refer to the person at hand, flag it using Bionomia’s “Not them” button.
21. Once all specimens are assessed and appropriately claimed or ignored, for those people who have Wikidata QID as their identifier, click “make public” to generate a public profile page on Bionomia.

The **seventh stage** of the process is completed by a Data Curator in BIOSPEX and R.

22. Export a dataset from BIOSPEX with the following fields: *BIOSPEX_id*, *gbifID, recordedBy, recordedByID, identifiedBy,* and *identifiedByID*.
23. Get relevant data from Bionomia:
    a. For each value of *gbifID*, fetch JSON-LD from Bionomia using the URL format “https://bionomia.net/occurrence/” + [*gbifID* value] + “.json,” (e.g., “https://bionomia.net/occurrence/477976412.json”).
    b. Extract data from the *Recorded* and *Identified* JSON arrays and merge into the dataset from BIOSPEX.
24. Reformat the data for ingestion back into BIOSPEX by completing the following:
    a. Concatenate the values for *givenName* and *familyName* into *recordedBy_rapid* for each item in the *Recorded* array from Bionomia, using a vertical bar and spaces (|) to separate items.
    b. Transfer the values for *sameAs* into *recordedByID_rapid* for each item in the *Recorded* array from Bionomia, using a vertical bar and spaces (|) to separate items.

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.

    c. Concatenate the values for *givenName* and *familyName* into *identifiedBy_rapid* for each item in the *Identified* array from Bionomia, using a vertical bar and spaces (|) to separate items.
    d. Transfer the values for *sameAs* into *identifiedByID_rapid* for each item in the *Identified* array, using a vertical bar and spaces (|) to separate items.

25. Import data back into BIOSPEX.
    a. Note: names provided in the *recordedBy* and *identifiedBy* are not recommended to replace what collections already have. Rather, they are reference values for the collections to refine what they know about these individuals including the IDs, where we managed to find or assign one.

The **eighth stage** of the process is completed by the informatics team.

26. Using the data in Bionomia, query the *eventDate* and *dateIdentified fields* against birth and death dates of the person indicated in *recordedByID* or *identifiedByID* by making use of Wikidata SPARQL queries, flagging suspect records (i.e., those collected or identified before birth or after death).

The **ninth stage** of the process is completed by Data Curators and a key collaborator.

27. Contact collection managers and direct them to the Frictionless Data downloads from Bionomia for their dataset (e.g., https://bionomia.net/dataset/b6015b60-6f96-43a9-88e5-2f41854e8f07). The UUID here is the same as that used by GBIF in the field *datasetKey*.

References used by the Data Curators when finding sources to make Wikidata entries:

- Family Search, https://www.familysearch.org (e.g., a search here for the England and Wales Death Registration Index 1837-2007)
- Find A Grave, https://www.findagrave.com
- Smithsonian Online Virtual Archives, https://sova.si.edu
- Posted obituaries (e.g., https://www.lensingfuneral.com/obituaries/obituary-listings?obId=1046659)
- Digital biographies of the collector, e.g., https://viaf.org/viaf/459154441719535460005/
- Hardcopy biographical literature (e.g., The Eponym Dictionary of Mammals, The Eponym Dictionary of Birds, and The Eponym Dictionary of Reptiles)
- Online regional encyclopedias (e.g., https://www.enciklopedija.hr/natuknica.aspx?id=16917)

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.

Communication

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

Results

The first stage was completed by Robert Bruhn and Deborah Paul on 2020-10-01 and resulted in the file *PEOPLE_13_2020-10-01_dpaul.csv*.

The second stage was completed by David Shorthouse and Deborah Paul on 2020-11-15 and resulted in two Google Sheet workbooks, 1) the Agent Strings Sheets Document which was then used to support the work to generate 2) the Strings and People Sheets Document with two sheets: GBIF Agent Strings and Known People.

The third stage was completed by David Shorthouse, Deborah Paul, Katelin Pearson, and Trevor Dalton on 2020-11-30.

The fourth stage was completed by Workshop Participants on 2020-12-01 and resulted in new entries in the Strings and People Sheets Document and new profiles in Bionomia.

The fifth stage was completed by David Shorthouse, Deborah Paul, and the Data Curators on 2021-01-05.

The sixth stage was completed by David Shorthouse, Deborah Paul, workshop participants, and the Data Curators over a period of time spanning from the workshop (2020-12-01) until 2021-01-05.

Steps #22–24 of the seventh stage were completed by Katelin Pearson on 2021-01-06 and took a total of 7 hours, including exploratory data analysis and writing the code to accomplish this stage. A total of 648 people records for specimen collectors and 69 people records for specimen identifiers were returned from Bionomia in Step #23. When reintegrated with the data from BIOSPEX in Step #24, the results from Bionomia applied to 27,075 rows in the *recordedBy_rapid* and *recordedByID_rapid* fields and 7,833 rows in the *identifiedBy_rapid* and *identifiedByID_rapid* fields. In the original data, 54,631 rows had 3,335 distinct values in the field *recordedBy* and 12,025 rows had 240 distinct values in the field *identifiedBy*. Only 2 rows in the original data had a value for *recordedByID* and no rows had a value for *identifiedByID*. It should be emphasized that the values in *recordedBy_rapid, recordedByID_rapid, identifiedBy_rapid,* and *identifiedByID_rapid* only contain enhancements to a subset of the original data because not all agent namestrings could be disambiguated. In some cases, this means that where the original data had agent namestrings representing more than one person, the RAPID enhanced data may only have been able to disambiguate one of those agent namestrings. Step #25 was completed by Robert Bruhn on 2021-02-02. Code associated with the seventh stage of this protocol can be found in ‘RAPID-code_people.R’ (archived at https://doi.org/10.5281/zenodo.3974999).

The eighth stage was completed by key collaborator David Shorthouse on 2021-02-25 using the records now in Bionomia. This stage resulted in 21 specimen records flagged as having a date collected or date identified that does not fit into the birth-death range for the person collecting or identifying the specimen referenced. The results can be found in the file ‘bionomia-problem-dates-all-datasets_2021-02-25.csv’ (archived at https://doi.org/10.5281/zenodo.3974999).

The ninth stage will be completed by key collaborator David Shorthouse and Deborah Paul at the end of this work, by contacting those collections (top 15) contributing the most records to the dataset, to share the resulting records now in Bionomia.

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.