Protocol for resolving collection date information

Authors

Katelin Pearson, Erica Krimmel, Austin Mast

Date last edited

2021-04-30

Goal

Standardize format for dates of collection. Assign and standardize previously unstandardized dates of collection for specimens for which additional evidence can be leveraged to resolve missing data (e.g., cases in which a duplicate specimen was collected and the collection date was included on that label) and ambiguities (e.g., cases in which the century of collection can be inferred using the known lifespan of the collector, but is otherwise unclear, such as “12 Dec 13”). Assign and standardize date ranges for specimens for which evidence can be leveraged from the lifetime or floruit of the specimen collector.

Relevant fields in the dataset

Data evaluated from

- BIOSPEXid
- eventDate_gbifP / eventDate_gbifR / eventDate_idbR / eventDate_idbP
- verbatimEventDate_gbifR / verbatimEventDate_gbifP / verbatimEventDate_idbR / verbatimEventDate_idbP
- year_gbifR / year_gbifP / year_idbR
- month_gbifR / month_gbifP / month_idbR
- day_gbifR / day_gbifP / day_idbR
- recordedBy_rapid
- recordedByID_rapid

Enhanced data recorded in

- eventDate_rapid
- flagEventDate_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 is completed by Data Curators in R and Excel. Each of the following steps is performed on the reduced number of records that remain without a value in *eventDate_rapid* after completing the previous step.

2. Attempt to parse dates from *eventDate_gbifP, eventDate_gbifR, eventDate_idbR, eventDate_idbP, year_gbifR, year_gbifP, year_idbR, month_gbifR, month_gbifP, month_idbR, day_gbifR, day_gbifP* and *day_idbR* fields into *eventDate_rapid*. When a standardized date can be retrieved from one of the fields, note this source in the *flagEventDate_rapid* field (e.g., "Date sourced from eventDate_gbifR"). Review for non-standardized or impossible values, including "9999," question marks, dashes, "NA," "/n," textual date descriptions, and years greater than 2020. Correct such values when enough information is present to do so. Use the standard phrase, "Reinterpreted original data" to note where this is done. Use the following formatting guidelines:
    a. Enforce the date format YYYY-MM-DD. For date ranges, separate the start and end dates with a forward slash (e.g., YYYY-MM-DD/YYY2-M2-D2 or YYYY/YYY2 or YYYY-MM-DD/D2) as recommended by the Darwin Core standard for this term.
    b. For partial dates, enter only what is present, e.g., "1987-02."
    c. Use the "1800" as a standard placeholder for start date in a range where this information is otherwise lacking (e.g., "1800-02-16/2020-02-16").
3. Attempt to parse dates from *verbatimEventDate* fields into *eventDate_rapid* following the formatting guidelines in Step #2a-c (above). This may involve converting Roman numerals, words, or question marks into digits. When Roman numerals are present, interpret them as the collection month unless evidence suggests otherwise. When a standardized date can be retrieved, note this in *flagEventDate_rapid* as, e.g., "Date sourced from verbatimEventDate_gbifR." Use the phrase "Reinterpreted original data" when a non-standardized date is re-interpreted into a standardized date.
4. Attempt to use collector metadata to assign collection dates to specimen records. Many agent namestrings in the *recordedBy_rapid* (collector) field are disambiguated and referenced to person records in Bionomia as part of another protocol in this project (see *RAPID-protocol_people.pdf*). Use information from the collector's unique ID (found in the *recordedByID_rapid* field) to assign a value to *eventDate_rapid* based on the following guidelines:
    a. For specimens for which a *verbatimEventDate* field indicates a specific date except for the century (e.g., "29-Oct-62"), cross-reference the date with the birth year(s) of the collector(s) as indicated in Bionomia (see Step #6 for how to do this). For specimens with the collector(s) born after 1880, assume the 20th century for *eventDate_rapid*. Record how this determination was made in *flagEventDate_rapid*, using the phrase, "Century confirmed from collector lifetime."
    b. Assign a value to *eventDate_rapid* based on the year range of the collector's lifetime (e.g., "1857/1941"). Record how this determination was made in

flagEventDate_rapid, using the standard phrases, “Date assigned based on collector birth and death years” for a single collector and, “Date assigned based on collector birth and death years (intersection between multiple collectors)” for specimens for which multiple collectors were referenced.

c. For collectors with ORCID IDs instead of Wikimedia QIDs in the *recordedByID_rapid* field, assign a value of “1920/2020” to *eventDate_rapid* and record “Date assigned based on living collector as of 2020” in *flagEventDate_rapid*.

5. Attempt to use associated images to assign collection dates to specimen records. For each specimen, open the corresponding record online in GBIF or iDigBio and determine whether a date is displayed on associated images.
    a. If you transcribe the collection date from a specimen label, note this in *flagEventDate_rapid* as “Transcribed from image of original specimen label.”
    b. If an image of a ledger is present and the ledger entries before and after the focal specimen indicate a certain collection year, assign the year of collection of ledger entries before and after to the specimen (e.g., enter only “1926” in *eventDate_rapid*). Note this in *flagEventDate_rapid* as “Transcribed from image of original ledger entry.”
6. Attempt to use collector activity to assign collection dates to specimen records. For each value of *recordedBy_rapid* for specimens without collection dates, look at the person profile (or agent string) in Bionomia by opening the URL produced by concatenating the value in the *recordedByID_rapid* field with “https://bionomia.net/”, or by searching for the person in Bionomia by entering the collector name string. Download the Bionomia specimen records attributed to this person/name string and sort the specimen records according to collection country and collection date. Use this information to assign a value to *eventDate_rapid* based on the following guidelines:
    a. Exclude obvious outliers from the downloaded Bionomia specimen records (e.g., one or two specimens in an entirely different century) only if there are no more than three. If there are more than three, skip to Step #7 (below).
    b. If fewer than 20 total Bionomia specimen records exist in the download for this person, skip to Step #7 (below).
    c. Determine the person’s range of collecting years in the country where a specimen from this project was collected. If this range is larger than 100 years, skip to Step d (below). If there are at least 20 Bionomia specimen records for a given country, use this range of collecting years to assign a value to *eventDate_rapid* (e.g., “1962/2017”). Record how this determination was made in *flagEventDate_rapid*, using the standard phrasing, “Date assigned based on minimum and maximum of collecting years in collecting location ([insert country]), n = [insert number of specimens upon which this assumption was based]”. For example, “Date assigned based on minimum and maximum of collecting years in collecting location (Indonesia), n = 575”.

d. If fewer than 20 specimens exist for this person in a given country, evaluate the minimum and maximum collecting years of the person in all countries. If this range is larger than 100 years, skip to Step #7 (below). Otherwise, use this range to assign a value to *eventDate_rapid*, e.g. “1962/2017.” Record how this determination was made in *flagEventDate_rapid*, using the standard phrasing, “Date assigned based on minimum and maximum of collecting years, n = [insert number of specimens upon which this assumption was based].”
e. When the museum catalog number can be used to set a maximum collecting date (e.g., the catalog number includes a year), use this to set the maximum collecting date in the *eventDate_rapid* range. Record how this determination was made in *flagEventDate_rapid*, using the standard phrasing, “Date assigned based on minimum of collecting years in collecting location ([insert country) and maximum year from museum catalog number.”
f. When the museum catalog number can be used to establish both a minimum and maximum collection year (i.e., the catalog numbers before and after the specimen have definitive dates of collection), assign the range of possible collection dates accordingly. For example, specimens with museum catalog numbers 1972.4161 and 1972.4236 by the same collector were both collected in the same country in 1960. A specimen in this project’s dataset from the same institution has catalog number 1972.4179 and was collected in the same country. The year of collection can be assumed to be 1960. Record how this determination was made in *flagEventDate_rapid*, using the standard phrasing, “Date assigned based on museum catalog number.”

7. Look at other specimen records with similar locality or collector information to see if patterns can be identified and a date from another record applied. Assign a value to *eventDate_rapid* based on the following guidelines:
    a. Search in GBIF and iDigBio for duplicate specimens using the collector name, collector number (where present), country, and/or other pertinent fields. Where found, assign the duplicate collecting date value to *eventDate_rapid* and record “Date assigned based on duplicate specimen” in *flagEventDate_rapid*.
    b. If a collector number exists, search in GBIF and iDigBio for slightly higher and slightly lower collector numbers used by the same collector. Determine whether a date can be confidently applied based upon these specimens’ collecting dates. If so, assign a value to *eventDate_rapid* and record how this determination was made in *flagEventDate_rapid* using the standard phrase “Date assigned based on collector numbers” along with additional details.
    c. For specimens where the collection date is present except for the century (e.g., “29-Oct-62”), yet the collector name was not disambiguated, search in GBIF and iDigBio for additional specimens collected by the same person. Determine whether a century can be confidently applied based upon these specimens’ collecting dates. Record how this determination was made in

flagEventDate_rapid, using the standard phrase, “Century confirmed from other collections.”

d. For specimens where the collection date is present except for the century (e.g., “29-Oct-62”), search in GBIF and iDigBio for additional specimens identified by the same person. Determine whether a century can be confidently applied based upon these specimens’ identification dates. Record how this determination was made in *flagEventDate_rapid*, using the standard phrase, “Century confirmed from other identifications.”

8. Flag remaining specimens as uninterpretable by recording “Unable to disambiguate date” in *flagEventDate_rapid*. Reasons for being uninterpretable included the following:
    a. Field *recordedBy_rapid* contains only a last name.
    b. Specimen record is completely lacking in date and collector data.

Communication

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

Results

This task was completed by Data Curator Katelin Pearson on 2021-01-24 and took approximately 40 hours, including exploratory data analysis and writing the code to accomplish this stage. In Step #2 (above), values for *eventDate_rapid* were sourced from the following fields in the original data: *eventDate_gbifP* (n = 65,458), *year_idbR* + *month_idbR* + *day_idbR* (n = 3,742), *eventDate_gbifR* (n = 177), and *verbatimEventDate_gbifP* (n = 15). Per Step #2c, the placeholder year 1800 was used in 14 records. Original date data was reinterpreted (Steps #2 and #3) in 2,033 records. Using disambiguated collector birth and death dates in Bionomia, 32 rows in this project’s data had the value of their century in *eventDate_rapid* confirmed by the disambiguated collector’s lifespan (Step #4a), 2,650 rows had date ranges assigned based on the collector(s)’ lifespan (Step #4b), and 61 rows had date ranges assigned based on the birth year of a currently living collector (Step #4c). Per Step #5, values for *eventDate_rapid* were transcribed from images related to the specimen in 103 rows, and from other fields in the specimen record in 29 rows. Using specimen data downloaded from Bionomia, 645 rows had date ranges assigned as values in *eventDate_rapid* based on the collector’s activity in a given country (Step #6c) and an additional 201 rows had date ranges assigned based on the collector’s total activity (Step #6d). A combination of the collector’s activity and museum catalog number was used to assign a date range to an additional 86 rows (Step #6e), and museum catalog number alone was used to assign a date range to an additional 3 rows (Step #6f). Using collector numbers, 11 rows in this project’s data had a value assigned to *eventDate_rapid* based on the collecting date of a duplicate specimen or other related specimen (Step #7a), and 1 row had a value assigned based on collector numbers (Step #7b). The century of an existing value in *eventDate_rapid* was confirmed in 214 rows by other specimens collected by the same person (Step #7c), and in 2 rows by other specimens identified by the same person (Step #7d). Date ranges were assigned based on the date of a collecting expedition in 2 rows.

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.

Prior to this protocol, 69,392 (77.2%) rows in the dataset contained a value in one of the *eventDate* fields.

After completion of this protocol, a value for *eventDate_rapid* was confirmed or assigned to 75,369 (83.9%) rows. The resolution and format of these dates is:

Date, YYYY-MM-DD = 69,135 rows
Date, YYYY-MM = 1,586 rows
Date, YYYY = 248 rows
Date range, YYYY-MM-DD/YYY2-M2-D2 = 439 rows
Date range, YYYY-MM/YYY2-M2 = 335 rows
Date range, YYYY/YYY2 = 3,576 rows
Date range, YYYY-MM-DD/D2 = 50 rows

This project was unable to disambiguate a collecting date for 14,468 (16.1%) rows.

Code associated with this protocol can be found in 'RAPID-code_collection-date.R' (archived at https://doi.org/10.5281/zenodo.3974999).