Protocol for identifying and recording references to external linkages in specimen records

Authors

Katelin Pearson, Erica Krimmel, Deborah Paul, Austin Mast

Date last edited

2021-04-15

Goal

Discover and standardize linkages between specimen records in biodiversity aggregators (this project's dataset) and genetic material records aggregated by the United States National Center for Biotechnology Information's GenBank® and the Barcode of Life project. Examine and summarize data from other relevant fields that may contribute to specimens' extended data network, such as gut, tissue, or DNA data.

Relevant fields in the dataset

Data evaluated from

- BIOSPEXid
- associatedSequences
- catalogNumber
- associatedReferences
- materialSampleID
- samplingProtocol
- preparations
- otherCatalogNumbers
- dynamicProperties

Enhanced data recorded in

- associatedSequences_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).

The second stage of this process is completed by one Data Curator in R.

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. Populate the field *associatedSequences_rapid* by standardizing existing values in the other *associatedSequences* fields. Convert NCBI sequence IDs and incorrect NCBI URLs to resolvable URLs (e.g. “https://www.ncbi.nlm.nih.gov/nuccore/GU328060”) as recommended by Darwin Core.
3. Download records from GenBank® (https://www.ncbi.nlm.nih.gov/genbank) associated with bats in the families Hipposideridae or Rhinolophidae.
4. Use the data acquired in Step #3 to compare catalog numbers (“specimen voucher” numbers) cited by GenBank® to catalog numbers present in this project’s dataset (in the field *catalogNumber*). Align formats of catalog numbers in this project’s dataset with the formats of catalog numbers cited by GenBank®, when necessary (e.g., append the institution code to the front of the raw catalog number). Add values to the field *associatedSequences_rapid* for any record for which this step identifies a new associated sequence based on matching catalog numbers.
5. Download records from Barcode of Life (http://www.barcodinglife.org) associated with bats in the families Hipposideridae or Rhinolophidae.
6. Use the data acquired in Step #5 to compare catalog numbers cited by Barcode of Life to catalog numbers present in this project’s dataset (in the field *catalogNumber*). Add values to the field *associatedSequences_rapid* for any record where this step identifies a new associated sequence based on matching catalog numbers.
7. Explore data contained in the other fields listed in “Data evaluated from” (above) and summarize this as a starting point for future work to standardize linkages to other data types.

Communication

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

Results

This task was completed by Data Curator Katelin Pearson on 2021-02-12 and took approximately 16 hours, including exploratory data analysis. Step #2 (above) resulted in standardized values for 4,957 records; four of these records had existing, properly formatted, resolvable URLs, two had properly formatted but non-resolving URLs, and the remaining 4,951 had GenBank® accession numbers, which were converted into properly formatted, resolvable URLs.

In Step #3, the downloaded GFF3 file was converted into a CSV file such that each line of the GFF3 file corresponded to a single row in a single column in the CSV. The resulting CSV files contained over 200,000 records from GenBank® associated with bats in the families Hipposideridae or Rhinolophidae. Step #4 resulted in adding standardized values to 1,109 records. In Step #5, the downloaded Darwin Core archive file contained 2,002 records from Barcode of Life (BOLD) associated with bats in the families Hipposideridae or Rhinolophidae. Step #6 identified 766 BOLD records associated with specimens in our dataset, and nine of

these had not been previously identified by Step #4. A total of 6,075 records (6.8% of the records in this project's dataset) are associated with sequence data. Of these, 227 records are associated with at least two sequences. The code used to accomplish Step #1 can be found in 'RAPID-code_external-linkages-standardize.R.' The code used to accomplish Steps #4 and #6 can be found in 'RAPID-code_external-linkages-genbank.R' and 'RAPID-code_external-linkages-bold.R,' respectively. These code files are archived in the project repository (https://doi.org/10.5281/zenodo.3974999). The results of Step #7 are summarized below:

Summary of data related to linked resources

We explored data contained in the several additional fields that may relate to external data, but which have less well-defined recommendations for standardization. The types and formats of data in these text fields are heterogeneous, and standardization of the values present was beyond the scope of this project. However, we provide a summary of the contents of these fields for potential future work.

The *associatedReferences* field was sparsely populated in our dataset. Only nine specimens cited a single published article, 44 cited the 2017 African Chiroptera Report, and 26 records listed the specimen's determiner in this field.

Similarly, the *materialSampleID* field contained data for only 29 records. For these records, the field contained UUIDs that corresponded with the specimens' occurrenceIDs.

The *samplingProtocol* field was populated for 298 records and did not contain data regarding genetic, tissue, gut, or other sampling methods. Rather, the values described how the specimen was collected (e.g., "hand net", "mist net", "killed by cat").

In contrast, the *preparations* field was more frequently populated in our dataset—35,301 records, or 39.3% of the records in this project's dataset, had values present in this field—though few records contained evidence of linked resources. Although not strictly standardized, most data in this field adhered to the Darwin Core recommendation of listing which method or methods were used to preserve the specimen. For example, 12,839 records included "skin", 11,792 included "skull" or "cranium", 3,988 included "whole animal" or "whole body," and 2,795 included "skeleton" or "bones." A large portion (~43%) of records with information in the *preparations* field included more than one preparation method, such as "skeleton, partial; skin, study." The type of preparation was also frequently mentioned in this field, e.g., "wet," "alcohol," "EtOH," or "fluid" was mentioned in 18,199 records, and "dry" or "mummified" in 1,050 records. There were no instances of "DNA," "gut," or "sequence" in this field for any records in the dataset. "Tissue" was listed in this field for 1,100 specimens in the dataset, and 51 of these records (all from the same institution) also included a tissue number.

Tissue numbers were also discovered in the *otherCatalogNumber* field for 102 records in the dataset. Otherwise, this field generally contained only preparator numbers (617 records), collector numbers (387 records), and other identifiers that appeared to identify the specimen itself rather than additional tissues or resources. A total of 25,708 records in the dataset (28.6%) had some value in the *otherCatalogNumber* field.

Lastly, we identified the *dynamicProperties* field as a potential repository for external data. This field was populated for 20,230 records (22.5%) in our dataset. This field most frequently included data on the specimen's sex (2,259 records), weight or mass (1,176 records), or length (of a limb, tail, or full body length; 1,090 records). No URIs or connections to trait or anatomical ontologies were observed with these measurements. For 9,956 records, this field included a "Related Resource ID", though the nature of this related resource was not specified, nor could this be determined from the collection's home data portal. The latter records, as well as 6,822 additional records, also contained metadata about the specimen such as donor name, determination names, date created, and data quality flags issued by GBIF. However, no potential linkages to genetic, gut, or tissue data were observed.