Protocol for standardizing country data

Authors

Erica Krimmel, Austin Mast

Date last edited

2021-05-05

Goal

Assign ISO 3166 three-letter code (alpha-3) to all unambiguous values for country in the data.

Relevant fields in the dataset

Data evaluated from

- country_gbifR / country_idbP / country_idbR
- countryCode_gbifR / countryCode_gbifP / countryCode_idbR / idigbio_isoCountryCode_idbP
- decimalLatitude_rapid
- decimalLongitude_rapid
- geodeticDatum_rapid

Enhanced data recorded in

- country_rapid
- countryCode_rapid

Process & Parties Responsible

The first stage in this process is completed by the Digitization Specialist in R and OpenRefine prior to the beginning of the georeferencing protocol.

1. Read the primary records dataset into R and generate a list of distinct values for the combination of data available in *country_gbifR, country_idbP, country_idbR, countryCode_idbR, countryCode_gbif*, and *idigbio_isoCountryCode_idbP*. Export these data out of R and into OpenRefine to facilitate edits happening at a cell level.
2. In OpenRefine, use a gazetteer (e.g. Geonames, https://www.geonames.org) to verify and resolve country names and codes where possible.
    a. Where raw data automatically resolve against the gazetteer, assign gazetteer values to *country_rapid* and *countryCode_rapid*, using ISO 3166 three-letter codes (alpha-3) for the country code.

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.

   b. Where raw data do not automatically resolve, determine a value for *country_rapid* and *countryCode_rapid* manually (as illustrated in the Results section below), based on gazetteer values.
   c. Where raw data cannot be resolved automatically or manually (as illustrated in the Results section below), record "[undetermined]" in *country_rapid* and leave *countryCode_rapid* blank.
3. Export data out of OpenRefine and read back into R.
4. Reintegrate data in *country_rapid* and *countryCode_rapid* at the row level in the primary records dataset.

The second stage in this process is completed by the Digitization Specialist in QGIS after the georeferencing protocol is complete.

5. Reverse geocode records by intersecting georeferenced points (as recorded in *decimalLatitude_rapid*, *decimalLongitude_rapid*, and *geodeticDatum_rapid*) with country (i.e., Level 0) polygons provided by GADM (Database of Global Administrative Areas, https://gadm.org).
6. Update data in *country_rapid* and *countryCode_rapid* based on Step #5.
7. Export as a CSV file and import data into BIOSPEX.

Communication

Communication for this task will be via weekly team meetings.

Results

Stage one was completed by Digitization Specialist Erica Krimmel on 2020-09-23 and took a total of 2 hours. Gazetteer information came from *countryInfo.txt*, accessed from https://download.geonames.org/export/dump/ on 2020-07-30. There were 805 distinct combinations in Step #1 (above). Of these, 709 resolved automatically against the gazetteer, either based on the country or country code values (Step #2a). 61 rows were resolved by manually looking up the appropriate gazetteer value for country names that were recorded either with a misspelling, in a non-english language, or using a non-preferred format (Step #2b). 35 rows were unable to be resolved, either because the data recorded was too vague, or because it referenced a country that no longer exists and does not have a modern equivalent in the same geographic footprint, e.g. "Yugoslavia" (Step #2c). There were a total of 125 countries represented in the data.

Stage two was completed by Digitization Specialist Erica Krimmel on 2021-03-01 and took a total of 2 hours. Geospatial data from the GADM database was version 3.4, created in April 2018 (*gadm36_levels_gpkg.zip*; accessed from https://gadm.org/download_world.html on 2021-02-26). Out of the 56,203 records for which coordinates were verified or assigned as part of the georeferencing protocol, the existing value for *country_rapid* was confirmed in 52,782 records (93.9%). A value for *country_rapid* was assigned to 2,799 records (5%) with a

previously undetermined country. The existing value for *country_rapid* was updated based on new information gained from the georeferencing process for 622 records (1.1%); these updates were manually verified by looking at the *georeferencingRemarks_rapid* field, and by plotting the coordinates on a map where even more verification was necessary (n = 77). 951 records with unreviewed legacy coordinates that were retained by this project were also reverse geocoded. Based on these coordinates, the existing value for *country_rapid* was confirmed in 868 records, the existing value for *country_rapid* was updated in 2 records, and a value for *country_rapid* was assigned to 72 records with a previously undetermined country. Additionally, the legacy coordinates for 9 records were deemed implausible and discarded.

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

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.