# Protocol for standardizing country data


## Authors

Erica Krimmel, Austin Mast


## Date last edited

2020-09-23

## 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_idbR / countryCode_gbif / idigbio_isoCountryCode_idbP

### Enhanced data recorded in

- country_rapid
- countryCode_rapid

## Process & Parties Responsible

This process will be completed entirely by Erica Krimmel in R and OpenRefine, following these steps:

1. Read the primary records dataset into R and generate a list of distinct values for the combination of raw data fields in "Data evaluated from" (above). Export this data out of R and into OpenRefine to facilitate edits happening at a cell level.
2. In OpenRefine, use a gazetteer (Geonames 'countryInfo.txt' file, accessed at https://download.geonames.org/export/dump/ on 2020-07-30) 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.
    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.

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.

4. Reintegrate data in *country_rapid* and *countryCode_rapid* at the row level in the primary records dataset.

### Communication

This is a short term task with no in-progress communication necessary.

### Results

This task was completed by Erica Krimmel on 2020-09-23 and took a total of 2 hours. 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.

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