Processing Country Centroids at the Global Biodiversity Information Facility

John Thomas Waller [‡]

[‡] GBIF, Copenhagen, Denmark

Corresponding author: John Thomas Waller (jhnwllr@gmail.com)

Abstract

The Global Biodiversity Information Facility (GBIF) is an international network and data infrastructure that promotes open access to biodiversity data. GBIF serves as a global hub for aggregating and disseminating biodiversity information from diverse sources, including museums, research institutions, and citizen science projects.

I will present recent additions to GBIF data quality measures, focusing on the introduction of the country centroid filtering feature. Additionally, I explore the functionality and significance of GBIF data quality flags, which aid in assessing the reliability and usability of the aggregated data. I will also discuss recent changes to the default filtering mechanism employed by GBIF to remove suspicious points.

GBIF Data Quality Flags

GBIF employs a large set of data quality flags to provide users with an assessment of the reliability and fitness for use of the aggregated data. These flags are assigned based on various quality-related criteria, such as geographic precision, taxonomic accuracy, and completeness of associated metadata. By analyzing these flags, users can make informed decisions regarding the suitability of data for their specific research or conservation needs. GBIF data quality flags serve as valuable indicators that facilitate data filtering and enhance data usability. These flags are assigned to individual occurrence records based on various quality-related criteria and are designed to help data users evaluate the trustworthiness and suitability of the data for their specific needs.

Country Centroid filter

Retrospective geocoding is a process in which historical occurrence data lacking precise geographic coordinates are assigned approximate mappable locations based on other available information, such as locality descriptions, gazetteers, or administrative boundaries. Retrospective geocoding introduces centroids into occurrence datasets wherever better information than a large named area is unknown. A Country centroid refers to the geographical center point of a country's geographic polygon.

Recently, GBIF has implemented country centroid filtering as a feature. By utilizing this filtering technique, GBIF users can identify and exclude occurrences that are country or area centroids. This ensures that data users receive more accurate and reliable information, reducing potential errors resulting from georeferencing issues.

The Catalogue of Centroids is a GitHub repository designed to collect and provide a comprehensive list of country centroids for the purpose of occurrence data filtering. The repository serves as a central resource for storing and sharing centroid coordinates associated with each country, area, or province. The main idea is to gather the from sources most often used during retrospective geo-referencing of museum collections.

Fig. 1

Default Filtering and Treatment of Suspicious Points

GBIF employs default filtering mechanisms to identify suspicious occurrences from its aggregated datasets. For example, the "Country Coordinate Mismatch" data quality flag in GBIF is a quality assessment indicator that highlights occurrences where there is a potential mismatch between the recorded coordinates and the associated country information. Similarly, zero coordinates are often encountered as placeholders or missing data points. GBIF's default filtering process identifies such occurrences from the dataset, as they can introduce inaccuracies when conducting spatial analyses or mapping exercises. This step ensures that only reliable and valid occurrences are available for further analyses and downstream applications.

GBIF continues to improve data quality flagging, which significantly enables users to assess the fitness of data for a particular use and increases the utility of the mediated biodiversity information.

Keywords

data quality, biodiversity, country centroid filtering, data quality flags, spatial errors, georeferencing, data filtering, GBIF

Presenting author

John Thomas Waller

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

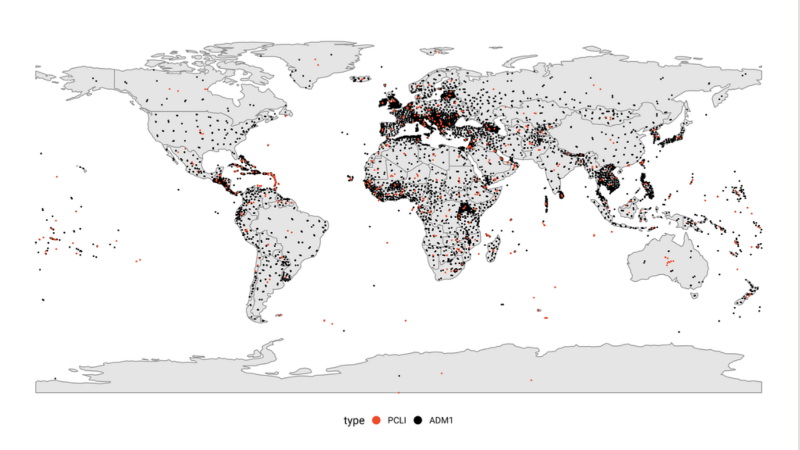

Figure 1.

This figure presents a map displaying the centroid coordinates of various countries, areas, or provinces. The centroids serve as reference locations for occurrence data filtering. **PCLI** means places with an iso-code. **ADM1** means roughly provinces, states, gadm1.