BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35829
doi: 10.3897/biss.3.35829

OPEN ACCESS

Conference Abstract

Data Location Quality at GBIF

John Waller ‡

‡ GBIF, Copenhagen, Denmark

Corresponding author: John Waller (jwaller@gbif.org)

Received: 29 Apr 2019 | Published: 13 Jun 2019

Citation: Waller J (2019) Data Location Quality at GBIF. Biodiversity Information Science and Standards 3: e35829.
https://doi.org/10.3897/biss.3.35829

Abstract

I will cover how the **Global Biodiversity Information Facility** (GBIF) handles **data quality issues**, with specific focus on coordinate location issues, such as gridded datasets (Fig. 1) and country centroids. I will highlight the challenges GBIF faces identifying potential data quality problems and what we and others (Zizka et al. 2019) are doing to discover and address them.

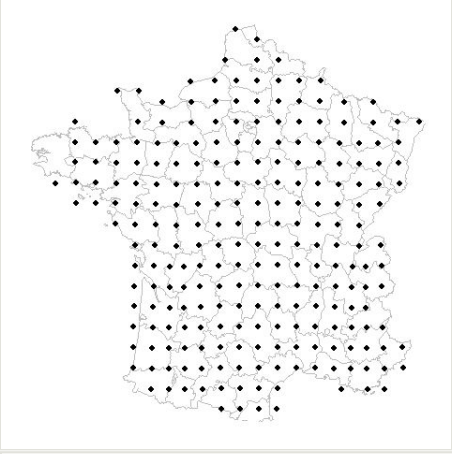

Figure 1.

A gridded dataset animation illustrating the expecation of users versus the reality of underlying occurrence data.

GBIF is the largest open-data portal of biodiversity data, which is a large network of individual datasets (> 40k) from various sources and publishers. Since these datasets are variable both within themselves and dataset-to-dataset, this creates a challenge for users wanting to use data collected from museums, smartphones, atlases, satellite tracking, DNA sequencing, and various other sources for research or analysis.

Data quality at GBIF will always be a moving target (Chapman 2005), and GBIF already handles many obvious errors such as zero/impossible coordinates, empty or invalid data fields, and fuzzy taxon matching. Since GBIF primarily (but not exclusively) serves lat-lon location information, **there is an expectation that occurrences fall somewhat close to where the species actually occurs**. This is not always the case. Occurrence data can be hundereds of kilometers away from where the species naturally occur, and there can be multiple reasons for why this can happen, which might not be entirely obvious to users.

One reasons is that many GBIF datasets are **gridded**. Gridded datasets are datasets that have low resolution due to equally-spaced sampling. This can be a data quality issue because a user might assume an occurrence record was recorded exactly at its coordinates. **Country centroids** are another reason why a species occurrence record might be far from where it occurs naturally. GBIF does not yet flag country centroids, which are records where the dataset publishers has entered the lat-long center of a country instead of leaving the field blank. I will discuss the challenges surrounding locating these issues and the current solutions (such as the CoordinateCleaner R package).

I will touch on how existing DWCA terms like coordinateUncertaintyInMeters and footprintWKT are being utilized to highlight low coordinate resolution. Finally, I will highlight some other emerging data quality issues and how GBIF is beginning to experiment with **dataset-level flagging**. Currently we have flagged around **500 datasets as gridded** and around **400 datasets as citizen science**, but there are many more potential dataset flags.

Keywords

GBIF, Data Quality, Gridded Datasets, Country Centroids, CoordinateCleaner, coordinateUncertaintyInMeters, footprintWKT

Presenting author

John Waller

References

- Chapman A (2005) Principles of Data Quality. Global Biodiversity Information Facility https://doi.org/10.15468/DOC.JRGG-A190

- Zizka A, Silvestro D, Andermann T, Azevedo J, Ritter CD, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, Svantesson S, Wengström N, Zizka V, Antonelli A (2019) CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods in Ecology and Evolution https://doi.org/10.1111/2041-210x.13152