GBIF Data Processing and Validation

John Waller[‡], Nikolay Volik[‡], Federico Mendez[‡], Andrea Hahn[‡]

‡ Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark

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

Abstract

GBIF (Global Biodiversity Information Facility) is the largest data aggregator of biological occurrences in the world. GBIF was officially established in 2001 and has since aggregated 1.8 billion occurrence records from almost 2000 publishers. GBIF relies heavily on Darwin Core (DwC) for organising the data it receives.

GBIF Data Processing Pipelines

Every single occurrence record that gets published to GBIF goes through a series of three processing steps until it becomes available on GBIF.org.

1. source downloading
2. parsing into verbatim occurrences
3. interpreting verbatim values

Once all records are available in the standard verbatim form, they go through a set of interpretations.

In 2018, GBIF processing underwent a significant rewrite in order to improve speed and maintainablility. One of the main goals of this rewrite was to improve the consistency between GBIF's processing and that of the Living Atlases. In connection with this, GBIF's current data validator fell out of sync with GBIF pipelines processing.

New GBIF Data Validator

The current GBIF data validator is a service that allows anyone with a GBIF-relevant dataset to receive a report on the syntactical correctness and the validity of the content contained within the dataset. By submitting a dataset to the validator, users can go through the validation and interpretation procedures usually associated with publishing in GBIF and quickly determine potential issues in data, without having to publish it. GBIF is planning to rework the current validator because the current validator does not exactly match current GBIF pipelines processing.

Planned Changes

The new validator will match the processing of the GBIF pipelines project.

- Validations will be saved and show up on user pages similar to the way downloads and derived datasets appear now (no more bookmarking validations!)
- A downloadable report of issues found will be produced.

Suggested Changes/Ideas

One of the main guiding philosophies for the new validator user interface will be avoiding information overload. The current validator is often quite verbose in its feedback, highlighting data issues that may or may not be fixable or particularly important. The new validator will:

- generate a map of record geolocations;
- give users issues by order of importance;
- give "What", "Where", "When" flags priority;
- give some possible solutions or suggested fixes for flagged records.

We see the hosted portal environment as a way to quickly implement a pre-publication validation environment that is interactive and visual.

Potential New Data Quality Flags

The GBIF team has been compiling a list of new data quality flags. Not all of the suggested flags are easy to implement, so GBIF cannot promise the flags will get implemented, even if they are a great idea. The advantage of the new processing pipelines is that almost any new data quality flag or processing step in pipelines will be available for the data validator.

Easy new potential flags:

- **country centroid flag**: Country/province centroids are a known data quality problem.
- **any zero coordinate flag**: Sometimes publishers leave either the latitude or longitude field as zero when it should have been left blank or NULL.
- **default coordinate uncertainty in meters flag:** Sometimes a default value or code is used for dwc:coordinateUncertaintyInMeters, which might indicate that it is incorrect. This is especially the case for values 301, 3036, 999, 9999.
- **no higher taxonomy flag:** Often publishers will leave out the higher taxonomy of a record. This can cause problems for matching to the GBIF backbone taxonomy..
- **null coordinate uncertainty in meters flag:** There has been some discussion that GBIF should encourage publishers more to fill in dwc:coordinateUncertaintyInMeters. This is because every record, even ones taken from a Global Positioning System (GPS) reading, have an associated dwc:coordinateUncertaintyInMeters

It is also nice when a data quality flag has an escape hatch, such that a data publisher can get rid of false positives or remove a flag through filling in a value.

Batch-type validations that are doable for pipelines, but probably not in the validator include:

- **outlier:** Outliers are a known data quality problem. There are generally two types of outliers: environmental outliers and distance outliers. Currently GBIF does not flag either type of outlier.
- **record is sensitive species:** A sensitive species would be a record where the species is considered vulnerable in some way. Usually this is due to poaching threat or the species is only found in one area.
- **gridded dataset:** Rasterized or gridded datasets are common on GBIF. These are datasets where location information is pinned to a low-resolution grid. This is already available with an experimental API (Application Programming Interface).

Conclusion

Data quality and data processing are moving targets. Variable source data will always be an issue when aggregating large amounts of data. With GBIF's new processing architecture, we hope that new features and data quality flags can be added more easily. Time and staffing resources are always in short supply, so we plan to prioritise the feedback we give to publishers, in order for them to work on correcting the most important and fixable issues. With new GBIF projects like the vocabulary server, we also hope that GBIF data processing can have more community participation.

Keywords

tool, API, data quality, issue flag, workflow, reporting, hosted portals, data publication, occurrence data, publisher support, Darwin Core

Presenting author

John Waller

Presented at

TDWG 2021

Conflicts of interest