# GBIF Integration of Open Data

Tim Robertson‡, Federico Mendez‡, Matthew Blissett‡, Morten Høfft‡, Thomas Stjernegaard Jeppese‡, Nikolay Volik‡, Marcos Lopez Gonzalez‡, Mikhail Podolskiy‡, Markus Döring‡

‡ Global Biodiversity Information Facility, Copenhagen, Denmark

Corresponding author: Tim Robertson (trobertson@gbif.org)

## Abstract

The Global Biodiversity Information Facility (GBIF) runs a global data infrastructure that integrates data from more than 1700 institutions. Combining data at this scale has been achieved by deploying open Application Programming Interfaces (API) that adhere to the open data standards provided by Biodiversity Information Standards (TDWG). In this presentation, we will provide an overview of the GBIF infrastructure and APIs and provide insight into lessons learned while operating and evolving the systems, such as long-term API stability, ease of use, and efficiency. This will include the following topics:

- The registry component provides RESTful APIs for managing the organizations, repositories and datasets that comprise the network and control access permissions. Stability and ease of use have been critical to this being embedded in many systems.
- Changes within the registry trigger data crawling processes, which connect to external systems through their APIs and deposit datasets into GBIF's central data warehouse. One challenge here relates to the consistency of data across a distributed network.
- Once a dataset is crawled, the data processing infrastructure organizes and enriches data using reference catalogues accessed through open APIs, such as the vocabulary server and the taxonomic backbone. Being able to process data quickly as source data and reference catalogues change is a challenge for this component.
- The data access APIs provide search and download services. Asynchronous APIs are required for some of these aspects, and long-term stability is a requirement for widespread adoption. Here we will talk about policies for schema evolution to avoid incompatible changes, which would cause failures in client systems.
- The APIs that drive the user interface have specific needs such as efficient use of the network bandwidth. We will present how we approached this, and how we are currently adopting GraphQL as the next generation of these APIs.

- There are several APIs that we believe are of use for the data publishing community. These include APIs that will help in data quality aspects, and new data of interest thanks to the data clustering algorithms GBIF deploys.

## Keywords

GBIF, API, GraphQL, occurrence, Darwin Core, species, schema, biodiversity

## Presenting author

Matthew Blissett

## Presented at

TDWG 2021

## Conflicts of interest