GBIF Global Biodiversity Information Facility

Cloud Services

John Waller | jwaller@gbif.org

https://techdocs.gbif.org/en/cloud-services/

Ways to get GBIF mediated occurrence records **…**

Downloads

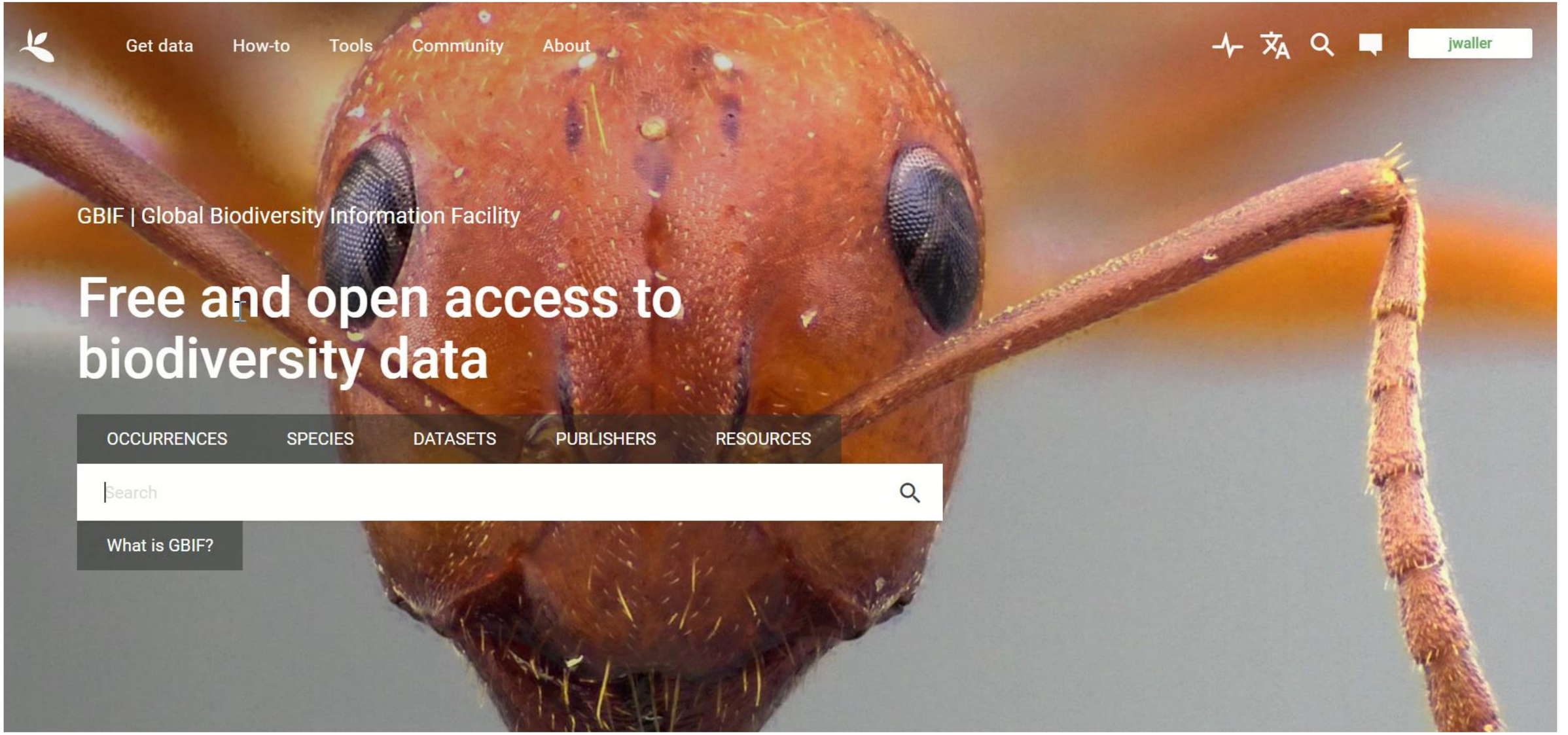
Occurrence search API

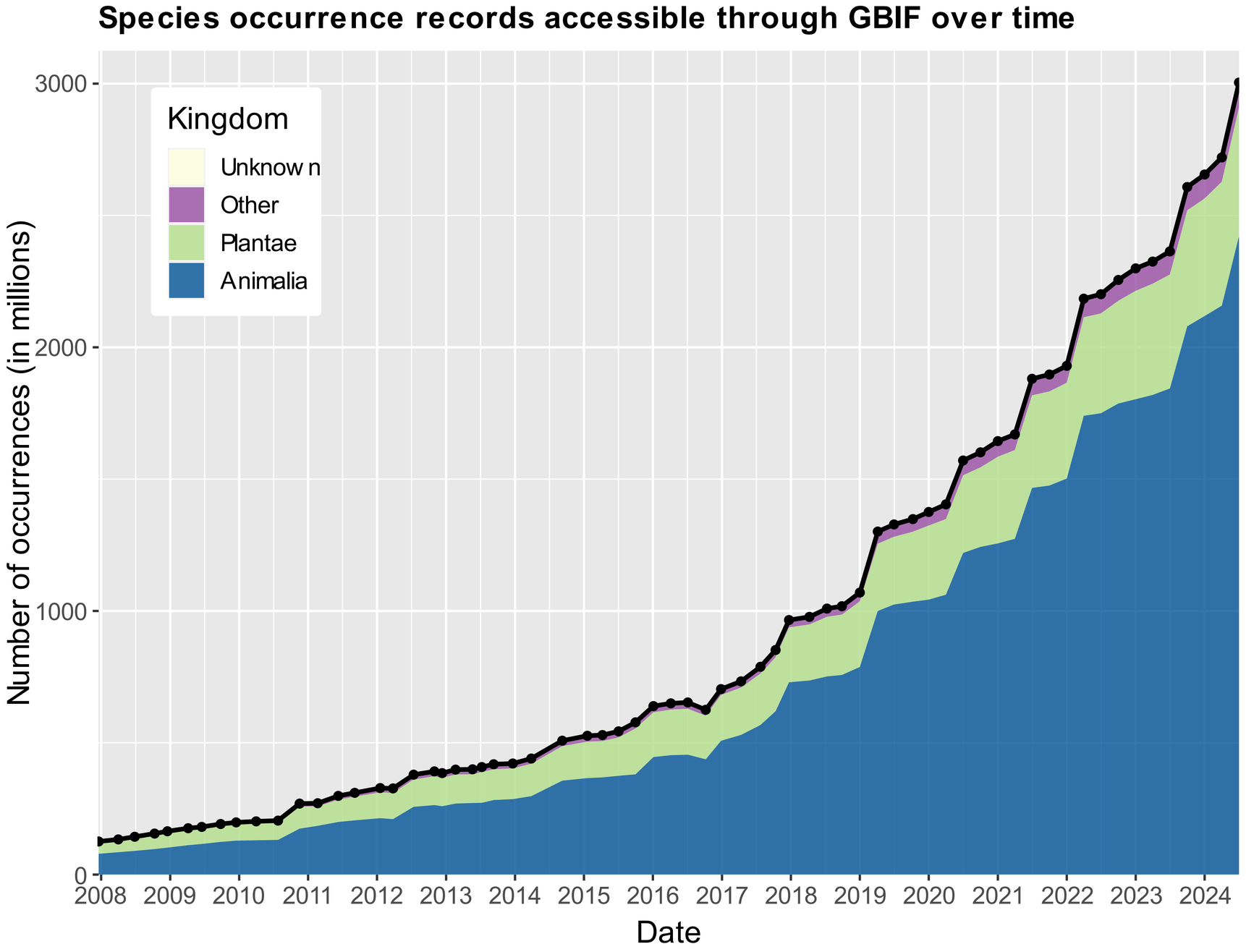
- rgbif
- pygbif

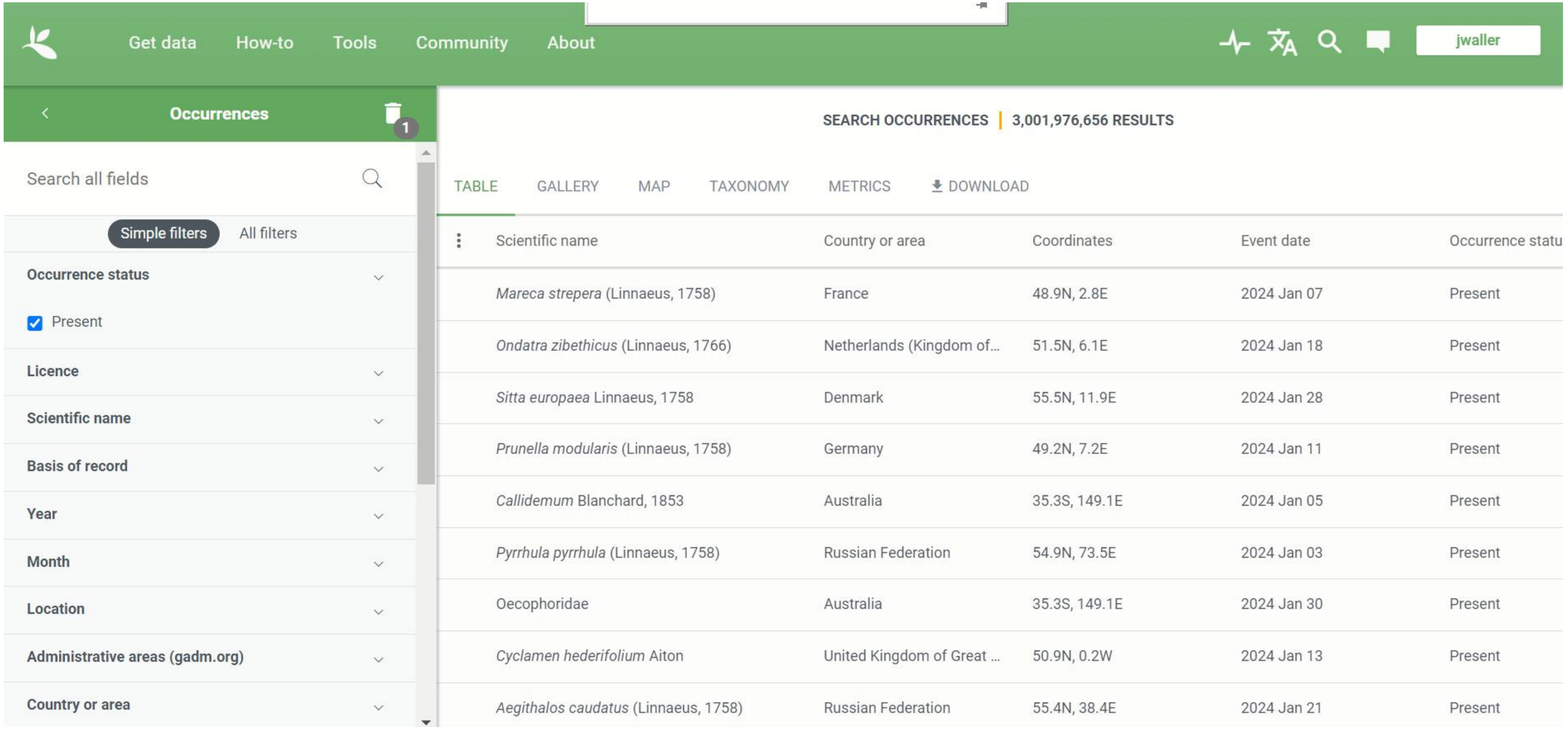
GBIF Cloud Exports

GBIF SQL Downloads

https://techdocs.gbif.org/en/data-use/
https://www.gbif.org/composition/4TlmnRvvPs2RxrPvLH6mOa/data-use-club-practical-session-3-recording-and-resources

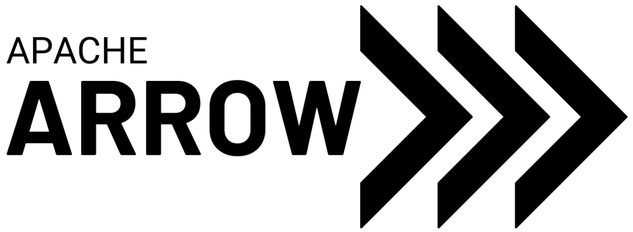

Get data
How-to
Tools
Community
About
jwaller
GBIF | Global Biodiversity Information Facility
Free and open access to biodiversity data
OCCURRENCES
SPECIES
DATASETS
PUBLISHERS
RESOURCES
Search
What is GBIF?


Species occurrence records accessible through GBIF over time
Kingdom
Unknow n
Other
Plantae
Animalia
Number of occurrences (in millions)
3000
2000
1000
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Date


Get data
How-to
Tools
Community
About
jwaller
Occurrences
1
Search all fields
Simple filters
All filters
Occurrence status
Present
Licence
Scientific name
Basis of record
Year
Month
Location
Administrative areas (gadm.org)
Country or area
SEARCH OCCURRENCES | 3,001,976,656 RESULTS
TABLE
GALLERY
MAP
TAXONOMY
METRICS
DOWNLOAD
Scientific name
Country or area
Coordinates
Event date
Occurrence statu
Mareca strepera (Linnaeus, 1758)
France
48.9N, 2.8E
2024 Jan 07
Present
Ondatra zibethicus (Linnaeus, 1766)
Netherlands (Kingdom of...
51.5N, 6.1E
2024 Jan 18
Present
Sitta europaea Linnaeus, 1758
Denmark
55.5N, 11.9E
2024 Jan 28
Present
Prunella modularis (Linnaeus, 1758)
Germany
49.2N, 7.2E
2024 Jan 11
Present
Callidemum Blanchard, 1853
Australia
35.3S, 149.1E
2024 Jan 05
Present
Pyrrhula pyrrhula (Linnaeus, 1758)
Russian Federation
54.9N, 73.5E
2024 Jan 03
Present
Oecophoridae
Australia
35.3S, 149.1E
2024 Jan 30
Present
Cyclamen hederifolium Aiton
United Kingdom of Great ...
50.9N, 0.2W
2024 Jan 13
Present
Aegithalos caudatus (Linnaeus, 1758)
Russian Federation
55.4N, 38.4E
2024 Jan 21
Present


Big Data
(something that doesn’t fit on your laptop)

Big Data **Tools**



Occurrence Snapshots

Periodic exports of GBIF occurrence data

Every month GBIF takes a full occurrence snapshot, saved in different formats to ease usage. All snapshots are issued with a DOI to simplify citation, and some formats are copied to public clouds for easy use on those environments.

Cloud-based datasets

GBIF makes data available on the Microsoft Planetary Computer (Azure), as an Amazon AWS Open Dataset and on a public Google GCS bucket and BigQuery table. When using cloud-based snapshots, we always recommend creating a Derived Dataset citation for the records that you use. When referring to the full dataset, please use the appropriate citation found below.

Date	Format	Citation	Filters
01 September 2024	Simple Parquet	GBIF.org (01 September 2024) GBIF Occurrence Data https://doi.org/10.15468/dl.v4njrj	
01 August 2024	Simple Parquet	GBIF.org (01 August 2024) GBIF Occurrence Data https://doi.org/10.15468/dl.t56p6n	

https://www.gbif.org/occurrence-snapshots

GBIF exports full snapshots to …

1. Google
2. Microsoft
3. Amazon

https://techdocs.gbif.org/en/cloud-services/

Google Cloud | My First Project

Product details

GBIF Species Occurrences

BigQuery Public Data

Global-scale records of organisms at a given time and place

VIEW DATASET

OVERVIEW | SAMPLES | RELATED PRODUCTS

Overview

GBIF—the Global Biodiversity Information Facility—is an international network and data infrastructure funded by the world's governments providing global data that document the occurrence of species. GBIF integrates datasets from around the world and currently documents more than two billion species occurrences. The GBIF occurrence dataset combines data

Additional details

Type: Data

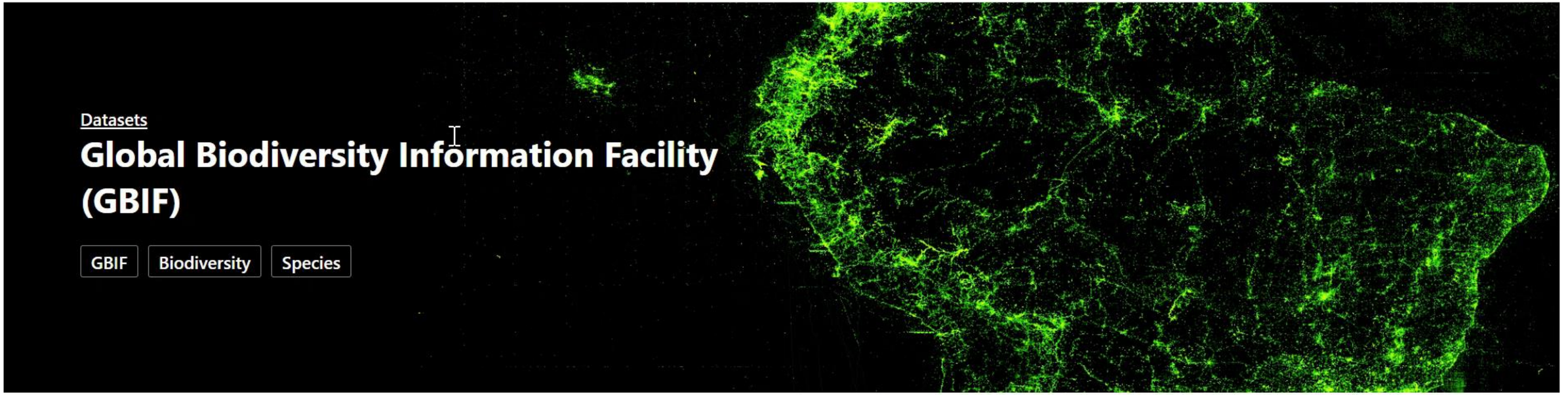
Category: Science & research

Dataset source: GBIF Species Occurrence snapshots

https://console.cloud.google.com/marketplace/product/bigquery-public-data/gbif-occurrences?pli=1&project=nodal-reserve-251311

https://earthengine.google.com/

Microsoft | Planetary Computer Explore Data Catalog Applications Documentation



Datasets

Global Biodiversity Information Facility (GBIF)

GBIF Biodiversity Species

Overview Example Notebook

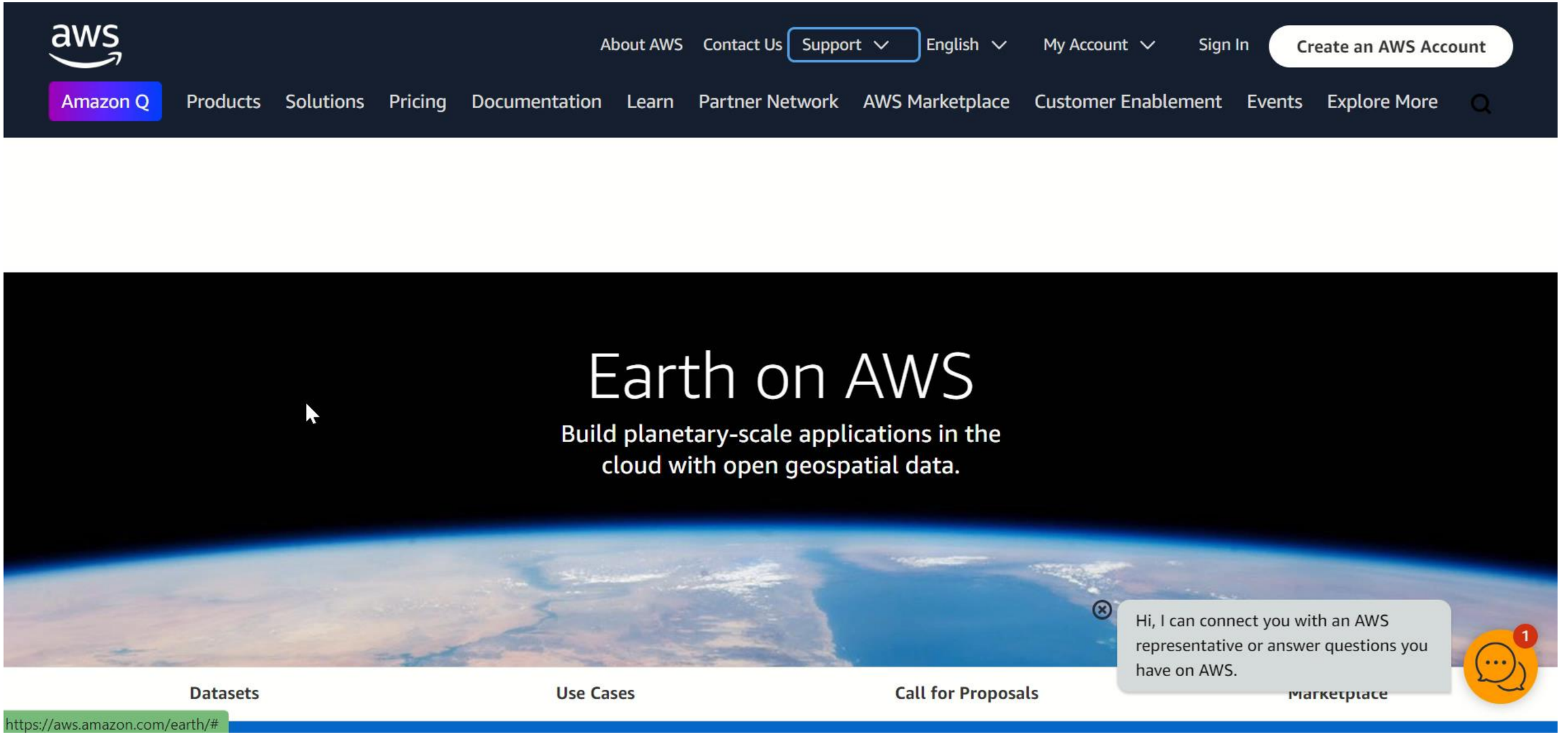
Overview

The Global Biodiversity Information Facility (GBIF) is an international network and data infrastructure funded by the world's governments, providing global data that document the occurrence of species. GBIF currently integrates datasets documenting over 1.6 billion species occurrences.

The GBIF occurrence dataset combines data from a wide array of sources, including specimen-related data from natural history museums

Spatial Extent

https://planetarycomputer.microsoft.com/dataset/gbif



https://aws.amazon.com/earth/

data blog

posts community-forum gbif.org about

GBIF and Apache-Spark on AWS tutorial

John Waller

2021-06-04 · GBIF

GBIF now has a snapshot of 1.3 billion occurrence† records on **Amazon Web Services** (AWS). This guide will take you through running **Spark notebooks** on AWS. The GBIF snapshot is documented : here.

https://data-blog.gbif.org/post/aws-and-gbif/
https://data-blog.gbif.org/post/microsoft-azure-and-gbif/

New Phytologist

Research

Plant diversity darkspots for global collection priorities

Ian Ondo[1,2]*, Kiran L. Dhanjal-Adams[1]*, Samuel Pironon[1,2,3]*, Daniele Silvestro[4,5], Matheus Colli-Silva[1], Victor Deklerck[1,6], Olwen M. Grace[1,7], Alexandre K. Monro[1], Nicky Nicolson[1], Barnaby Walker[1] and Alexandre Antonelli[1,5,8]

[1]Royal Botanic Gardens, Kew, Richmond, TW9 3AE, UK; [2]UN Environment Programme World Conservation Monitoring Centre (UNEP-WCMC), Cambridge, CB3 0DL, UK; [3]School of Biological and Behavioural Sciences, Queen Mary University of London, London, E1 4DQ, UK; [4]Department of Biology, University of Fribourg, Fribourg, 1700, Switzerland; [5]Department of Biological and Environmental Sciences, Gothenburg Global Biodiversity Centre, University of Gothenburg, Gothenburg, 41319, Sweden; [6]Meise Botanic Garden, Meise, 1860, Belgium; [7]Royal Botanic Garden Edinburgh, Edinburgh, EH3 5LR, UK; [8]Department of Biology, University of Oxford, Oxford, OX1 3RB, UK

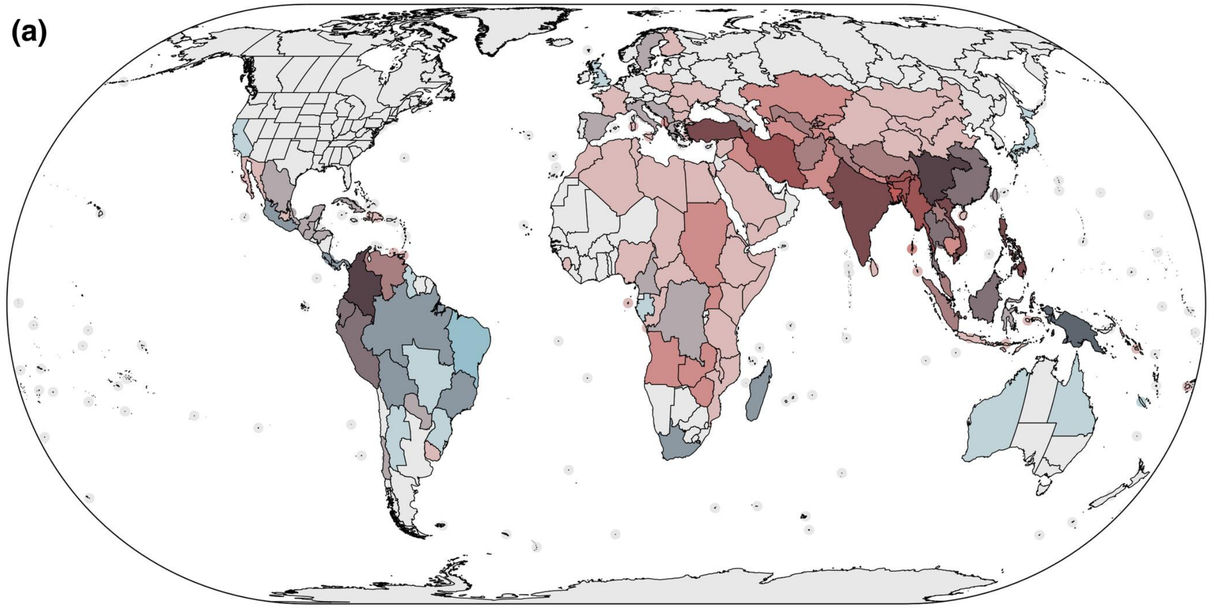
Authors for correspondence:
Samuel Pironon
Email: s.pironon@kew.org

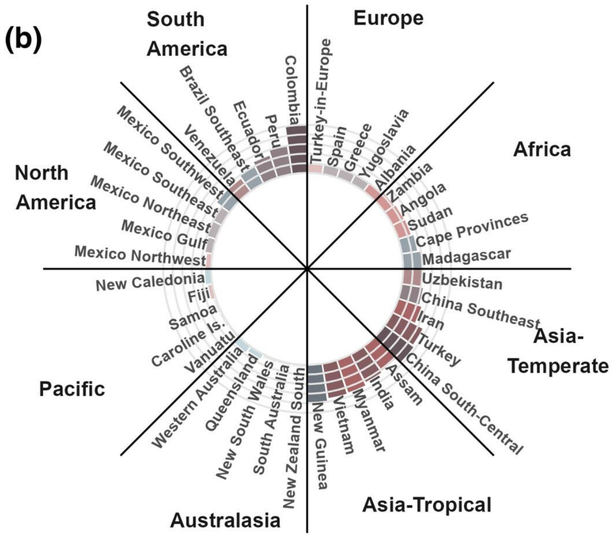
Alexandre Antonelli
Email: a.antonelli@kew.org

Summary

• More than 15% of all vascular plant species may remain scientifically undescribed, and many of the >350 000 described species have no or few geographic records documenting their distribution. Identifying and understanding taxonomic and geographic knowledge shortfalls is key to prioritising future collection and conservation efforts.

• Using extensive data for 343 523 vascular plant species and time-to-event analyses, we conducted multiple tests related to plant taxonomic and geographic data shortfalls, and iden-



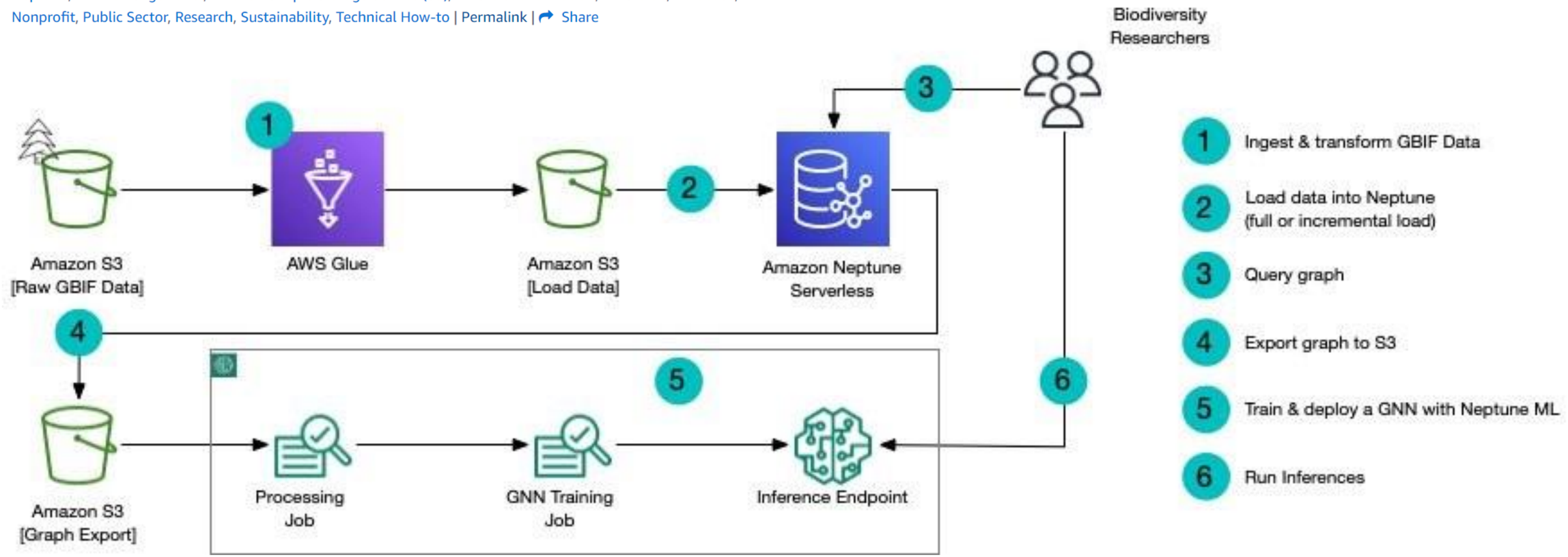


https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.20024

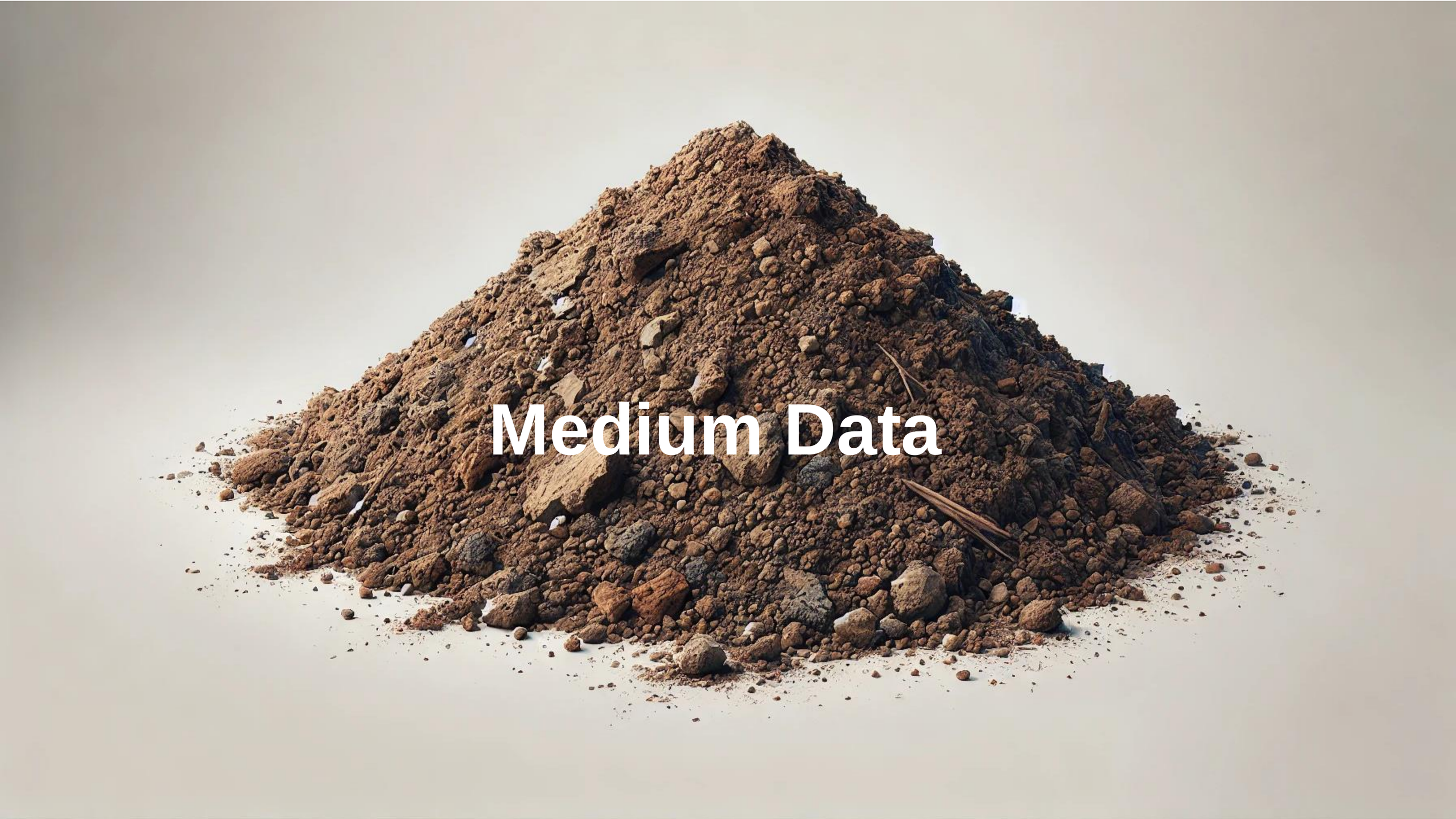
AWS Public Sector Blog

Hydrating the Natural History Museum's Planetary Knowledge Base with Amazon Neptune and Open Data on AWS

by Nishant Casey, Ilan Gleiser, Karsten Schroer, and Sam Bydlon | on 13 SEP 2024 | in Amazon Machine Learning, Amazon Neptune, Amazon SageMaker, Amazon Simple Storage Service (S3), AWS CloudFormation, AWS Glue, Database, Nonprofit, Public Sector, Research, Sustainability, Technical How-to | Permalink | Share



https://aws.amazon.com/blogs/publicsector/hydrating-the-natural-history-museums-planetary-knowledge-base-with-amazon-neptune-and-open-data-on-aws/


Medium Data

GBIF saves the snapshots it exports in a **columnar data** format known as **Parquet**. This format allows for **certain types** of queries to run very quickly.

Parquet contains row group level statistics that contain the minimum and maximum values for each column chunk. Queries that fetch specific column values need not read the entire row data thus improving performance.

https://data-blog.gbif.org/post/apache-arrow-and-parquet/

Run a query on your laptop with R

The R package **arrow** allows large queries to run locally by only downloading the parts of the dataset necessary to perform the query.

```
# get occurrence counts from all species in Sweden since 1990
library(arrow)
library(dplyr)
gbif_snapshot <- "s3://gbif-open-data-eu-central-1/occurrence/2021-11-01/occurrence.parquet"
df <- open_dataset(gbif_snapshot)
df %>%
filter(
countrycode == "SE",
class == "Mammalia",
year > 1990
) %>%
group_by(species) %>%
count() %>%
collect()
```

https://data-blog.gbif.org/post/apache-arrow-and-parquet/

Download your own local parquet files from GBIF

Local parquet files will allow **apache arrow** queries to run much faster.

```
library(rgbif)
library(arrow)

# all Botswana occurrences
occ_download(pred("country", "BW"),format = "SIMPLE_PARQUET")
# unzip files first…
arrow::open_dataset("occurrence.parquet")
```

https://data-blog.gbif.org/post/apache-arrow-and-parquet/

Register a Derived Dataset (With DOI)

Derived datasets are a new citation feature on GBIF. Derived datasets are citable records of GBIF-mediated occurrence data.

To register a derived dataset, you will need to create a simple text file with two columns:

1. A GBIF datasetkey (uuid)
2. A count of the number of occurrences from each dataset

datasetkey	Count
4fa7b334-ce0d-4e88-aaae-2e0c138d049e	213
906e6978-e292-4a8b-9c39-adf6bb0f3323	35

https://data-blog.gbif.org/post/derived-datasets/

GBIF

SQL Downloads

This is an **experimental feature**, and the implementation may change throughout 2024. The feature is currently only available for preview by **invited users**. Contact helpdesk@gbif.org to **request access**.

Features of GBIF SQL Downloads

- Access to **most** SQL statements
- Grouped-by counts and other aggregations
- + 400 columns available
- Reduce the size of large queries with **select statements**
- **Citable DOI** that gives attribution to all publishers without needing a derived dataset

https://techdocs.gbif.org/en/data-use/api-sql-downloads

Basic Usage

query.json

```
{
    "sendNotification": true,
    "notificationAddresses": [
        "userEmail@example.org"
    ],
    "format": "SQL_TSV_ZIP",
    "sql": "SELECT datasetKey, countryCode, COUNT(*) FROM occurrence WHERE continent = 'EUROPE' GROUP BY datasetKey, countryCode"
}
```

CURL

```
curl --include --user YOUR_GBIF_USERNAME:YOUR_PASSWORD --header "Content-Type: application/json" --data @query.json
https://api.gbif.org/v1/occurrence/download/request
```

https://api.gbif.org/v1/occurrence/download/request

DOWNLOAD | 27 JUNE 2024

2,851 occurrences included in download

DOI 10.15468/dl.nxmesk

DOWNLOAD

PLEASE USE THIS CITATION IN PUBLICATIONS

GBIF.org (27 June 2024) GBIF Occurrence Download https://doi.org/10.15468/dl.nxmesk

Copy BibTex RIS TELL US ABOUT USAGE

FILTER APPLIED 27 JUNE 2024

Licence: CC0 1.0

File: 86 KB SQL TSV zip

Involved datasets: 74,648

Using rgbif

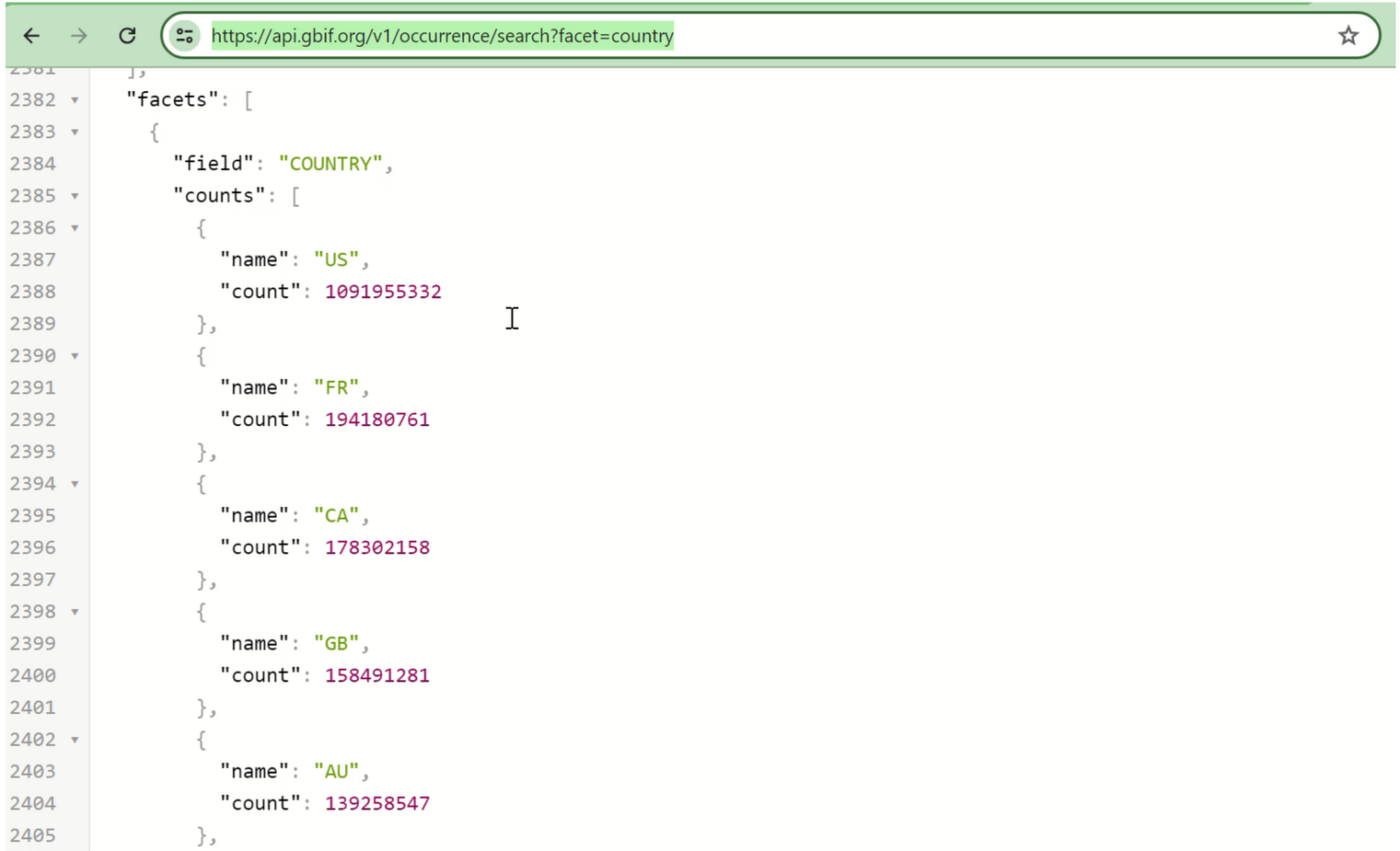
```
install_github("ropensci/rgbif", ref = "occ_download_sql")


library(rgbif)


occ_download_sql("SELECT datasetKey, countryCode, COUNT(*) FROM
occurrence WHERE continent = 'EUROPE' GROUP BY datasetKey,
countryCode")
```

https://github.com/ropensci/rgbif

Don’t Forget about Facet Queries



https://api.gbif.org/v1/occurrence/search?facet=country

Final thoughts …

Reasons to use public cloud computing

1. If you are already familiar with some system (AWS, Google, Microsoft).
2. You have a complicated model that requires a lot of computing power that cannot be reduced before the compute stage.
3. Flexibility and freedom.
4. Combine **GBIF mediated data** with existing **spatial layers** or datasets.

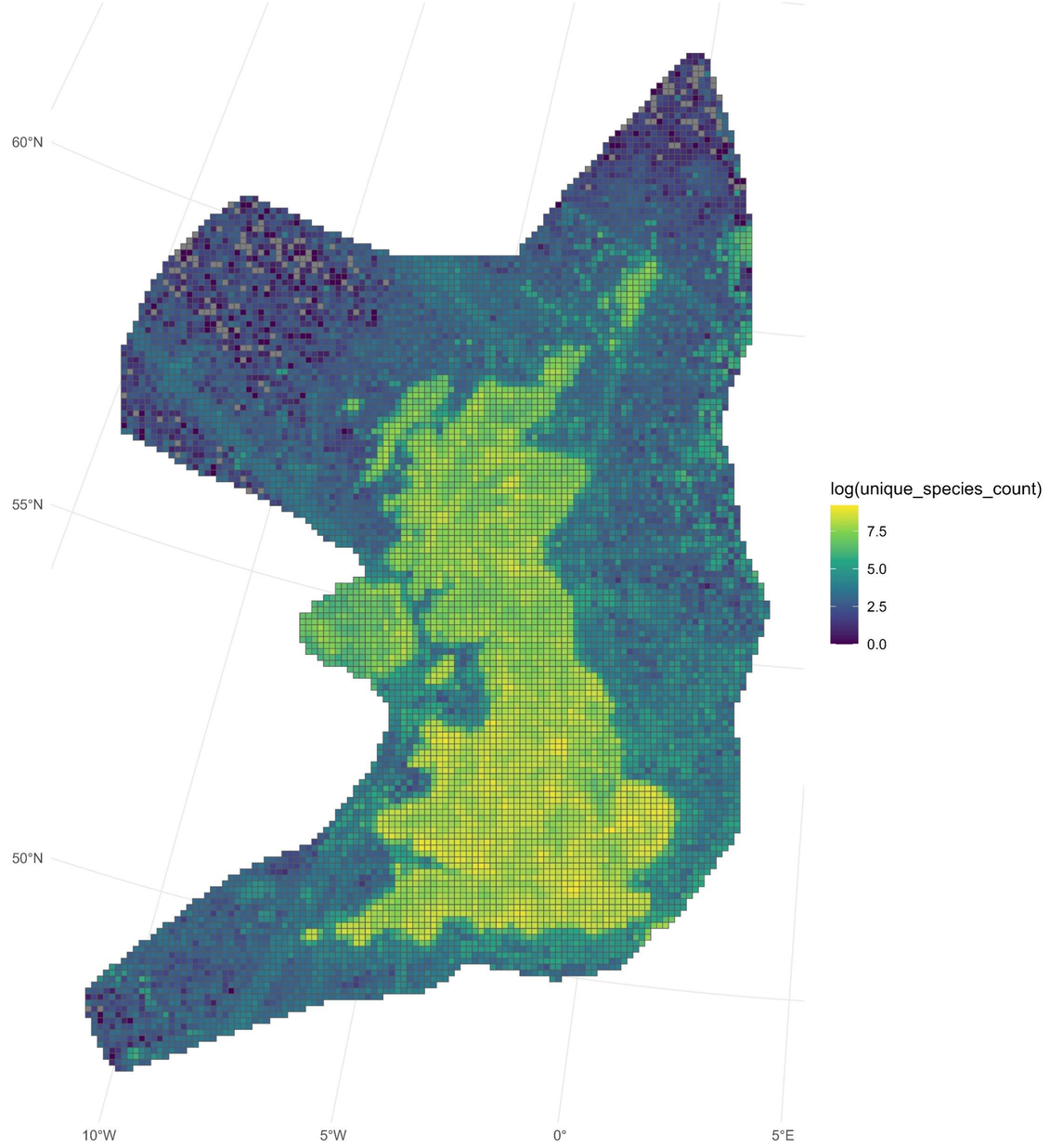
 Only invest the **time** and **money** into setting up a cloud computing system if **SQL downloads** and **traditional downloads** don’t work for you.

Useful links

Amazon
https://registry.opendata.aws/gbif/
Google
https://earthengine.google.com/
https://console.cloud.google.com/storage/browser/public-datasets-gbif
https://console.cloud.google.com/marketplace/product/bigquery-public-data/gbif-occurrences?project=nodal-reserve-251311
Microsoft
https://planetarycomputer.microsoft.com/dataset/gbif

Apache Arrow
https://data-blog.gbif.org/post/apache-arrow-and-parquet/
SQL Downloads
https://data-blog.gbif-uat.org/post/2024-06-24-gbif-sql-downloads/
https://techdocs.gbif.org/en/data-use/api-sql-downloads
Traditional GBIF Downloads
https://techdocs.gbif.org/en/data-use/api-downloads
https://docs.ropensci.org/rgbif/articles/getting_occurrence_data.html
Facets
https://techdocs.gbif.org/en/openapi/v1/occurrence#/Searching%20occurrences/searchOccurrence
https://api.gbif.org/v1/occurrence/search?facet=country

```
SELECT
  GBIF_EEARGCode(
    10000,
    decimalLatitude,
    decimalLongitude,
    COALESCE(coordinateUncertaintyInMeters, 0)
  ) AS cellcode,
  COUNT(DISTINCT speciesKey) AS
unique_species_count
FROM
  occurrence
GROUP BY
  cellcode
```



https://techdocs.gbif.org/en/data-use/data-cubes
https://sdi.eea.europa.eu/data/93315b78-089d-43a5-ac76-b3df627b2e4cf