

## Biodiversity Data Use Module 1: Data Processing

Andrew Rodrigues | Programme Officer



Global Biodiversity  
Information Facility

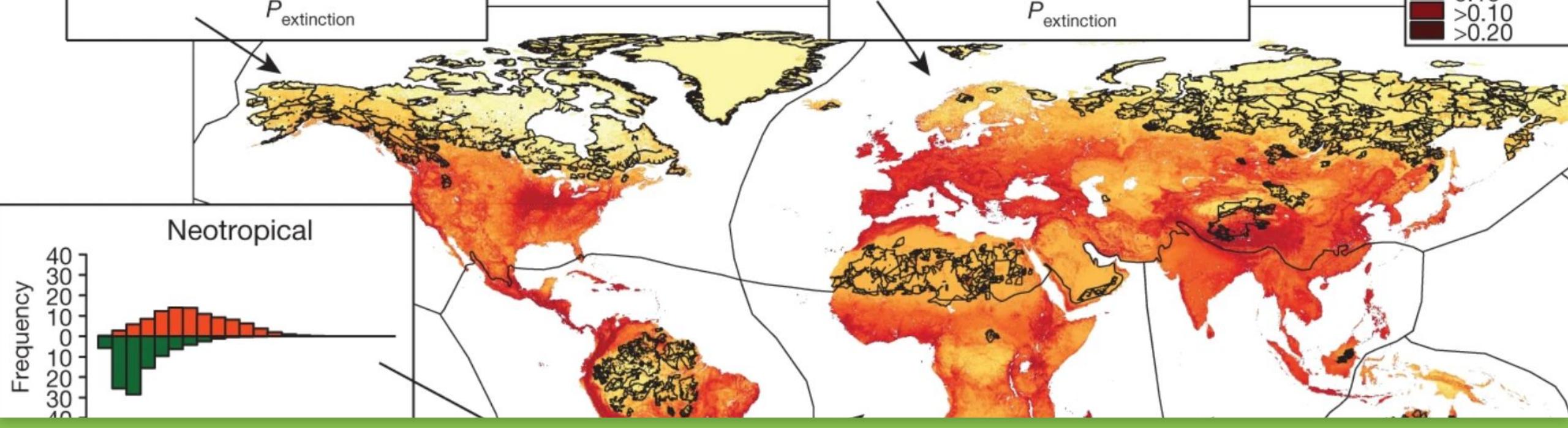
$P_{\text{extinction}}$

Afrotropical

rosette lichen (*Physcia tenella*), Isenvad, Jutland, Denmark, 6 January 2019. Photo by Lars Jørgen Grønbjerg  
CC BY-NF-NC via Danish Mycological Society <https://www.gbif.org/occurrence/2233587087>

Indomalay

Australasia



## COURSE OUTLINE

- Part 1 - Navigating [www.gbif.org](http://www.gbif.org)
- Part 2- Common Data Quality Issues
- Part 3- The API
- Part 4 - Using R

Resources: <https://docs.gbif.org/course-data-use/en/key-documentation.html>

Community Forum: [https://discourse.gbif.org/g/BID\\_DataUse](https://discourse.gbif.org/g/BID_DataUse)

# Trainers and Mentors

## Trainers



Andrew Rodrigues – GBIF Secretariat, Programme Office for Participation and Engagement



John Waller - GBIF Secretariat, Data Analyst

## Mentors



Anabela Plos  
Museo Argentino de Ciencias Naturales Bernardino Rivadavia



Arman Pili  
Monash University



Leonardo Buitrago  
GBIF Caribbean Regional Support Contractor

Vijay Barve  
University of Florida

# Free and open access to biodiversity data

Occurrences

Species

Datasets

Publishers

Resources

Search



WHAT IS GBIF? | ABOUT GBIF DENMARK

*Montastraea cavernosa* observed in Cayman Islands by Kerry Lewis (CC-BY-NC 4.0)

Occurrence records  
1,676,825,999

Datasets  
57,757

Publishing institutions  
1,665

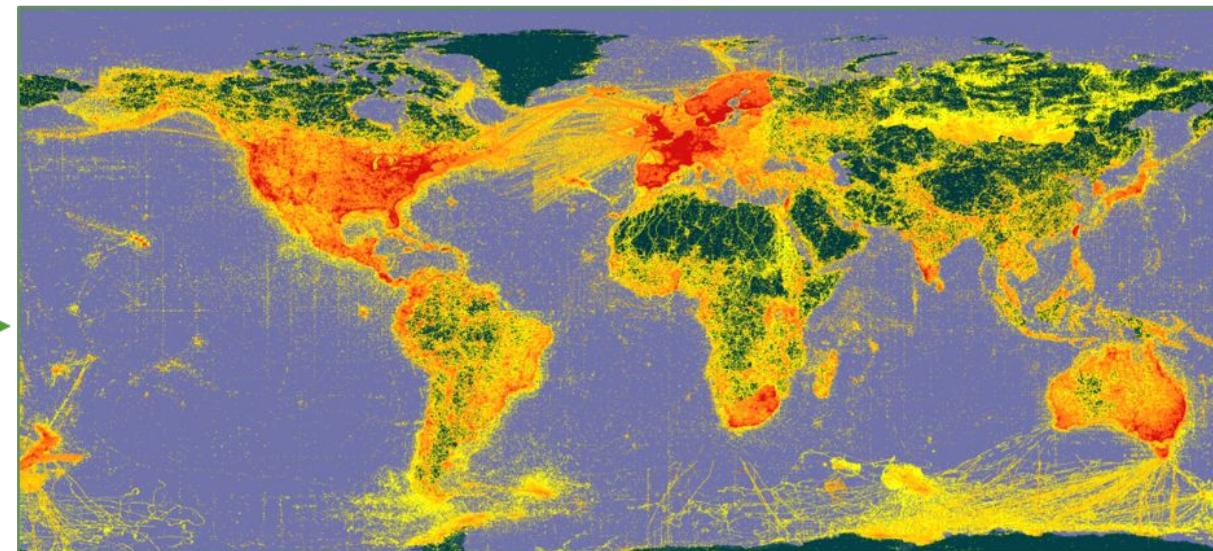
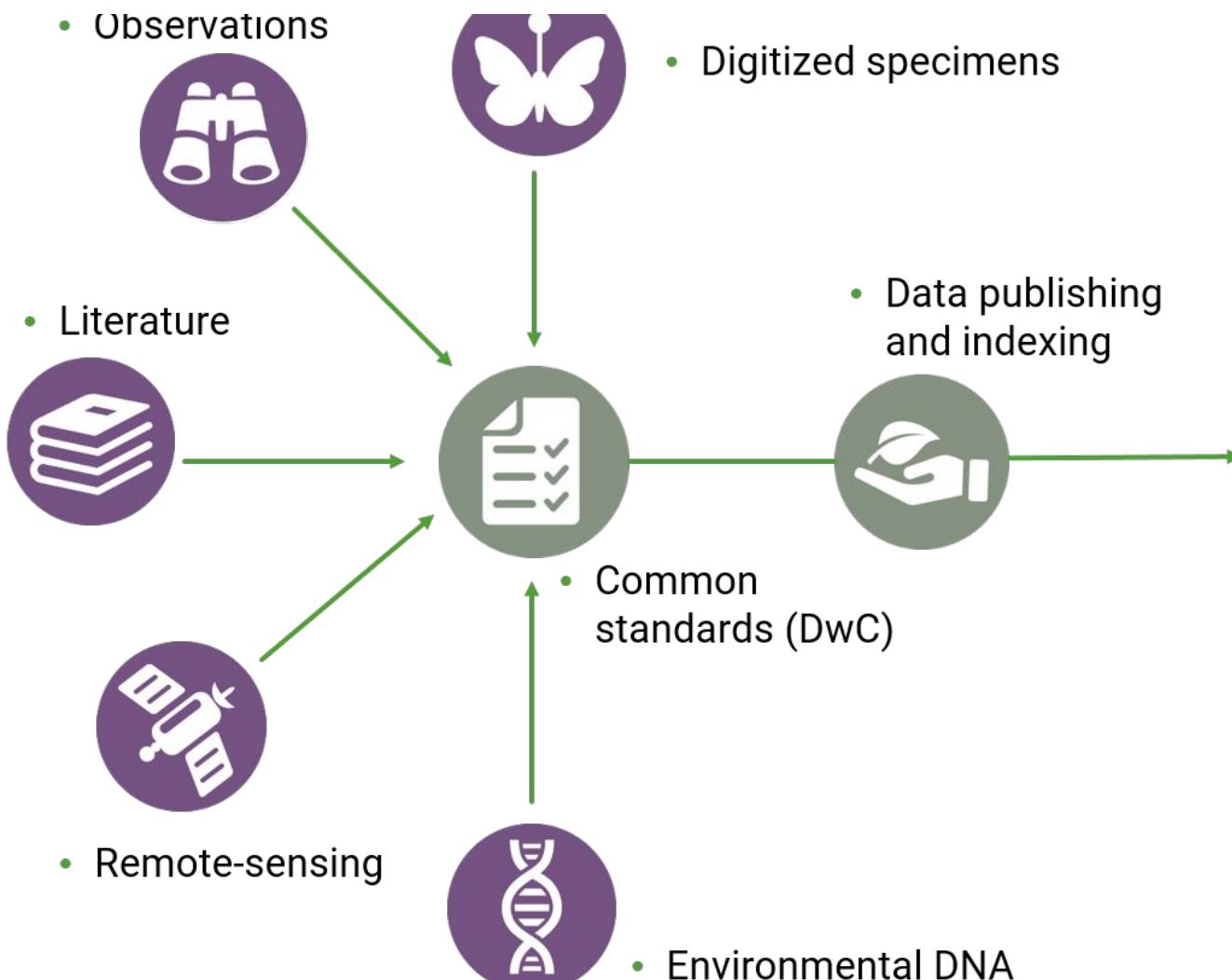
Peer-reviewed papers using data  
5,703

## WHY DO WE PROCESS DATA?

We want to make our download fit for our own purposes

1. Remove erroneous data e.g. outliers
2. Ensure sufficient level of precision in the data for our purpose

# A WINDOW ON EVIDENCE ABOUT WHERE SPECIES HAVE LIVED, AND WHEN



- Data discovery and use

# Free and open access to biodiversity data

Occurrences

Species

Datasets

Publishers

Resources

Search



WHAT IS GBIF? | ABOUT GBIF DENMARK

*Montastraea cavernosa* observed in Cayman Islands by Kerry Lewis (CC-BY-NC 4.0)

Occurrence records  
1,676,825,999

Datasets  
57,757

Publishing institutions  
1,665

Peer-reviewed papers using data  
5,703

## WHY DO WE PROCESS DATA?

Each time you process a dataset for use you will have to consider

1. Requirements of your analysis
2. Balance between data quality and the robustness of your analysis

This may be an iterative process

# Free and open access to biodiversity data

Occurrences

Species

Datasets

Publishers

Resources

Search



WHAT IS GBIF? ABOUT GBIF DENMARK

*Montastraea cavernosa* observed in Cayman Islands by Kerry Lewis (CC BY-NC 4.0)

Occurrence records  
1,676,825,999

Datasets  
57,757

Publishing institutions  
1,665

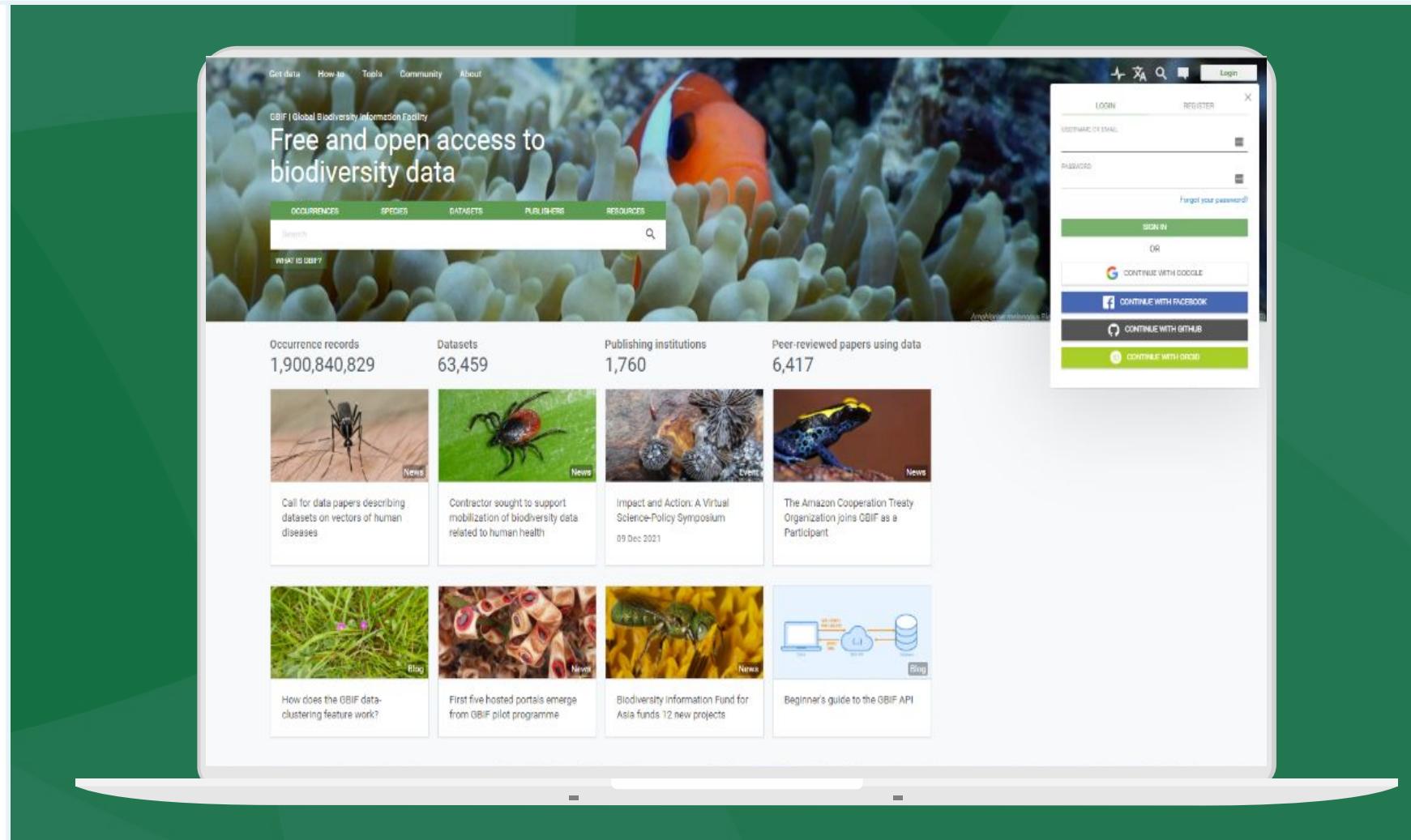
Peer-reviewed papers using data  
5,703

## GOLDEN RULES OF GBIF-MEDIATED DATA USE

1. Must have an account on [www.gbif.org](http://www.gbif.org)
2. Must agree to the Data User Agreement - <https://www.gbif.org/terms/data-user>
3. Document how you process your data
4. Correctly cite the data you use
5. Deposit used data in a public repository

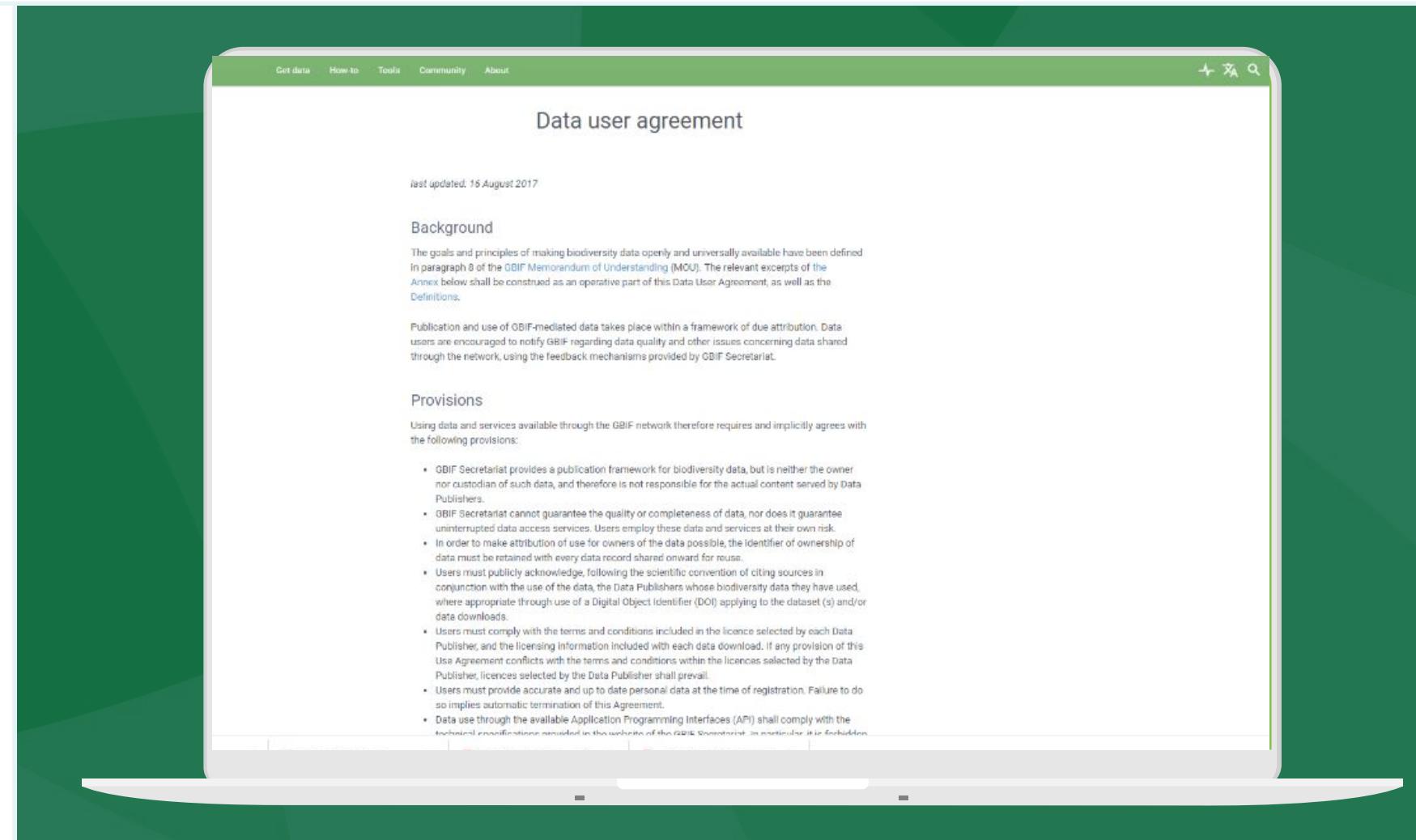
# GOLDEN RULES OF GBIF-MEDIATED DATA USE

- Must have an account on [www.gbif.org](http://www.gbif.org)



# GOLDEN RULES OF GBIF-MEDIATED DATA USE

- Must have an account on [www.gbif.org](http://www.gbif.org)
- Must agree to the Data User Agreement - <https://www.gbif.org/terms/data-user>
  - Non-binding
  - Sets out guiding principles of data use including citation of data



# GOLDEN RULES OF GBIF-MEDIATED DATA USE

- Must have an account on [www.gbif.org](http://www.gbif.org)
- Must agree to the Data User Agreement - <https://www.gbif.org/terms/data-user>
- Document how you process your data

## Step 1

Download occurrence records for species x with an associated DOI

## Step 2

Remove all records from outside its native range

## Step 3

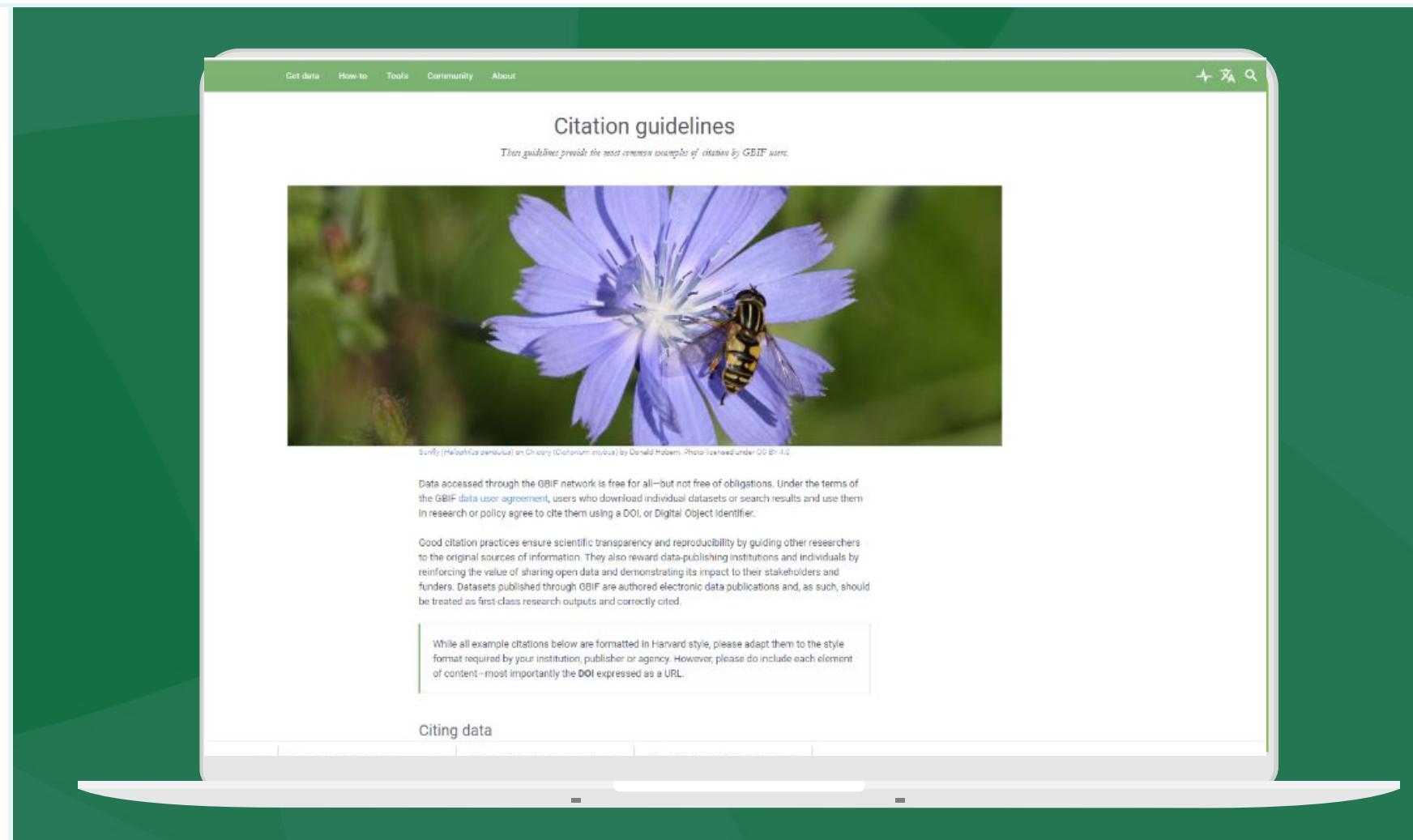
Remove all records collected before 1950

## Step 4

Final clean dataset

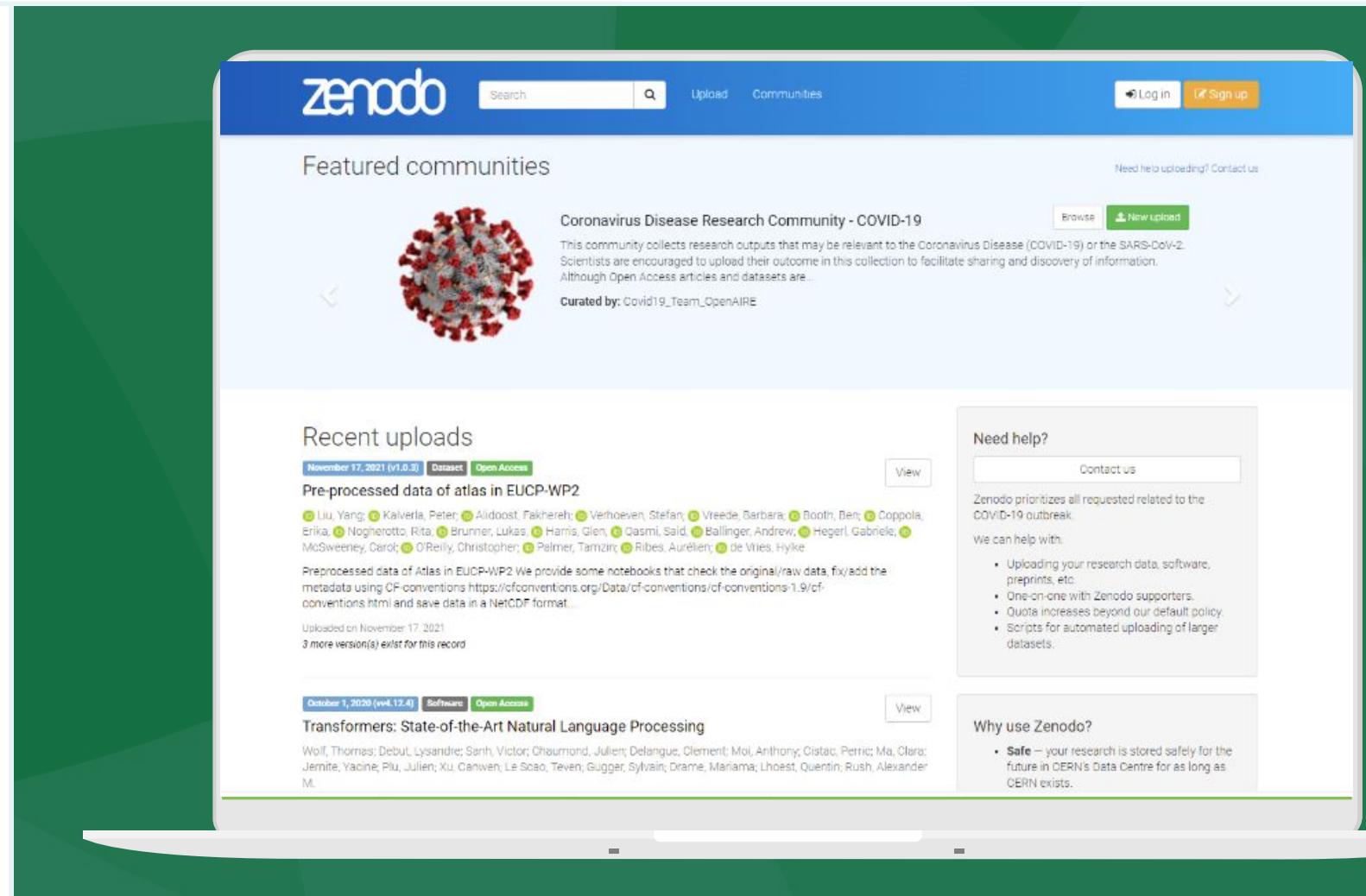
# GOLDEN RULES OF GBIF-MEDIATED DATA USE

- Must have an account on [www.gbif.org](https://www.gbif.org)
- Must agree to the Data User Agreement - <https://www.gbif.org/terms/data-user>
- Document how you process your data
- Correctly cite the data you use
  - Guidelines- <https://www.gbif.org/citation-guidelines>
  - Derived dataset DOIs - <https://www.gbif.org/derived-dataset/about>



# GOLDEN RULES OF GBIF-MEDIATED DATA USE

- Must have an account on [www.gbif.org](https://www.gbif.org)
- Must agree to the Data User Agreement - <https://www.gbif.org/terms/data-user>
- Document how you process your data
- Correctly cite the data you use
  - Guidelines: <https://www.gbif.org/citation-guidelines>
  - Derived dataset DOIs
- Deposit used data in a public repository e.g. Zenodo



	Raw data	Interpreted data	Multimedia	Coordinates	Format	Estimated data size
<b>⬇ SIMPLE</b>	X	✓	X	✓ (if available)	Tab-delimited CSV <small>?</small>	<b>1 MB</b> (167 KB zipped for download)
<b>⬇ DARWIN CORE ARCHIVE</b>	✓	✓	✓ (links)	✓ (if available)	Tab-delimited CSV <small>?</small>	<b>3 MB</b> (423 KB zipped for download)
<b>⬇ SPECIES LIST</b>	X	✓	X	X	Tab-delimited CSV <small>?</small>	

## DATA DOWNLOADS

Data can be downloaded in three formats

**Simple:** Tab delimited CSV. Only contains the data after GBIF interpretation. No multimedia included. [More information about CSV](#)

**Darwin Core Archive:** The Darwin Core Archive (DwC-A) contains both the original data as publisher provided it and the GBIF interpretation. Links (but not files) to multimedia included. [More information about DwC-A](#)

**Species list:** Tab delimited CSV with the distinct list of names in the search result.

# Free and open access to biodiversity data

[Occurrences](#)[Species](#)[Datasets](#)[Publishers](#)[Resources](#)[WHAT IS GBIF?](#) | [ABOUT GBIF DENMARK](#)

*Montastraea cavernosa* observed in Cayman Islands by Perry Lewis (CC BY-NC 4.0)

Occurrence records

1,676,825,999



Datasets

57,757



Publishing institutions

1,665



Peer-reviewed papers using data

5,703



## WEBSITE OVERVIEW

[www.gbif.org](http://www.gbif.org)

# Exercise 1: Navigating [www.gbif.org](http://www.gbif.org)

What are the total number of occurrences for Tongatapu islands in Tonga?

How many records for the kingdom Plantae are there on the islands?

How many of these records are from the BID programme?

How many of these records are under a CC-BY licence?

How many have images?

# Exercise 1: Navigating [www.gbif.org](http://www.gbif.org)

What are the total number of occurrences for Tongatapu islands in Tonga? (3,372 - 7th Dec 2021 only GADM, 3,362 - 7th Dec 2021 with GADM and country filter )

How many records for the kingdom Plantae are there on the islands? (459 - 7th Dec2021)

How many of these records are from the BID programme? (36 - 7th Dec2021)

How many of these records are under a CC-BY licence? (0 - 7th Dec2021)

How many have images? (0 - 7th Dec2021)

# Common Data Quality Issues

John Waller | Data Analyst



**GBIF**

Global Biodiversity  
Information Facility

Your **GBIF download** will not always be ‘perfect’ for what you want do with it. **There are a few things you should be aware of...**



Occurrences 1

SEARCH OCCURRENCES | 1,794,432,303 RESULTS

Occurrence status 1

Licence

Scientific name

Basis of record

Location

No preference

Including coordinates

Without coordinates

Include records where coordinates are flagged as suspicious

TABLE GALLERY MAP TAXONOMY METRICS DOWNLOAD

**Default Geospatial Issues Button**

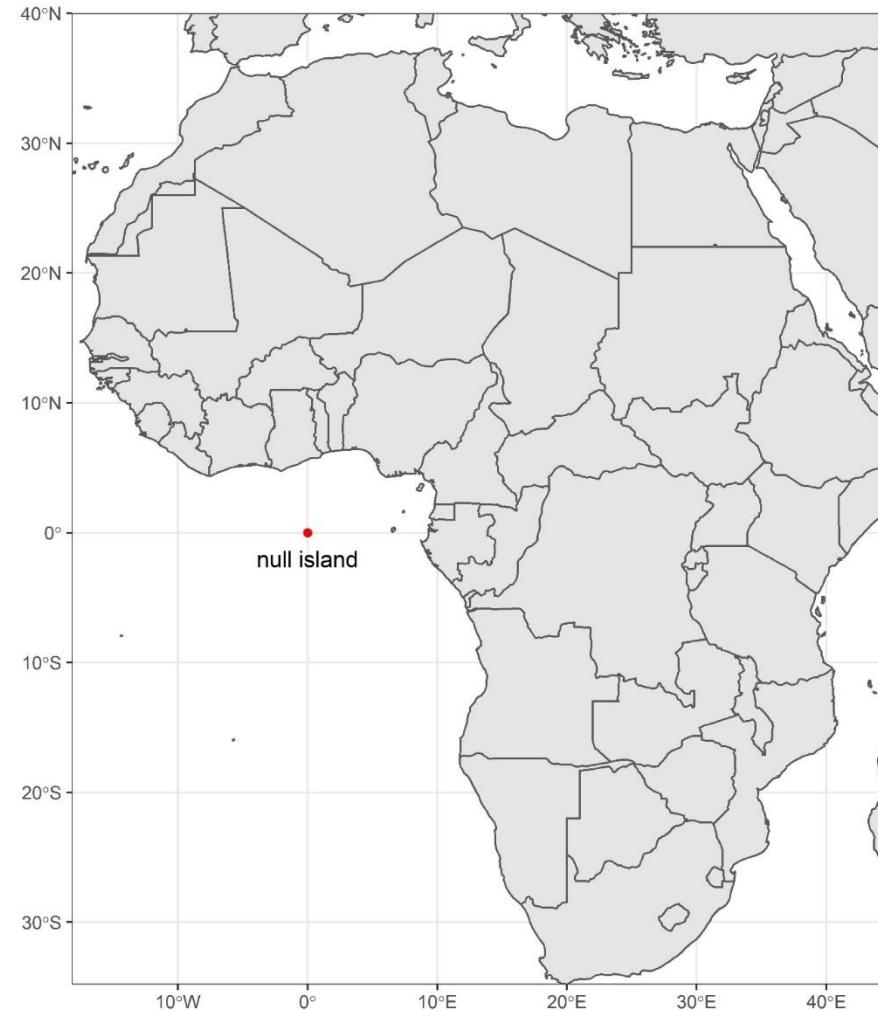
Scientific name	Country or area	Coordinates	Month & year	Basis of record	Dataset
	Viet Nam	21.9N, 104.3E	2021 January	Living specimen	<a href="#">Royal Botanic Garden</a>
	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
Asteraceae	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
<i>Tibouchina</i> Aubl.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
<i>Calibrachoa</i> Cerv.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
<i>Polygala moquiniana</i> A.St.-Hil.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
Cyperaceae	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
<i>Hyptis</i> Jacq.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De</a>
● <i>Belenois java teutonia</i>	Australia	34.9S, 138.6E	2021 January	Preserved specimen	<a href="#">South Australian Mus</a>
● <i>Belenois java teutonia</i>	Australia	34.9S, 138.6E	2021 January	Preserved specimen	<a href="#">South Australian Mus</a>
<i>Evermannella balbo</i> (Risso, 1820)	Spain	41.3N, 2.6E	2021 January	Material sample	<a href="#">Colección de referenc</a>

# Default geospatial issues

GBIF removes common geospatial issues by default if you choose to have data with a location.

- **Zero coordinate** : Coordinates are exactly (0,0). null island
- **Country coordinate mismatch** : The coordinates fall outside of the given country's polygon.
- **Coordinate invalid** : GBIF is unable to interpret the coordinates.
- **Coordinate out of range** : The coordinates are outside of the range for decimal lat/lon values ((-90,90), (-180,180)).

# GBIF removes zero coordinates (0,0) “null island”

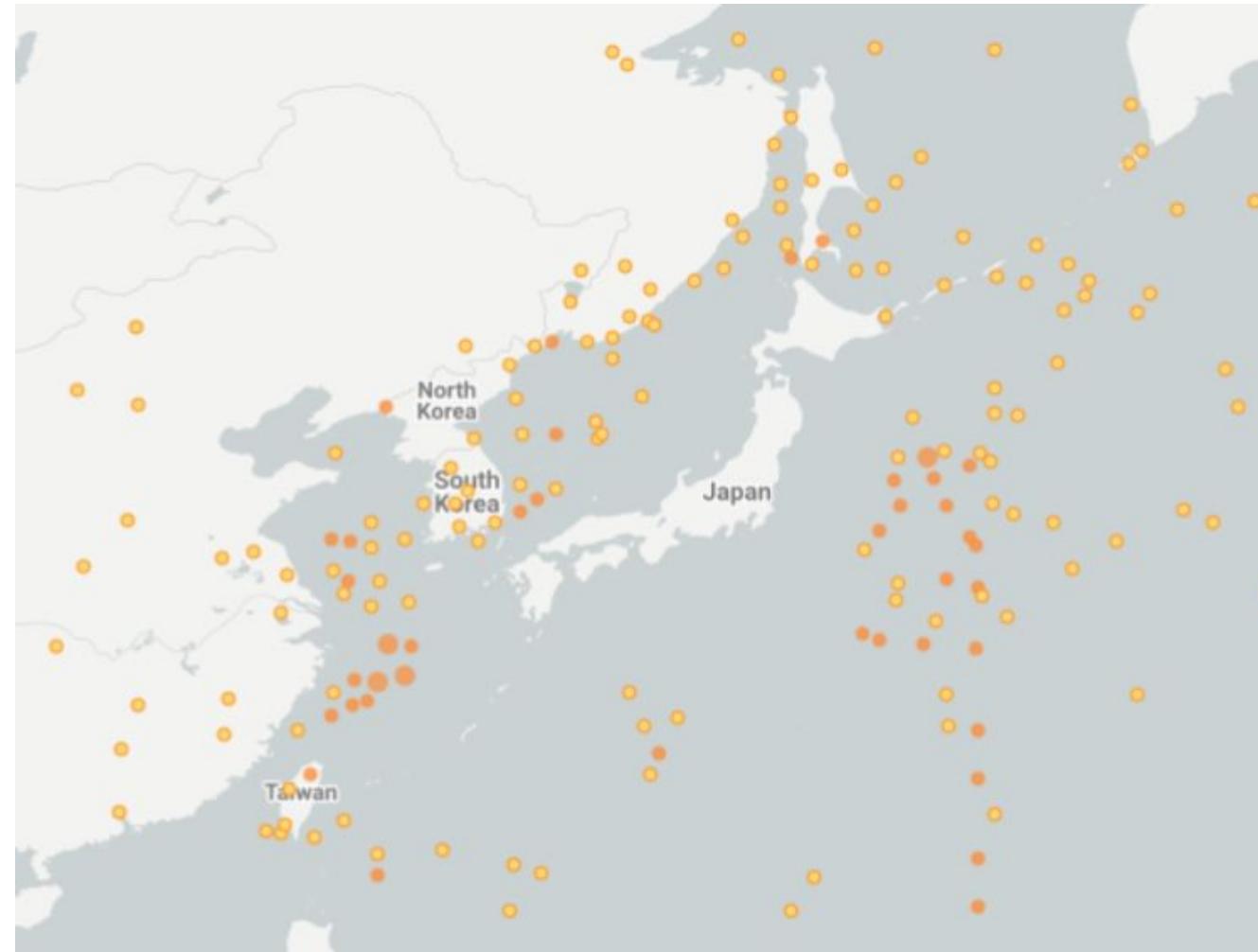


[https://www.gbif.org/occurrence/map?issue=ZERO\\_COORDINATE](https://www.gbif.org/occurrence/map?issue=ZERO_COORDINATE)

# GBIF removes country coordinate mismatch

GBIF removes records that  
do not match their  
countrycode.

All of **these records** claim to  
be located in Japan.



[https://www.gbif.org/occurrence/search?issue=COUNTRY\\_COORDINATE\\_MISMA](https://www.gbif.org/occurrence/search?issue=COUNTRY_COORDINATE_MISMA)  
TCH

# GBIF removes absence records

Sometimes data publishers will include **absence records** (where they verify that a species is not present). Most of users don't want these records.

```
gbif_download %>%
  filter(occurrenceStatus == "PRESENT")
```

[https://www.gbif.org/occurrence/search?occurrence\\_status=present](https://www.gbif.org/occurrence/search?occurrence_status=present)



Occurrences  1

SEARCH OCCURRENCES | 1,902,174,240 RESULTS

TABLE GALLERY MAP TAXONOMY METRICS DOWNLOAD

Simple Advanced

Occurrence status

Present

License

Scientific name

Basis of record

Location

Administrative areas (gadm.org)

Coordinate uncertainty in metres

Year

Month

Dataset

Country or area

Continent

This button is ticked by default

Scientific name	Country or area	Coordinates	Month & year	Basis of record	Dataset
	Viet Nam	21.9N, 104.3E	2021 January	Living specimen	Royal Botanic Garden Edinburgh Li
	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
Asteraceae	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
Tibouchina Aubl.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
Calibrachoa Cerv.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
Polygala moquiniana A.St.-Hil.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
Cyperaceae	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
Hyptis Jacq.	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	FLOR - Herbário do Departamento
Anacardium occidentale L.	Brazil		2021 January	Preserved specimen	MBM - Herbário do Museu Botânic
● Belenois java teutonia	Australia	34.9S, 138.6E	2021 January	Preserved specimen	South Australian Museum Australi
● Belenois java teutonia	Australia	34.9S, 138.6E	2021 January	Preserved specimen	South Australian Museum Australi

**Other issues you have to filter yourself...**

# Fossils and Living Specimens

GBIF has **Fossils** and **Living Specimens** (usually a plant inside a botanical garden or sometimes and animal in a zoo).

```
gbif_download %>%
  filter(!basisOfRecord %in%
  c("FOSSIL_SPECIMEN", "LIVING_SPECIMEN"))
```

# establishmentMeans

**dwc:establishmentMeans** : The process by which the biological individual(s) represented in the Occurrence became established at the location.

```
gbif_download %>%
  filter(!establishmentMeans %in% c("MANAGED",
  "INTRODUCED", "INVASIVE", "NATURALISED"))
```

Unfortunately not used very often.

<https://terms.tdwg.org/wiki/dwc:establishmentMeans>

# Old Records

GBIF has many museum records that might be **older than what is desired** for some studies.

```
gbif_download %>%
  filter(year >= 1900)
```

[https://www.gbif.org/occurrence/search?year=1000,1650&occurrence\\_status=present](https://www.gbif.org/occurrence/search?year=1000,1650&occurrence_status=present)

# Uncertain location example

**Species:** *Lophodytes cucullatus* (Linnaeus, 1758)

**Location:** United States of America

**Basis of record:** Human observation

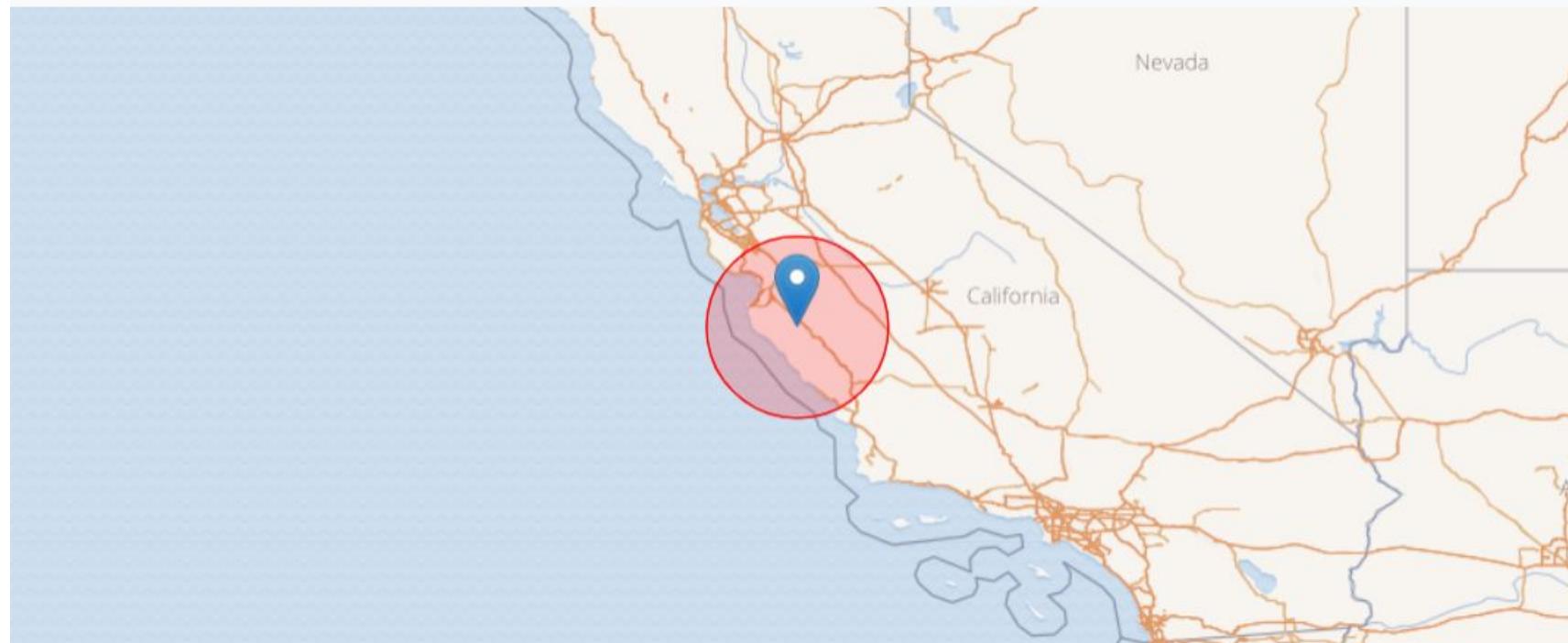


**Dataset:** iNaturalist Research-grade Observations

**Publisher:** iNaturalist.org

**Reference:** <https://www.inaturalist.org/observations/67427035>

**Issues:** Institution match none Collection match none



<https://www.gbif.org/occurrence/3017942707>

# Uncertain location

Often you will want to be sure that the coordinates give a certain location and are not really 1000s of km away from where the organism was observed or collected.

```
gbif_download %>%
  filter(coordinatePrecision > 0.01 |
  is.na(coordinatePrecision)) %>%
  filter(coordinateUncertaintyInMeters < 10000 |
  is.na(coordinateUncertaintyInMeters))
```

I recommend not filtering out missing values, since the value is often not filled in by publishers if they think the occurrence is fairly certain (from a GPS).

# Bad default values for coordinate uncertainty

```
gbif_download %>%
  filter(!coordinateUncertaintyInMeters %in%
  c(301, 3036, 999, 9999))
```

There are a few “fake” values for coordinate uncertainty that you should be aware of. These values are errors produced by geocoding software and do not represent real uncertainty values. In the case of 301, the uncertainty is often much-much greater than 301 and actually represents a **country centroid**.

# Points along the equator or prime meridian

Some publishers consider zero and NULL to be equivalent, empty latitude and longitude end up being plotted along these two lines.

```
gbif_download %>%
  filter(!decimalLatitude == 0 |
  !decimalLongitude == 0)
```

+

-

□

Atlantic Ocean

Mal

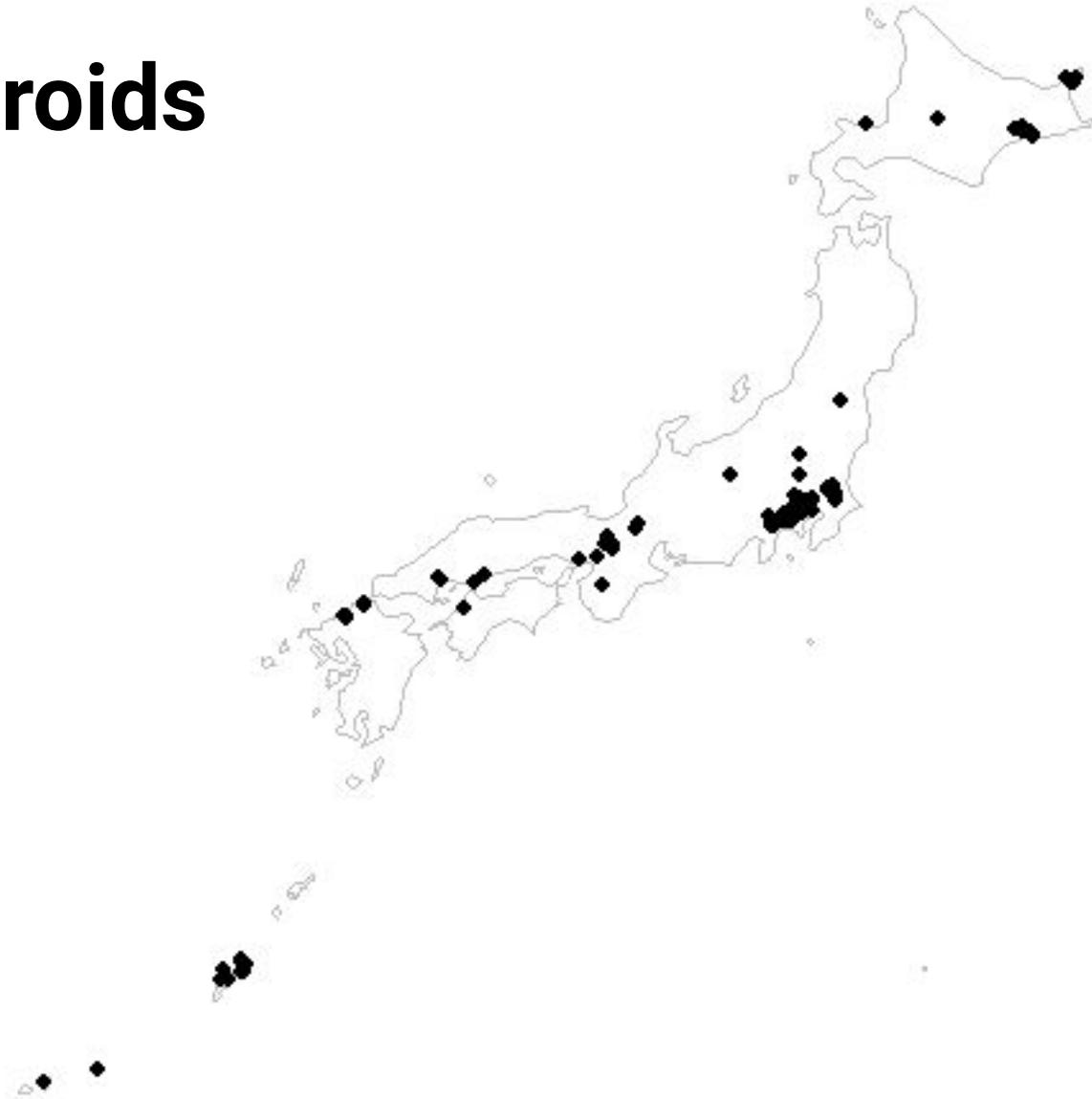
GRNHE

WCO

LAP

Generated 2 days ago © OpenStreetMap contributors, © OpenMapTiles, GBIF

# Country Centroids



# Retrospective geo-coding

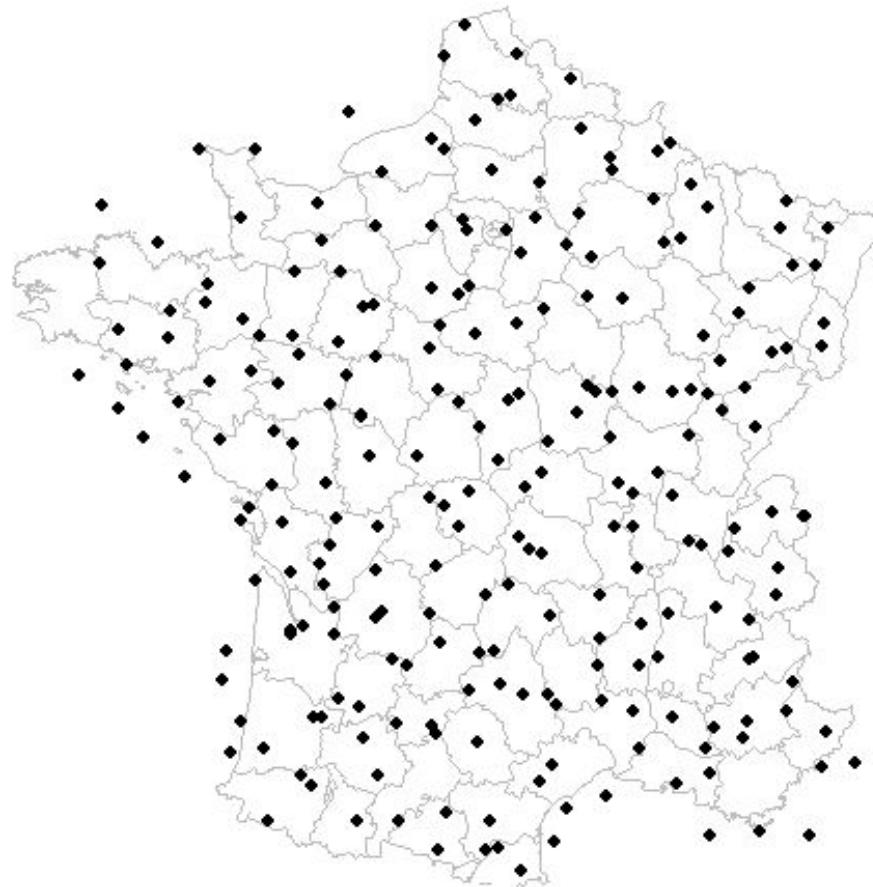
**Retrospective geocoding** is the process where lat-lon values are given to older records which only have **locality information**.

**Locality information** is sometimes only a country, city or a text description like

*“10 miles SW of main road, Austin TX”.*

Often museum records (preserved specimens), will have been retrospectively geocoded.

# Gridded datasets



<https://www.gbif.org/dataset/c779b049-28f3-4daf-bbf4-0a40830819b6>

# Gridded datasets filtering

Most publishers of gridded datasets actually fill in one of the following columns:

- coordinateuncertaintyinmeters
- coordinateprecision
- footprintwkt (only in dwca downloads)

So filtering by these columns **can be a good way to remove gridded datasets.**

GBIF has an [experimental API](#) for identifying datasets which exhibit a certain about of "griddyness". You can read more [here](#).

# Exercise 2: Filtering data for improved data quality

Using [www.gbif.org](http://www.gbif.org) filter the data for *Calopteryx splendens* in the following way:

- *Filter for records with coordinate uncertainty between 0 and 10,000m*
- *Filter for records between 1955 and 2017*
- *Exclude occurrence records where the establishment mean is indicated as managed, introduced or invasive*

How many records did you have at the start?

How many do you have after filtering?

How did you find the taxon?

What are the limitations of the filters?

This API works against the GBIF Occurrence Store, which handles occurrence records and makes them available through the web service and download files. In addition we also provide a Map API that offers spatial services.

internally we use a Java web service client for the consumption of these HTTP-based RESTful web services.

### Occurrences

This API provides services related to the retrieval of single document records.

Resource URL	Method	Response	Description
http://www.example.com/api/v1/resources	GET	200 OK	Get a list of resources

# The GBIF API

Andrew Rodrigues | Programme Officer



# Global Biodiversity Information Facility

61 reasons why you have a hard time

Please be aware that the following parameters are in a experimental phase and its definition could change in the future: `g, facet, facetOffset, facets, limit, facetsInCount` and `facetsInLastCount`.

Resource URL	Method	Response	Description	Paging	Parameters
/occurrence/search	GET	Occurrence	Full search across all occurrences. Results are ordered by relevance.	true	q, basisOfRecord, catalogNumber, classificatory, date, dateAccepted, datePublished, identifier, locality, specimenID, taxonID

# What is the GBIF API?

GBIF Application Programming Interface (API) gives users access to GBIF databases in a **secure way**

Usually the main reason you would want to use an API is because you want **software**, to interact with GBIF somehow

The GBIF API can be accessed via:

- A web browser by visiting a URL e.g. [https://api.gbif.org/v1/species/match?name=Passer domesticus](https://api.gbif.org/v1/species/match?name=Passer%20domesticus)
- Or using a command line program called `curl` that you need to install

GBIF has a few API group/namespace:

- **Registry API** - makes all registered Datasets, Installations, Organizations, Nodes, and Networks discoverable.
- **Species API** - works against data kept in the GBIF Checklist Bank which taxonomically indexes all registered checklist datasets in the GBIF network.
- **Occurrence API** - works against the GBIF Occurrence Store, which handles occurrence records and makes them available through the web service and download files.
- **Maps API** - web map tile service making it straightforward to visualize GBIF content on interactive maps, and overlay content from other sources.
- **Literature API** - search for literature indexed by GBIF, including peer-reviewed papers, citing GBIF datasets and downloads.

# What is the GBIF API?

- 

The **basic pattern** of an API call:

- **base url** : this will always be <https://api.gbif.org/v1/>
- **api** : this is the GBIF API group/namespace you want to query.
- **function** : the functionality you want to use.
- **parameter** : the parameters for your API call. A **?** is sometimes used.
- **query** : the query you fill in. Sometimes will be free text and sometimes will be a predefined argument.

Example

[https://api.gbif.org/v1/species/match?name=Passer domesticus](https://api.gbif.org/v1/species/match?name=Passer%20domesticus)

# Exercise 3 - Finding GBIF taxonkeys

- Taxon keys are issued at the species, genus family, order, phylum and kingdom level.
- Unique identifiers are issued to accepted names with synonyms of those accepted names issued the same identifier. Usage keys vs acceptedusagekeys
- Allow for user to ensure that they are collecting all the data they need
- Also facilitate multiple species downloads
- Taxon keys can be found through:
  - [www.gbif.org](http://www.gbif.org)
  - Species API
  - Species Matching tool - <https://www.gbif.org/tools/species-lookup>

<https://api.gbif.org/v1/species/match?name=Calopteryx%20splendens>

# Exercise 3 - Finding GBIF taxonkeys

Using the Species API find the GBIF **taxonkeys** for these scientific names:

- *Lepus saxatilis* F.Cuvier, 1823
- Aves
- Magnoliophyta
- *Aegithalos caudatus* (Linnaeus, 1758)

What are the taxonomic status of each?

<https://api.gbif.org/v1/species/match?name=Calopteryx%20splendens>

# Exercise 3 - Finding GBIF taxonkeys

Using the Species API find the GBIF **taxonkeys** for these scientific names:

- *Lepus saxatilis* F.Cuvier, 1823 = 2436775 (ACCEPTED)
- 
- Aves = 212 (ACCEPTED) Note: that this is not a species, so if you want occurrences for an entire group you just need one taxonkey and not a list of every species in the group.
- 
- Magnoliophyta = 49 (SYNONYM). Note: that if you used the API directly GBIF will give you the acceptedUsageKey : 7707728, which is Tracheophyta. This is the accepted name of this group. If you decided to just use the old name, you would be missing millions of occurrences so be careful.
- 
- *Aegithalos caudatus* (Linnaeus, 1758) = 2495000 (DOUBTFUL) Note: this is an interesting case because this “doubtful” name has millions of occurrences tied to it. There is some apparently interesting taxonomic history behind this case...

What are the taxonomic status of each?

<https://api.gbif.org/v1/species/match?name=Calopteryx%20splendens>

# R and rgbif

John Waller | Data Analyst & rgbif maintainer



**GBIF**

Global Biodiversity  
Information Facility



**is a programming language.**

**It is commonly used for statistics and research.**

**There are thousands of R packages.**

```
# basic math (use # for comments)

x <- 2 # assign a variable

x + 2

x*x

(x - 10)/2

# some data types

v <- c(1,2,3,4)

l <- list(1,"cat",c(1,2,3))

d <- data.frame(pets = c("dog", "cat"), num = c(1,2))

pet <- "dog"

class(v) # use class to see type
```

```
# functions  
print("dog")  
class(1)  
getwd()  
?getwd # get help
```

```
# write your own function  
test_fun <- function(a,b) a + b  
test_fun(2,2)
```

```
# R packages are collections of functions  
  
install.packages("tidyverse")  
  
install.packages("rgbif")  
  
install.packages("CoordinateCleaner", dependencies = TRUE)  
  
# load packages  
  
library(tidyverse)  
  
library(rgbif)  
  
library(CoordinateCleaner)  
  
  
.libPaths() # where the packages were installed  
  
rgbif:: # type this in Rstudio to get list of functions
```

```
d <- data.frame(x=c(1,2,3))  
View(d) # view like in excel
```

```
library(dplyr) # for %>% and filter  
"dog" %>% print() # pipe  
print("dog") # same as above
```

```
# useful for filters  
d %>%  
  filter(x > 1) %>%  
  glimpse()
```

```
# read in an external table
library(readr)
table <- read_tsv("C:/Users/John/Desktop/some_file.tsv")

# basic data manipulation
library(dplyr)
d <- data.frame(x=c(1,2,3),y=c("cat", "dog", "dog"))
d$x # select single column
d %>% pull(x) # select single column

d %>%
group_by(y) %>%
count()
```





**rgbif** is a R package.

**rgbif** uses the [\*\*GBIF API\*\*](#) to access GBIF mediated data from within R.

It is useful for to **downloading** and **looking up species names** among other things.

```
library(rgbif)

name_backbone("Lepus saxatilis") # look up a taxonkey

occ_search(taxonKey=2436775) # preview some records

# preview a download request

occ_download_prep(pred("taxonKey") , 2436775)

# run an actual download

k <-occ_download(pred("taxonKey") , 2436775)

occ_download_wait(k) # wait for a download to finish

# download list of species

occ_download(pred_in( ("taxonKey") , c(2436775 , 10903982) )
```

# Exercise 4: Setting up R Environment

```
# only need to run once
install.packages("tidyverse")
install.packages("rgbif")
install.packages("CoordinateCleaner")
# run every time you restart Rstudio
library(tidyverse)
library(rgbif)
library(CoordinateCleaner)
```

Using **RStudio**, run the set up code above.

# Setting up R Environment (optional)

```
install.packages("usethis")
```

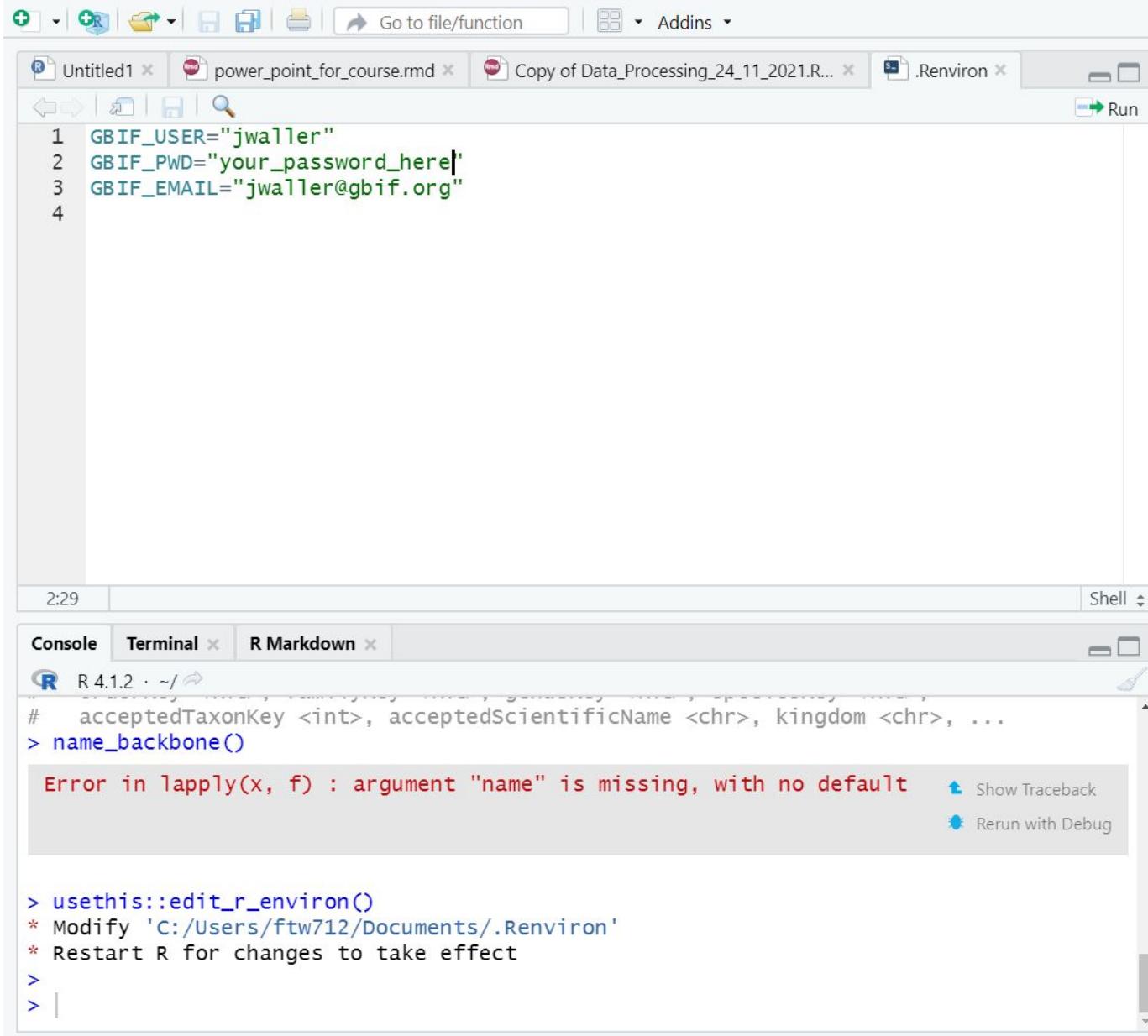
```
usethis::edit_r_environ()
```

*# Edit your .Renviron to look like this but with your information:*

```
GBIF_USER="jwaller"
```

```
GBIF_PWD="safe_fake_password123"
```

```
GBIF_EMAIL="jwaller@gbif.org"
```



The screenshot shows the RStudio interface. The top bar includes 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help' menus. The toolbar contains icons for file operations like 'New', 'Open', 'Save', and 'Print', as well as 'Go to file/function' and 'Addins'. The code editor tab bar shows 'Untitled1', 'power\_point\_for\_course.rmd', 'Copy of Data\_Processing\_24\_11\_2021.R...', and '.Renvironment'. The code editor window contains the following R code:

```
1 GBIF_USER="jwaller"
2 GBIF_PWD="your_password_here"
3 GBIF_EMAIL="jwaller@gbif.org"
4
```

The R Markdown tab is active in the bottom tab bar. The console tab shows the following output:

```
2:29 | Shell
Console Terminal R Markdown
R 4.1.2 · ~/r
#   acceptedTaxonKey <int>, acceptedScientificName <chr>, kingdom <chr>, ...
> name_backbone()
Error in lapply(x, f) : argument "name" is missing, with no default
  Show Traceback
  Rerun with Debug
```

Below the console, the R environment tab shows:

```
> usethis::edit_r_environment()
* Modify 'C:/Users/ftw712/Documents/.Renvironment'
* Restart R for changes to take effect
>
>
```

# Exercise 5: Downloading a dataset using **rgbif**

Download a dataset (simple csv) using **rgbif** that has the following properties:

1. Taxon is *Lepus saxatilis*
2. Found in South Africa (ZA)
3. That is a preserved specimen or human observation
4. That has latitude and longitude coordinates
5. Does not have common geospatial issues

In other words:

*Lepus saxatilis* + in ZA + specimen or human observation + has coordinates + no geoissues

---

```
name_backbone("Lepus saxatilis") # look up a taxonkey
```



# Exercise 5: Downloading a dataset using rgbif

```
library(rgbif)

user <- ""
pwd <- ""
email <- ""

occ download(
  pred("taxonKey", ?),
  pred(in("basisOfRecord",
  c('PRESERVED_SPECIMEN', 'HUMAN_OBSERVATION')),
  pred("country", "ZA"),
  pred("hasCoordinate", TRUE),
  pred("hasGeospatialIssue", FALSE),
  format = "SIMPLE_CSV",
  user=user, pwd=pwd, email=email
)
```

# Exercise 5: 1st Step - provide credentials to GBIF

```
user <- ""  
pwd <- ""  
email <- ""
```

```
install.packages("usethis")  
usethis::edit_r_environ()  
# Edit your .Renviron to look like this but with your information:  
GBIF_USER="jwaller"  
GBIF_PWD="safe_fake_password123"  
GBIF_EMAIL="jwaller@gbif.org"
```

# Exercise 5: 2nd Step - Use the occ\_download function

```
gbif_download_key <- occ_download(  
  pred("taxonKey", 2436775),  
  pred_in("basisOfRecord", c('PRESERVED_SPECIMEN', 'HUMAN_OBSERVATION'),  
  pred("country", "ZA"),  
  pred("hasCoordinate", TRUE),  
  pred("hasGeospatialIssue", FALSE),  
  format = "SIMPLE_CSV",  
  user=user, pwd=pwd, email=email  
)
```

# Exercise 5: 3rd Step - Importing your download into R

```
# import your download into R
data_download <- occ_download_get(gbif_download_key,
overwrite = TRUE) %>%
occ_download_import()
View(data_download)

# obtain a DOI for your dataset
res <- occ_download_meta(gbif_download_key)
gbif_citation(res)
```

# Exercise 6: Downloading a long species list

```
gbif_taxon_keys <- c(3189834, 3189801, 2876099, 2888580)

occ_download(
  pred_in("taxonKey", gbif_taxon_keys),
  pred("hasCoordinate", TRUE),
  pred("hasGeospatialIssue", FALSE),
  format = "SIMPLE_CSV",
  user=user, pwd=pwd, email=email
)
```

# Resources

PDF file  
español  
français

Search  

Table of Contents

[Biodiversity Data Use](#)

Description

Data Processing  
Key documentation  
Glossary

Acknowledgements  
Colophon

Contribute

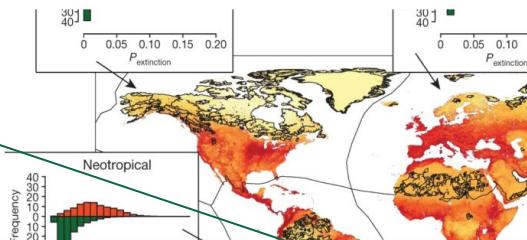
 [Improve this document](#)

## Biodiversity Data Use

GBIF Secretariat – [training@gbif.org](mailto:training@gbif.org) – Version 80af48f, 2021-12-14 12:40:04 UTC

This document is also available in [PDF format](#) and in other languages: [español](#), [français](#).

Di Marco M, Ferrier S, Harwood TD, Hoskins AJ and Watson JEM (2019) Wilderness areas halve the extinction risk of terrestrial biodiversity. *Nature*. Springer Science and Business Media LLC 573(7775): 582–585. Available at: <https://doi.org/10.1038/s41586-019-1567-7>



## Data Processing example

Lemur\_catta\_project.rmd

## Community Forum

[https://discourse.gbif.org/g/BID\\_DataUse](https://discourse.gbif.org/g/BID_DataUse)



### Other Formats

PDF file

español

français

### Search

### Table of Contents

[Biodiversity Data Use](#)

Description

Data Processing

Key documentation

API

Cloud Computing

Download Options

## Biodiversity Data Use

GBIF Secretariat – [training@gbif.org](mailto:training@gbif.org) – Version 80af48f, 2021-12-14 12:40:04 UTC

### R

- [Data Carpentry-Data Analysis and Visualization in R for Ecologists](#)
- [Datacamp- Range of courses in R, Python and SQL](#)
- [Introduction to R Manual](#)
- [rgbif Manual](#)
- [CoordinateCleaner Manual](#)
- [Downloading occurrences from a long list of species in R and Python](#)
- [Common things to look out for when post-processing GBIF downloads](#)
- [Finding gridded datasets & Gridded Datasets Update](#)

# Supplementary Exercise: Data cleaning with R

Clean your *Lepus saxatilis* download by

- Removing occurrence records where the **establishmentmeans** is indicated as **managed**, **introduced** or **invasive**
- Filtering **year** for records between 1955 and 2010
- Filtering for records with **coordinate uncertainty** of less than 10,000 and **coordinate precision** of greater than 0.01
- Removing points within 2 km of **country centroids** and **capital centroids**
- Removing points within 2 km of a **zoo** or **herbarium**

```
library(rgbif)

library(dplyr) # for filter and %>%

library(CoordinateCleaner) # for cc_cen,cc_cap,cc_inst

# gbif_download_key <- "0071981-210914110416597"

# first import data

gbif_download <- occ_download_get(gbif_download_key, overwrite = TRUE) %>%
  occ_download_import()

gbif_download %>%
  setNames(tolower(names(.))) %>% # set lowercase column names to work with CoordinateCleaner
  filter(!establishmentmeans %in% c("MANAGED", "INTRODUCED", "INVASIVE")) %>%
  filter(year >= 1955 & year <= 2010) %>%
  filter(coordinateprecision < 0.01 | is.na(coordinateprecision)) %>%
  filter(coordinateuncertaintyinmeters < 10000 | is.na(coordinateuncertaintyinmeters)) %>%
  cc_cen(buffer = 2000) %>% # remove country centroids within 2km
  cc_cap(buffer = 2000) %>% # remove capitals centroids within 2km
  cc_inst(buffer = 2000) %>% # remove zoo and herbaria within 2km
  glimpse() # look at results of pipeline
```

# Step 1 - Load libraries

This code will load the functions we need to clean the data.

**CoordinateCleaner** is an R package written specifically for **cleaning GBIF occurrence data**. <https://github.com/ropensci/CoordinateCleaner>

```
library(rgbif)
library(dplyr) # for filter and %>%
library(CoordinateCleaner) # for cc_cen, cc_cap, cc_inst
```

## Step 2 - Import Data

This code will import that data from GBIF into R.

Remember that the pipe (`%>%`) just passes the results of one function into another function.

```
# gbif_download_key <- "0071981-210914110416597"  
# first import data  
gbif_download <- occ_download_get(gbif_download_key, overwrite  
= TRUE) %>%  
  occ_download_import()
```

## Step 3 - Import data and clean column names

Set the column names to lowercase. The dot (“.”) here is special pipe code which refers back to the `gbif_download` object.

<https://magrittr.tidyverse.org/reference/pipe.html>

```
gbif_download %>%  
  setNames(tolower(names(.))) %>% # set lowercase column  
  names to work with CoordinateCleaner
```

## Step 4 - filter establishmentMeans

Here we use `filter` from the package `dplyr` to remove not naturally established records. “!“ means **negation** in R. The operator `%in%` checks if value in the column is in the vector `c ("MANAGED", "INTRODUCED", "INVASIVE")`.

```
filter(!establishmentmeans %in% c("MANAGED", "INTRODUCED",  
"INVASIVE")) %>%
```

## Step 5 - filter year

Here we use `filter` to keep only records between 1955 and 2010. The operator `>=` greater than or equal and means `<=` less than or equal. The `&` operator combines and checks that **both conditions** are TRUE.

```
filter(year >= 1955 & year <= 2010) %>%
```

# Step 6 - filter uncertain records

Here we use `filter` to keep only certain records. The `|` operator combines and checks that **only one condition** is TRUE. The `is.na` function checks if record is missing. **NA** means missing values in R.

```
filter(coordinateprecision < 0.01 | is.na(coordinateprecision))  
%>%  
filter(coordinateuncertaintyinmeters < 10000 |  
is.na(coordinateuncertaintyinmeters)) %>%
```

# Step 7 - filter with CoordinateCleaner

Here we use **CoordinateCleaner** to remove country centroids and records near zoos and botanical gardens.

```
cc_cen(buffer = 2000) %>% # remove country centroids
cc_cap(buffer = 2000) %>% # remove capitals centroids
cc_inst(buffer = 2000) %>% # remove zoo and herbaria
```

# RESOURCE LINKS

[GBIF occurrence search](#) (GBIF occurrence search)

<https://data-blog.gbif.org/post/downloading-long-species-lists-on-gbif/>  
(download)

<https://data-blog.gbif.org/post/gbif-filtering-guide/> (filtering guide)

<https://data-blog.gbif.org/categories/gbif/> (data blog)

<https://data-blog.gbif.org/post/outlier-detection-using-dbscan/> (outlier detection)

<https://data-blog.gbif.org/post/gbif-molecular-data-quality/> (metagenomics)