

## Uso de datos de biodiversidad 1er módulo: el procesamiento de datos

Andrew Rodrigues | Oficial de Programas



Global Biodiversity  
Information Facility

$P_{\text{extinction}}$

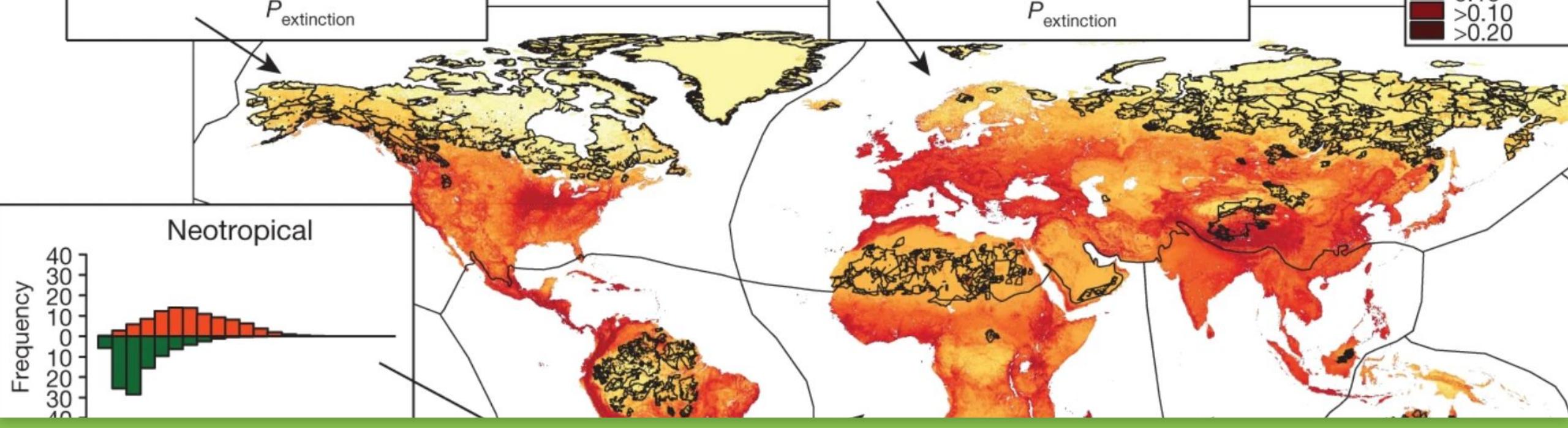
Afrotropical

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



## ESQUEMA DEL CURSO

Parte 1: Navegación por [www.gbif.org](http://www.gbif.org)

Parte 2: Problemas habituales de calidad de los datos

Parte 3- La API

Parte 4 - Uso de R

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

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

# Entrenadores y mentores

## Entrenadores



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



John Waller - GBIF Secretariat, Data Analyst

## Mentores



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

# Acceso libre y gratuito a los datos de biodiversidad

REGISTROS

ESPECIES

CONJUNTOS DE DATOS

PUBLICADORES

RECURSOS

Buscar



¿QUÉ ES GBIF?

SOBRE GBIF ARGENTINA

Registros biológicos  
1.904.771.045

Juegos de datos  
64.100

Instituciones que publican  
1.769

Artículos científicos usando datos  
6.528

## ¿POR QUÉ PROCESAMOS LOS DATOS?

Queremos que nuestra descarga se ajuste a nuestros propios fines.

1. Eliminar datos erróneos, p. ej. valores atípicos
2. Asegurar un nivel suficiente de precisión en los datos para nuestro propósito.

# UNA VENTANA SOBRE LA EVIDENCIA SOBRE DÓNDE Y CUÁNDO HAN VIVIDO LAS ESPECIES

- Observaciones



- Especímenes digitalizados

- Literatura



- Publicación de datos e indexación



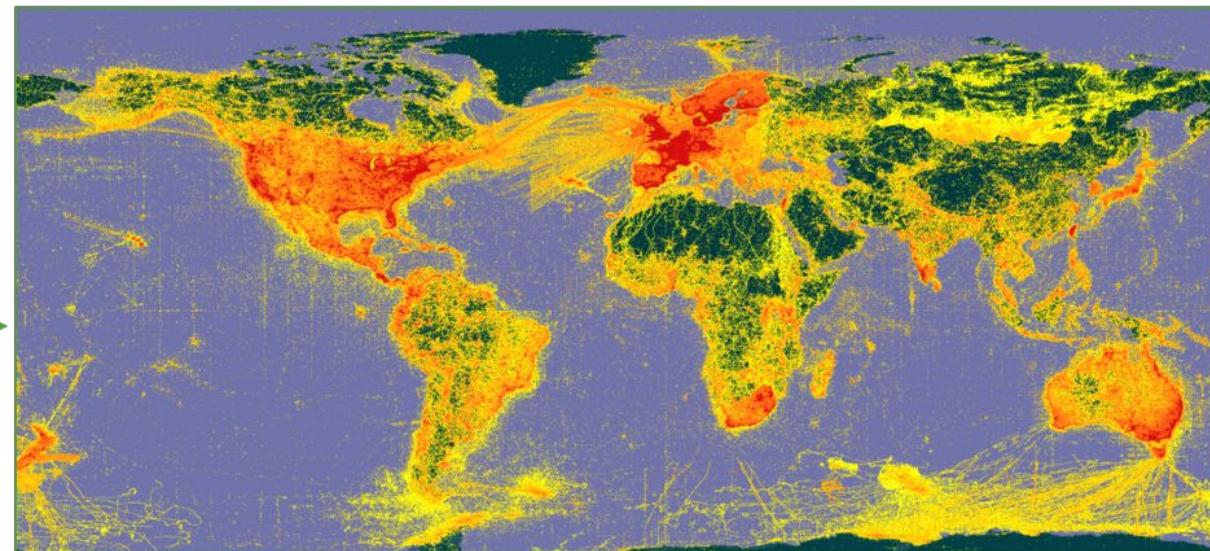
- Estándares comunes (DwC)



- Detección remota



- ADN Ambiental



- Descubrimiento de datos y uso

# Acceso libre y gratuito a los datos de biodiversidad

REGISTROS

ESPECIES

CONJUNTOS DE DATOS

PUBLICADORES

RECURSOS

Buscar



¿QUÉ ES GBIF?

SOBRE GBIF ARGENTINA

Registros biológicos  
1.904.771.045

Juegos de datos  
64.100

Instituciones que publican  
1.769

Artículos científicos usando datos  
6.528

*Cedrela odorata* L. observed in Haiti by Brian Oakes Haiti Hunter (CC BY 4.0)

## ¿POR QUÉ PROCESAMOS LOS DATOS?

Cada vez que procese un conjunto de datos para su uso, tendrá que considerar

1. Requisitos de su análisis
2. Equilibrio entre la calidad de los datos y la solidez de su análisis

Este puede ser un proceso iterativo

# Acceso libre y gratuito a los datos de biodiversidad

REGISTROS

ESPECIES

CONJUNTOS DE DATOS

PUBLICADORES

RECURSOS

Buscar



¿QUÉ ES GBIF?

SOBRE GBIF ARGENTINA

Registros biológicos  
1.904.771.045

Juegos de datos  
64.100

Instituciones que publican  
1.769

Artículos científicos usando datos  
6.528

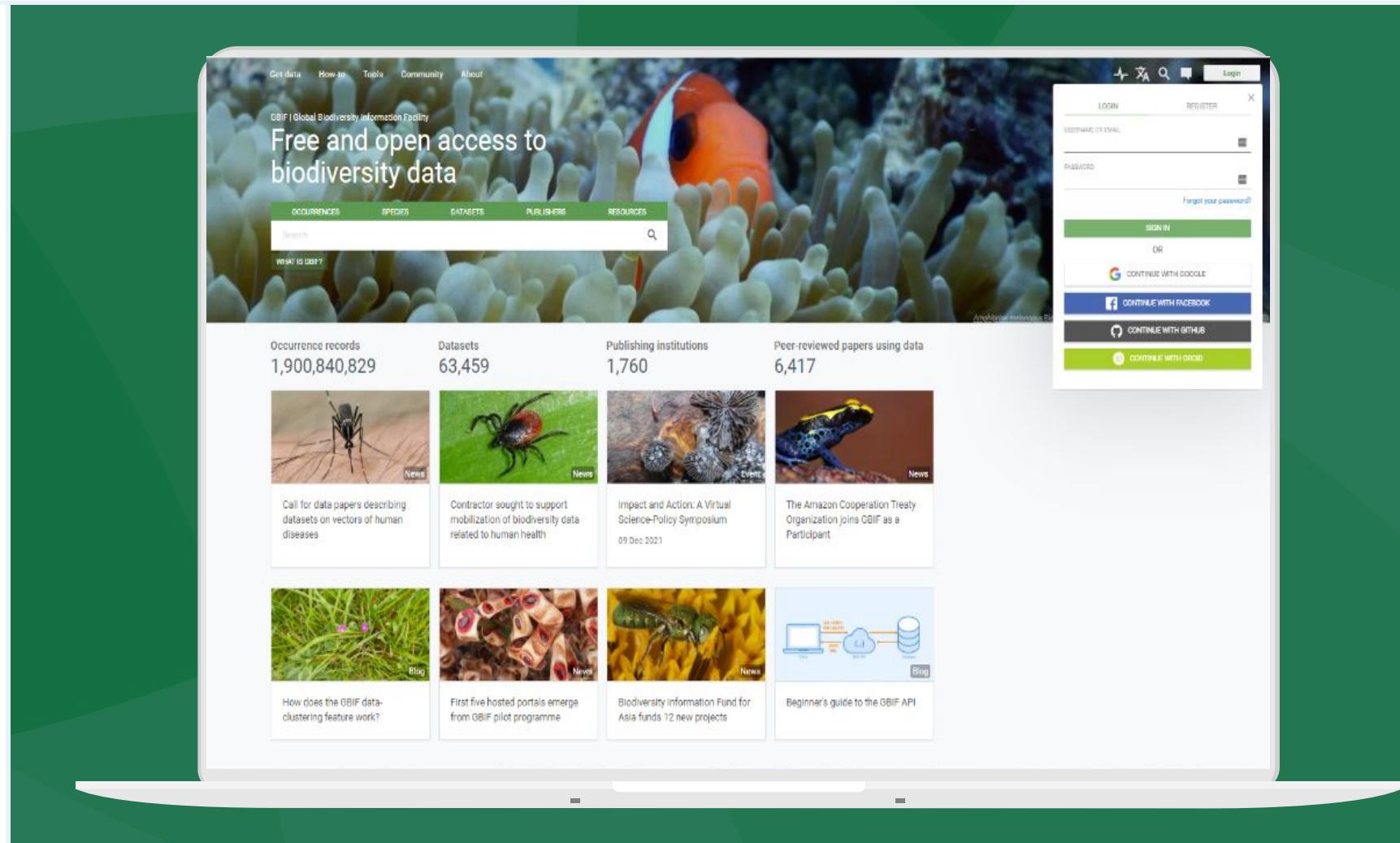
*Cedrela odorata* L. observed in Haiti by Brian Oakes Haiti Hunter (CC BY 4.0)

## REGLAS DE ORO DEL USO DE DATOS MEDIADOS POR GBIF

1. Debe tener una cuenta en [www.gbif.org](http://www.gbif.org)
2. Debe aceptar el Acuerdo de usuario de datos: <https://www.gbif.org/terms/data-user>
3. Documente cómo procesa sus datos
4. Cite correctamente los datos que usa
5. Depositar datos usados en un repositorio público

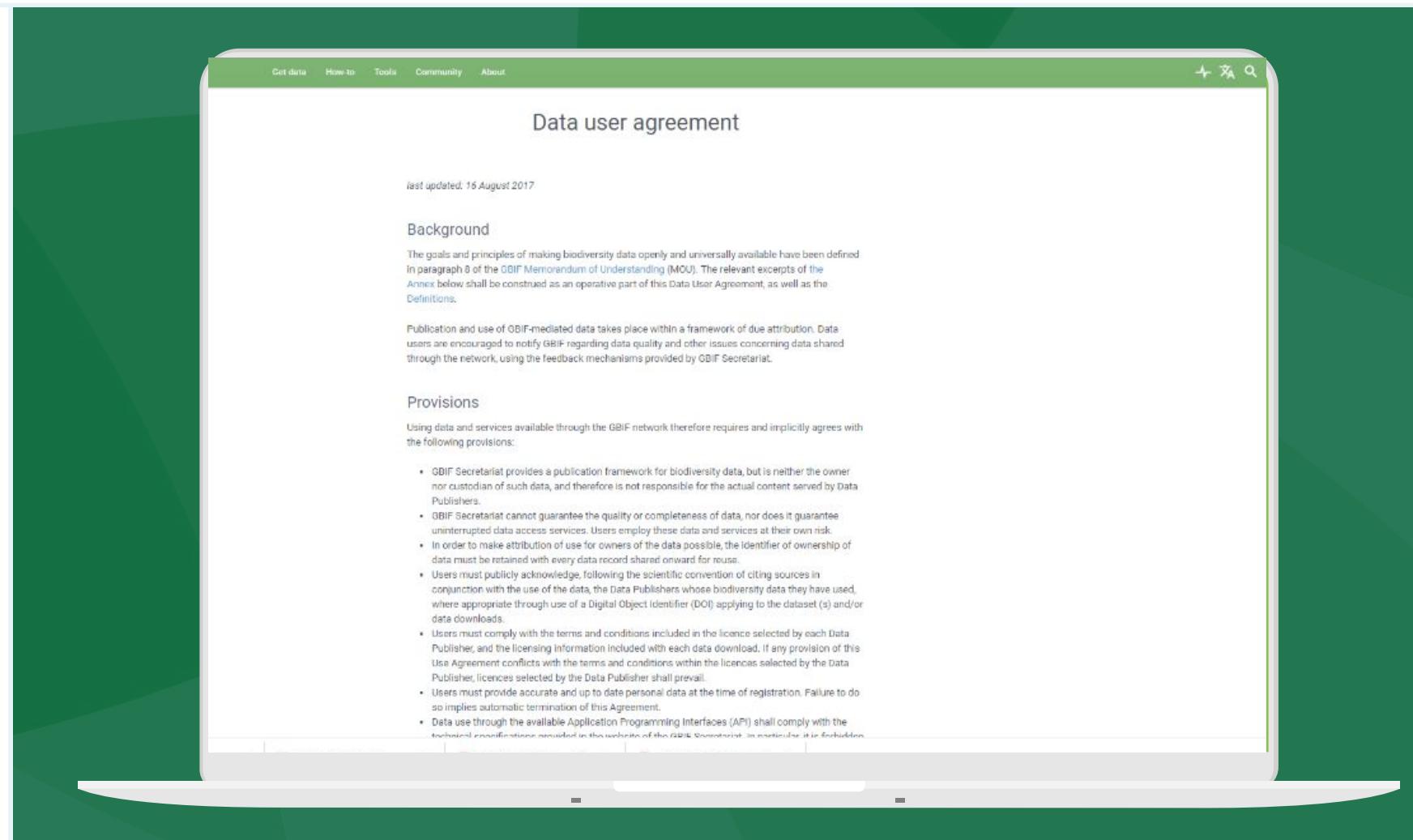
# REGLAS DE ORO DEL USO DE DATOS MEDIADOS POR GBIF

- Debe tener una cuenta en [www.gbif.org](http://www.gbif.org)



# REGLAS DE ORO DEL USO DE DATOS MEDIADOS POR GBIF

- Debe tener una cuenta en [www.gbif.org](http://www.gbif.org)
- Debe aceptar el Acuerdo de usuario de datos: <https://www.gbif.org/terms/data-user>
  - No vinculante
  - Establece los principios rectores del uso de datos, incluida la cita de datos.



# REGLAS DE ORO DEL USO DE DATOS MEDIADOS POR GBIF

- Debe tener una cuenta en [www.gbif.org](http://www.gbif.org)
- Debe aceptar el Acuerdo de usuario de datos:  
<https://www.gbif.org/terms/data-user>
- Documente cómo procesa sus datos

## Paso 1

Descargar registros de ocurrencia de la especie x con un DOI asociado

## Paso 2

Eliminar todos los registros fuera de su rango nativo

## Paso 3

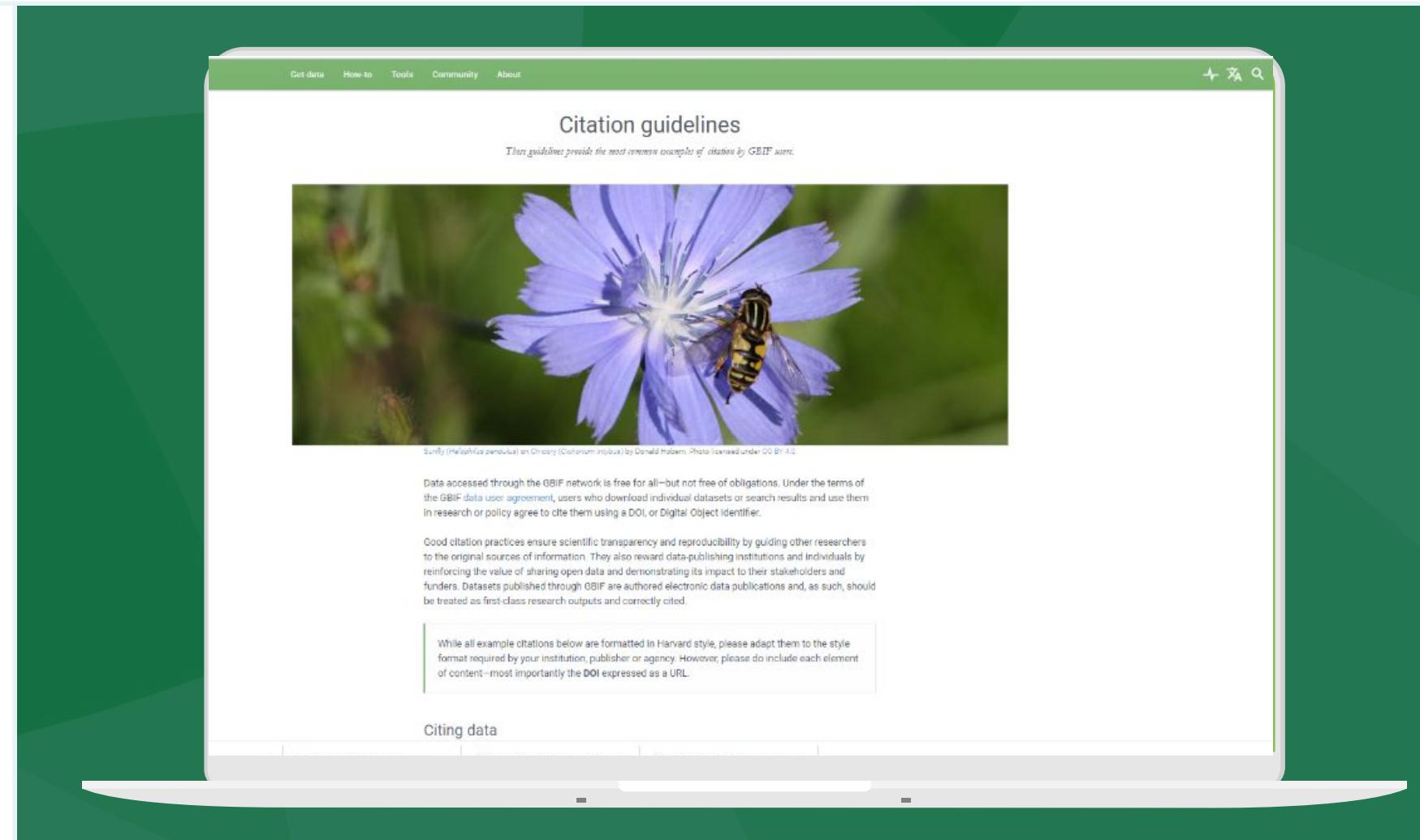
Eliminar todos los registros recopilados antes de 1950

## Paso 4

Conjunto de datos final limpio

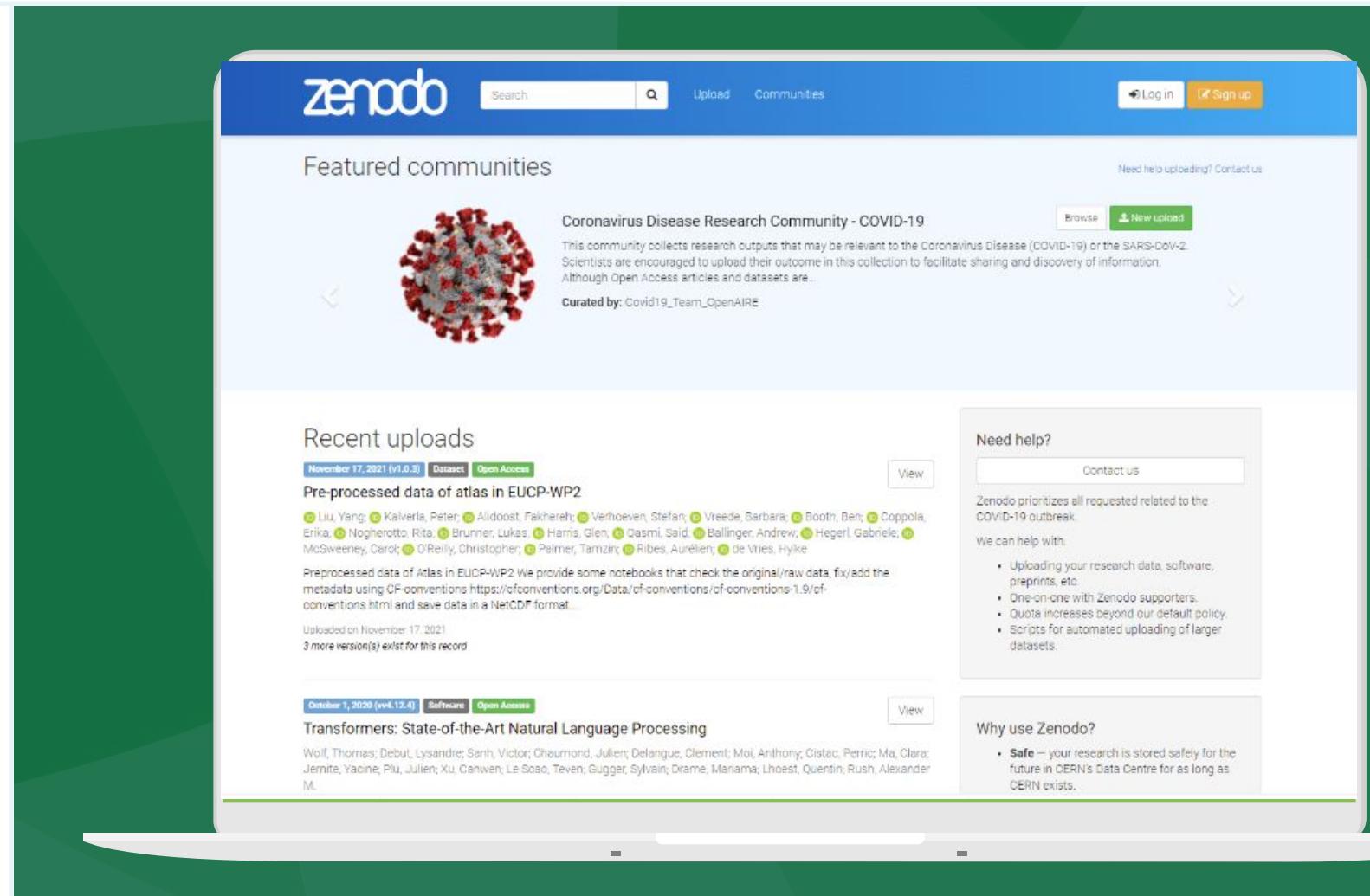
# REGLAS DE ORO DEL USO DE DATOS MEDIADOS POR GBIF

- Debe tener una cuenta en [www.gbif.org](https://www.gbif.org)
- Debe aceptar el Acuerdo de usuario de datos:  
<https://www.gbif.org/terms/data-user>
- Documente cómo procesa sus datos
- Cite correctamente los datos que usa
  - Directrices:  
<https://www.gbif.org/citation-guidelines>
  - DOI de conjuntos de datos derivados:  
<https://www.gbif.org/derived-dataset/about>



# REGLAS DE ORO DEL USO DE DATOS MEDIADOS POR GBIF

- Debe tener una cuenta en [www.gbif.org](https://www.gbif.org)
- Debe aceptar el Acuerdo de usuario de datos: <https://www.gbif.org/terms/data-user>
- Documente cómo procesa sus datos
- Cite correctamente los datos que usa
  - Directrices: <https://www.gbif.org/citation-guidelines>
  - DOI de conjuntos de datos derivado
- Deposite los datos usados en un repositorio público, ej. Zenodo



	Datos sin procesar	Interpreted data	Multimedia	Coordinadas	Formato	Estimated data size
<a href="#"> CSV</a>	X	✓	X	✓ (if available)	Tab-delimited CSV <a href="#">?</a>	<b>470 GB</b> (comprimido para descargar)
<a href="#"> ARCHIVO DARWIN CORE</a>	✓	✓	✓ (links)	✓ (if available)	Tab-delimited CSV <a href="#">?</a>	<b>1 TB</b> (comprimido para descargar)
<a href="#"> LISTAS DE ESPECIES</a>	X	✓	X	X	Tab-delimited CSV <a href="#">?</a>	

## DESCARGAS DE DATOS

Los datos se pueden descargar en tres formatos

**Simple: CSV delimitado por tabuladores.** Solo contiene los datos después de la interpretación de GBIF. No incluye multimedia. [Más información sobre CSV](#)

**Archivo Darwin Core:** El Archivo Darwin Core (DwC-A) contiene tanto los datos originales como los proporcionó el editor y la interpretación de GBIF. Vínculos (pero no archivos) a multimedia incluidos. [Más información sobre DwC-A](#)

**Lista de especies:** CSV delimitado por tabulaciones con la lista distinta de nombres en el resultado de la búsqueda.

# Acceso libre y gratuito a los datos de biodiversidad

REGISTROS

ESPECIES

CONJUNTOS DE DATOS

PUBLICADORES

RECURSOS

Buscar



¿QUÉ ES GBIF?

SOBRE GBIF ARGENTINA

Registros biológicos  
1.904.771.045

Juegos de datos  
64.100

Instituciones que publican  
1.769

Artículos científicos usando datos  
6.528

*Cedrela odorata* L. observed in Haiti by Brian Oakes Haiti Hunter (CC BY 4.0)

## RESUMEN DEL SITIO WEB

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

# Ejercicio 1: Navegando [www.gbif.org](http://www.gbif.org)

¿Cuál es el número total de ocurrencias de las islas Tongatapu en Tonga?

¿Cuántos registros del reino Plantae hay en las islas?

¿Cuántos de estos registros son del programa BID?

¿Cuántos de estos registros están bajo una licencia CC-BY?

¿Cuántos tienen imágenes?

# Ejercicio 1: Navegando [www.gbif.org](http://www.gbif.org)

¿Cuál es el número total de ocurrencias de las islas Tongatapu en Tonga?

(3,372 - 7 Diciembre 2021 sólo GADM, 3,362 - 7 Diciembre 2021 con GADM y filtro de país )

¿Cuántos registros del reino Plantae hay en las islas? (459 - 7 Diciembre 2021)

¿Cuántos de estos registros son del programa BID? (36 - 7 Diciembre 2021)

¿Cuántos de estos registros están bajo una licencia CC-BY? (0 - 7 Diciembre 2021)

¿Cuántos tienen imágenes? (0 - 7 Diciembre 2021)

# Problemas habituales de Calidad de Datos

John Waller | Analista de Datos



Tu **descarga de GBIF** no siempre será "perfecta" para lo que quieras hacer con ella. **Hay algunas cosas que debe tener en cuenta ...**



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

**Botón predeterminado de problemas geoespaciales**

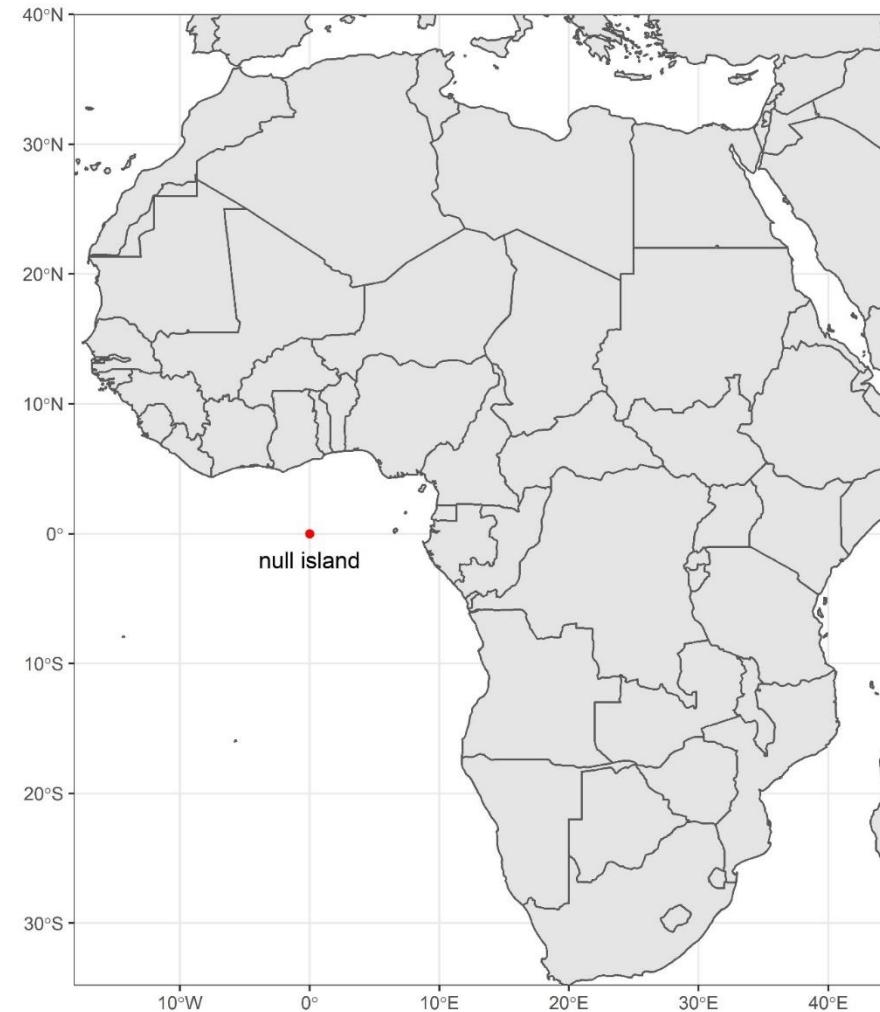
Scientific name	Country or area	Coordinates	Month & year	Basis of record	Dataset
<i>Asteraceae</i>	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De...</a>
<i>Tibouchina Aubl.</i>	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De...</a>
<i>Calibrachoa Cerv.</i>	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De...</a>
<i>Polygala moquiniana A.St.-Hil.</i>	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De...</a>
<i>Cyperaceae</i>	Brazil	24.4S, 49.8W	2021 January	Preserved specimen	<a href="#">FLOR - Herbário do De...</a>
<i>Hyptis Jacq.</i>	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>

# Problemas geoespaciales predeterminados

GBIF elimina los problemas geoespaciales comunes de forma predeterminada si elige tener datos con una ubicación.

- **Coordenada cero:** las coordenadas son exactamente (0,0). isla nula
- **Discrepancia de coordenadas de país:** las coordenadas quedan fuera del polígono de un país determinado.
- **Coordenadas no válidas:** GBIF no puede interpretar las coordenadas.
- **Coordenadas fuera de rango:** las coordenadas están fuera del rango de los valores decimales lat / lon ((-90,90), (-180,180)).

# GBIF elimina las coordenadas cero (0,0) "isla nula"

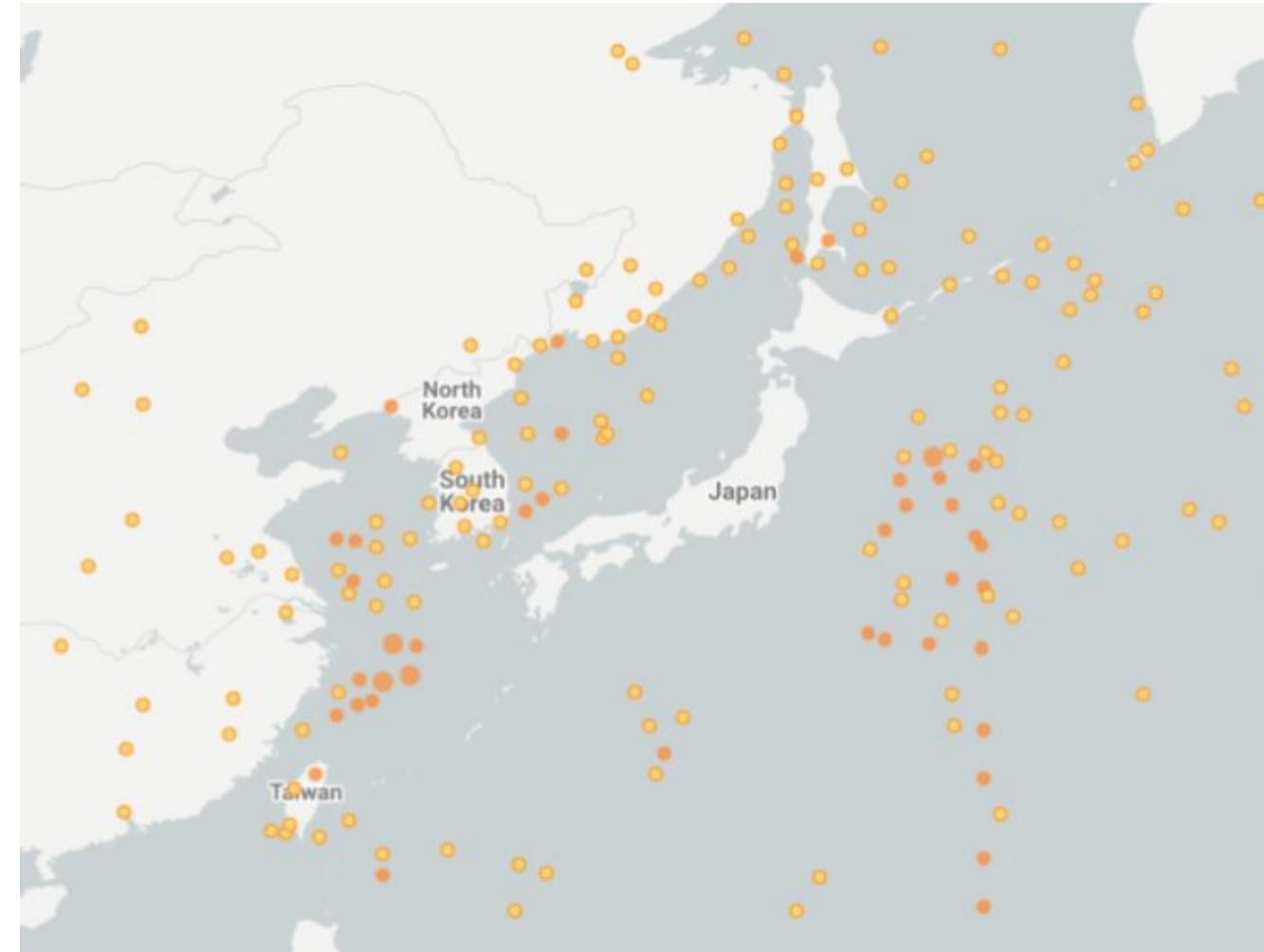


[https://www.gbif.org/occurrence/map?issue=ZERO\\_COORDINATE](https://www.gbif.org/occurrence/map?issue=ZERO_COORDINATE)

# GBIF elimina la falta de coincidencia de las coordenadas del país

GBIF elimina los registros que **no coinciden con su código de país**.

Todos **estos registros** afirman estar ubicados en Japón.



[https://www.gbif.org/occurrence/search?issue=COUNTRY\\_COORDINATE\\_MISMATCH](https://www.gbif.org/occurrence/search?issue=COUNTRY_COORDINATE_MISMATCH)

# GBIF elimina los registros de ausencias

A veces, los publicadores de datos incluirán **registros de ausencia** (donde verifican que una especie no está presente). La mayoría de los usuarios no quieren estos registros.

```
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

Search all fields

TABLE GALLERY MAP TAXONOMY METRICS DOWNLOAD

Simple Advanced

Occurrence status

Present

Este botón se encuentra activo por defecto

	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

Otras cuestiones que **tienes que filtrar tú mismo...**

# Fósiles y Especímenes Vivos

GBIF tiene **fósiles** y **especímenes vivos** (generalmente una planta dentro de un jardín botánico o, a veces, un animal en un zoológico).

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

# establishmentMeans

**dwc:establishmentMeans** : El proceso por el cual los individuos biológicos representados en el registro se establecieron en el lugar.

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

Desafortunadamente, no se usa con mucha frecuencia.

# Registros antiguos

GBIF tiene muchos registros de museos que **pueden ser más antiguos de lo que se desea** para algunos estudios.

```
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)

# Ejemplo de ubicación incierta

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

**Location:** United States of America

**Basis of record:** Human observation

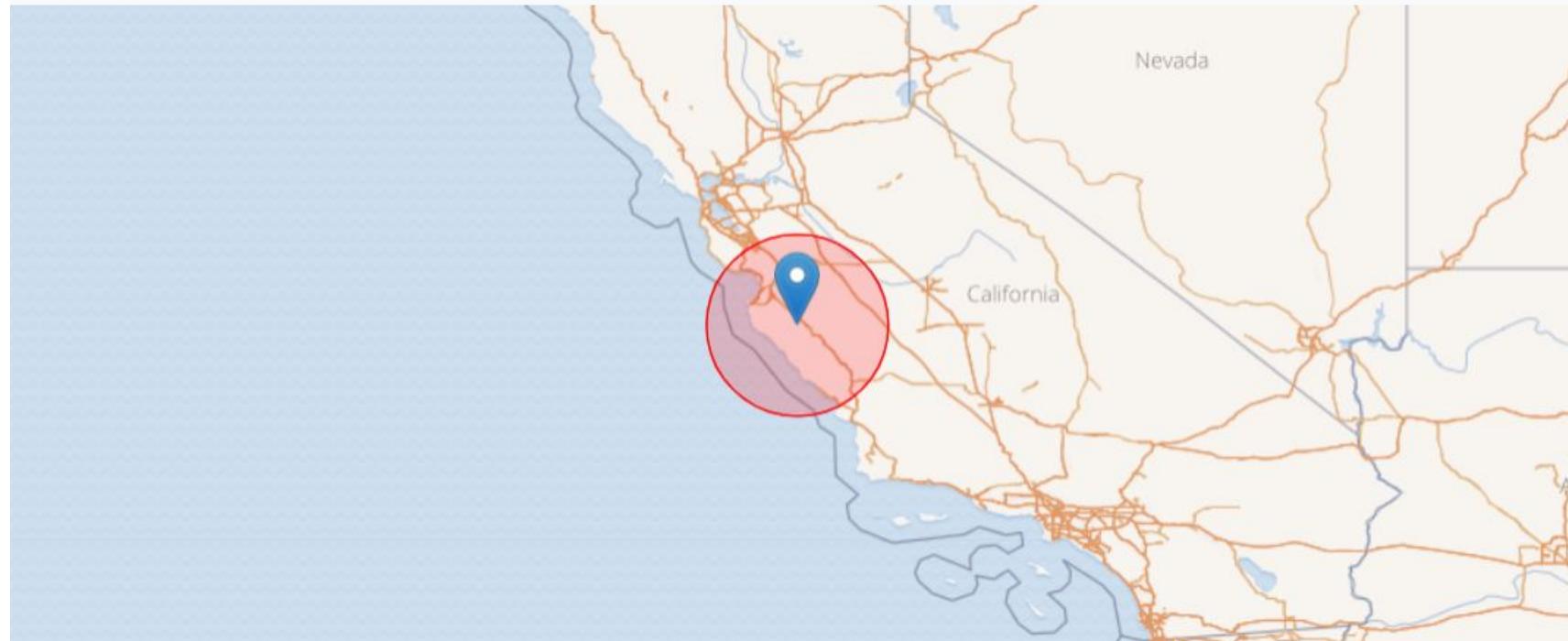


**Dataset:** iNaturalist Research-grade Observations

**Publisher:** iNaturalist.org

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

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



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

# Ubicación incierta

A menudo, querrá asegurarse de que las coordenadas den una ubicación determinada y no estén realmente a miles de kilómetros de donde se observó o se recogió el organismo.

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

**Recomiendo no filtrar los valores perdidos**, ya que los editores a menudo no completan el valor si creen que la ocurrencia es bastante segura (de un GPS).

# Valores predeterminados incorrectos para la incertidumbre de las coordenadas

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

Hay algunos valores "falsos" para la incertidumbre de coordenadas que debe conocer. Estos valores son errores producidos por el software de codificación geográfica y no representan valores de incertidumbre reales. En el caso de **301**, la incertidumbre es a menudo mucho mayor que 301 y en realidad representa **el centroide de un país**.

# Puntos a lo largo del ecuador o el primer meridiano

Algunos editores consideran que cero y NULL son equivalentes, la latitud y la longitud vacías terminan trazándose a lo largo de estas dos líneas.

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

+

-

□

Atlantic Ocean

Mal

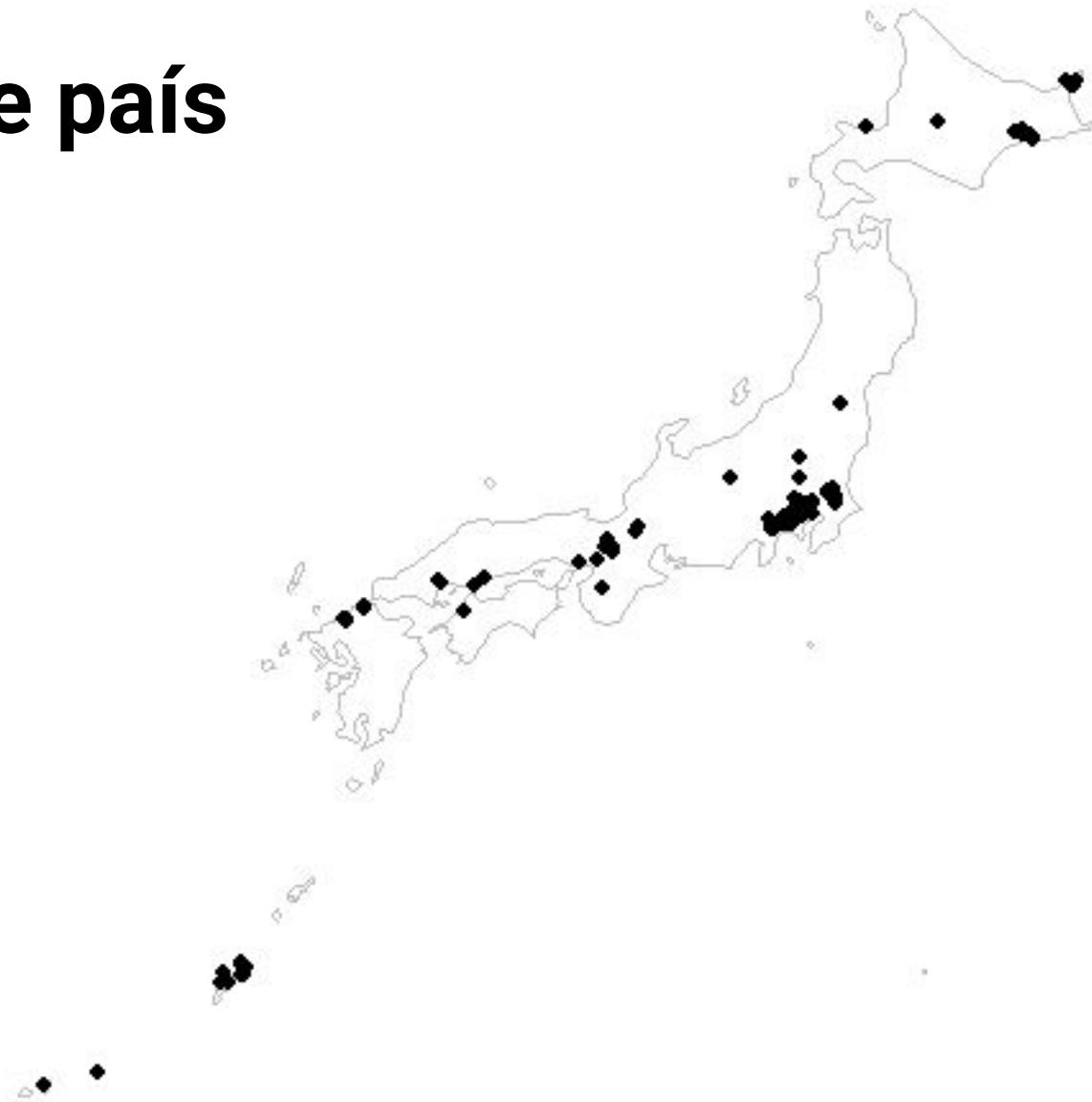
GRNHE

WCO

LAPE

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

# Centroides de país



# Georeferenciación retrospectiva

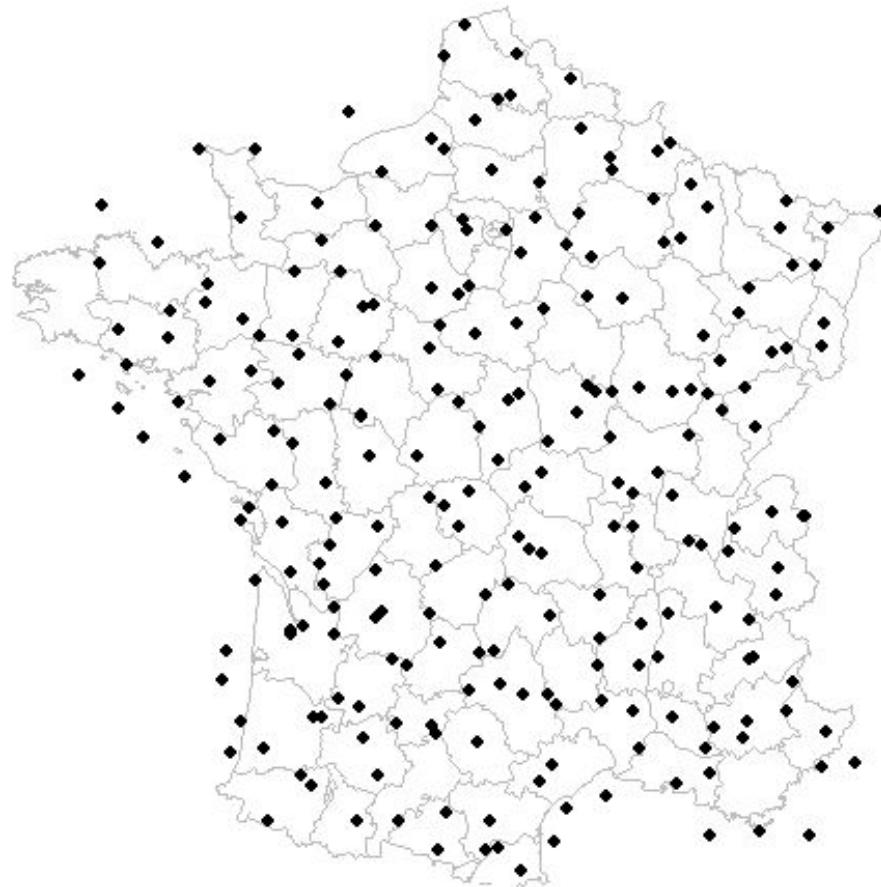
La **georreferenciación retrospectiva** es el proceso en el que se dan valores lat-lon a registros más antiguos que solo tienen información de localidad.

La información de la localidad a veces es solo un país, una ciudad o una descripción de texto como

*"10 millas al SO de la carretera principal, Austin TX".*

A menudo, los registros de los museos (especímenes conservados) se habrán georreferenciado retrospectivamente.

# Conjuntos de datos en cuadrícula



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

# Filtrado de conjuntos de datos en cuadrícula

La mayoría de los editores de conjuntos de datos cuadriculados en realidad completan una de las siguientes columnas:

- coordinateuncertaintyinmeters
- coordinateprecision
- footprintwkt (solo en descargas DwCA-A)

Por lo tanto, filtrar por estas columnas puede ser una **buenas forma de eliminar conjuntos de datos en cuadrícula**.

GBIF tiene una [API experimental](#) para identificar conjuntos de datos que exhiben un cierto grado de "cuadrícula". Puedes leer más [aquí](#).

# Ejercicio 2: filtrado de datos para mejorar la calidad de los datos

Usando [www.gbif.org](http://www.gbif.org) filtrar los datos de *Calopteryx splendens* de la siguiente manera:

- *Filtrar registros con incertidumbre de coordenadas entre 0 y 10,000 m*
- *Filtrar por registros entre 1955 y 2017*
- *Excluir registros de sucesos donde la media del establecimiento se indique como administrada, introducida o invasiva.*

¿Cuántos registros tenías al principio?

¿Cuántos tienes después de filtrar?

¿Cómo encontraste el taxón?

¿Cuáles son las limitaciones de los filtros?

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

## La API de GBIF

Andrew Rodrigues | Oficial de Programas



# 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

# ¿Qué es la API de GBIF?

La **interfaz de programación de aplicaciones (API)** de GBIF brinda a los usuarios acceso a las bases de datos de GBIF de manera segura.

Por lo general, la razón principal por la que querría usar una API es porque desea **software**, para interactuar con GBIF de alguna manera.

Se puede acceder a la API de GBIF a través de:

- Un navegador web visitando una URL, por ej. [https://api.gbif.org/v1/species/match?name=Passer domesticus](https://api.gbif.org/v1/species/match?name=Passer%20domesticus)
- O usando un programa de línea de comando llamado `curl` que necesita instalar

GBIF tiene algunos grupos de API / espacio de nombres:

- **API de registro:** hace que todos los conjuntos de datos, instalaciones, organizaciones, nodos y redes registrados sean detectables.
- **API de especies:** funciona con los datos guardados en el Banco de listas de verificación de GBIF, que indexa taxonómicamente todos los conjuntos de datos de listas de verificación registrados en la red de GBIF.
- **API de incidencias:** funciona contra el almacén de incidencias de GBIF, que gestiona los registros de incidencias y los pone a disposición a través del servicio web y los archivos de descarga.
- **API de mapas:** servicio de mosaicos de mapas web que simplifica la visualización del contenido GBIF en mapas interactivos y la superposición de contenido de otras fuentes.
- **API de literatura:** busque literatura indexada por GBIF, incluidos artículos revisados por pares, que citan conjuntos de datos de GBIF y descargas.

# ¿Qué es la API de GBIF?

El patrón básico de una llamada a la API:

- **URL base (base URL):** siempre será <https://api.gbif.org/v1/>
- **api:** este es el grupo de API GBIF / espacio de nombres que desea consultar.
- **función (function):** la funcionalidad que desea utilizar.
- **parámetro (parameter):** los parámetros para su llamada a la API. Un ? se utiliza, a veces.
- **consulta (query):** la consulta que usted completa. A veces será texto libre y, a veces, será un argumento predefinido.

Ejemplo

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

# Ejercicio 3 - Encontrar taxonkeys de GBIF

Las claves de taxón se generan a nivel de especie, familia de género, orden, filo y reino.

- Los identificadores únicos se emiten para los nombres aceptados con sinónimos de esos nombres aceptados emitidos con el mismo identificador. Claves de uso frente a acceptedusagekeys
- Permite que el usuario se asegure de que está recopilando todos los datos que necesita
- También facilita la descarga de múltiples especies
- Taxonkeys se puede encontrar a través de:
  - [www.gbif.org](http://www.gbif.org)
  - API de especies
  - Herramienta de comparación de especies: <https://www.gbif.org/tools/species-lookup>

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

# Ejercicio 3 - Encontrar taxonkeys de GBIF

Usando la **API de Especies** encuentre el **taxonkeys** de GBIF para estos nombres científicos:

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

¿Cuál es el estatus taxonómico de cada nombre?

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

# Ejercicio 3 - Encontrar taxonkeys de GBIF

Usando la **API de Especies** encuentre el **taxonkeys** de GBIF para estos nombres científicos:

- **Lepus saxatilis** F.Cuvier, 1823 = 2436775 (ACEPTADO)
- **Aves** = 212 (ACEPTADO) *Nota: esto no es una especie, por lo que si desea ocurrencias para un grupo completo, solo necesita un taxonkey y no una lista de todas las especies del grupo.*
- **Magnoliophyta** = 49 (SINÓNIMO) *Nota: que si usó la API directamente, GBIF le dará la clave de uso aceptada: 7707728, que es Tracheophyta. Este es el nombre aceptado de este grupo. Si decidiera usar el nombre antiguo, se perderían millones de apariciones, así que tenga cuidado.*
- **Aegithalos caudatus** (Linnaeus, 1758) = 2495000 (DUDOSO) *Nota: este es un caso interesante porque este nombre "dudoso" tiene millones de registros vinculados a él. Hay una historia taxonómica aparentemente interesante detrás de este caso ...*

¿Cuál es el estatus taxonómico de cada nombre?

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

**R y rgbif**

**John Waller | Analista de Datos**



**GBIF**

Global Biodiversity  
Information Facility



es un lenguaje de programación.

Comúnmente usado en **estadística e investigación**.

Existen **miles** de paquetes de R.

```
# matemática básica (use # para comentarios)

x <- 2 # asignar una variable

x + 2

x*x

(x - 10)/2

# algunos tipos de datos

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' para ver el tipo de dato
```

```
# funciones  
print("dog")  
class(1)  
getwd()  
?getwd # obtener ayuda
```

```
# escribir una función propia  
test_fun <- function(a,b) a + b  
test_fun(2,2)
```

```
# los paquetes de R son colecciones de funciones  
install.packages("tidyverse")  
install.packages("rgbif")  
install.packages("CoordinateCleaner", dependencies = TRUE)  
# cargar paquetes  
library(tidyverse)  
library(rgbif)  
library(CoordinateCleaner)  
  
.libPaths() # dónde fueron instalados los paquetes  
rgbif:: # escribir esto en Rstudio para obtener las funciones
```

```
d <- data.frame(x=c(1,2,3))  
View(d) # vista como en Excel
```

```
library(dplyr) # para %>% y filtro  
"dog" %>% print() # pipe  
print("dog") # igual que arriba
```

```
# útil para filtros  
d %>%  
  filter(x > 1) %>%  
  glimpse()
```

```
# leer en una tabla externa
library(readr)
table <- read_tsv("C:/Users/John/Desktop/some_file.tsv")

# manipulación básica de datos
library(dplyr)
d <- data.frame(x=c(1,2,3),y=c("cat", "dog", "dog") )
d$x # seleccionar una sola columna
d %>% pull(x) # seleccionar una sola columna

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





**rgbif es un paquete de R.**

**rgbif usa la API de GBIF para acceder a los datos mediados por GBIF dentro de R.**

Es útil para **descargar** y **buscar nombres de especies**, entre otras cosas.

```
library(rgbif)

name_backbone("Lepus saxatilis") # buscar un taxonKey

occ_search(taxonKey=2436775) # vista previa de algunos
registros

# vista previa de una solicitud de descarga
occ_download_prep(pred("taxonKey") , 2436775)

# ejecutar una descarga
k <-occ_download(pred("taxonKey") , 2436775)
occ_download_wait(k) # esperar para que la descarga finalice

# descarga de listas de especies
occ_download(pred_in(("taxonKey") , c(2436775,10903982))
```

# Ejercicio 4: Configuración del Entorno R

```
# solo necesita ser ejecutado una vez
install.packages("tidyverse")
install.packages("rgbif")
install.packages("CoordinateCleaner")
# ejecutar cada vez que se reinicia RStudio
library(tidyverse)
library(rgbif)
library(CoordinateCleaner)
```

Usando **RStudio**, ejecutar el código de configuración anterior.

# Configuración del entorno R (opcional)

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

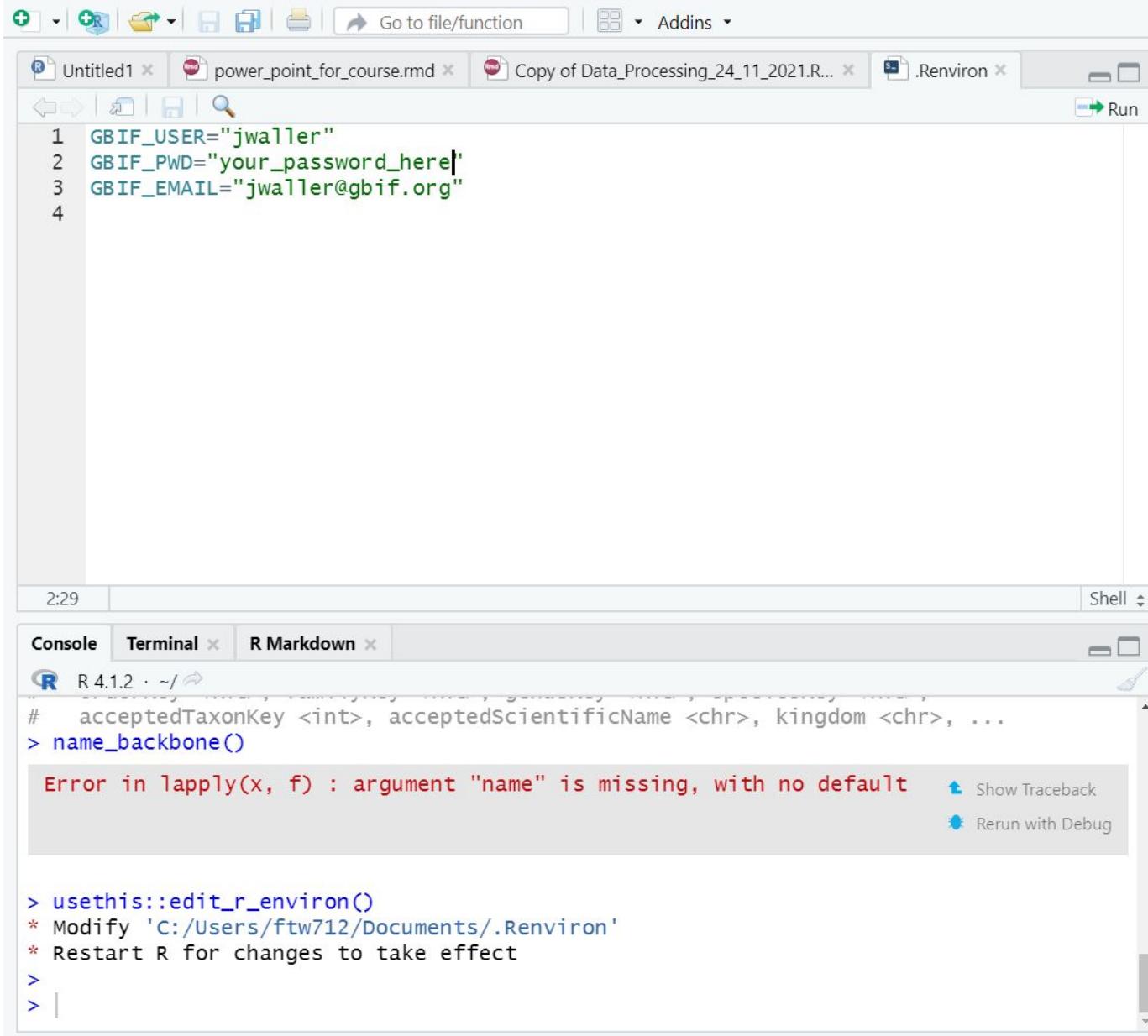
```
usethis::edit_r_environ()
```

*# Edite su entorno de R para para que se vea de esta forma, pero con su información personal:*

```
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 console tab bar shows 'Console', 'Terminal', and 'R Markdown'. The console window shows the following R session:

```
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 error message, the session continues:

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

# Ejercicio 5: Descargar un conjunto de datos usando rgbif

Descargar un conjunto de datos (simple csv) usando `rgbif`, que contenga las siguientes propiedades:

1. El taxon es *Lepus saxatilis*
2. Registros tomados en Sudáfrica (ZA)
3. Corresponden a especímenes preservados u observaciones humanas
4. Contienen coordenadas de latitud y longitud
5. No tienen problemas geoespaciales comunes

En otras palabras:

*Lepus saxatilis* + en ZA + especímenes preservados u observaciones humanas + coordenadas presentes + sin errores espaciales

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



# Ejercicio 5: Descargar un conjunto de datos usando 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
)
```

# Ejercicio 5: 1<sup>er</sup> Paso - proporcionar credenciales a GBIF

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

```
install.packages("usethis")  
usethis::edit_r_environ()  
# Edite su entorno de R para para que se vea de esta forma, pero con su información personal  
GBIF_USER="jwaller"  
GBIF_PWD="safe_fake_password123"  
GBIF_EMAIL="jwaller@gbif.org"
```

# Ejercicio 5: 2<sup>do</sup> Paso - Use la función occ\_download

```
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  
)
```

# Ejercicio 5: 3<sup>er</sup> Paso - Importar su descarga en R

```
# importar su descarga en R  
  
data_download <- occ_download_get(gbif_download_key,  
overwrite = TRUE) %>%  
  
occ_download_import()  
  
View(data_download)
```

```
# obtener un DOI para el conjunto de datos  
  
res <- occ_download_meta(gbif_download_key)  
  
gbif_citation(res)
```

# Ejercicio 6: Descargar una larga lista de especies

```
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
)
```

# Recursos

PDF file  
español  
français

Search  

Table of Contents

[Biodiversity Data Use](#)

Description

Data Processing  
Key documentation  
Glossary

Acknowledgements  
Colophon

Contribute

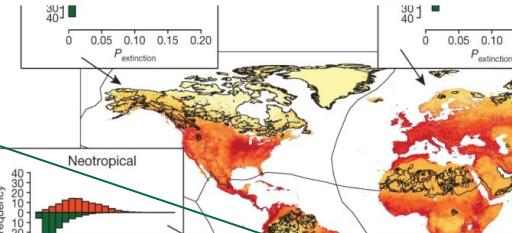
 [Improve this document](#)

## 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>



**Ejemplo del procesamiento de datos**  
Lemur\_catta\_project.rmd

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

 Global Biodiversity Information Facility

Other Formats

PDF file  
español  
français

Search  

Table of Contents

[Biodiversity Data Use](#)

Description

Data Processing  
Key documentation

API  
Cloud Computing  
Downline

## 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](#)

# Ejercicio suplementario: Limpieza de Datos con R

Limpiar su descarga de *Lepus saxatilis*:

- Removiendo ocurrencias donde los medios de establecimiento (**establishmentmeans**) indiquen: controlada (**managed**), introducida (**introduced**) o invasora (**invasive**)
- Filtrando año (**year**) para registros entre 1955 y 2010
- Filtrando registros para registros con Incertidumbre en coordenadas (**coordinate uncertainty**) menor a 10,000 y Precisión de coordenadas (**coordinate precision**) mayor a 0.01
- Removiendo puntos a menos de 2 km de los centroides de país (**country centroids**) y de la capital (**capital centroids**)
- Removiendo puntos a menos de 2 km de un zoológico o herbario

```
library(rgbif)

library(dplyr) # para filtrar y %>%
library(CoordinateCleaner) # para cc_cen,cc_cap,cc_inst

# gbif_download_key <- "0071981-210914110416597"

# primera importación de datos

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

gbif_download %>%
  setNames(tolower(names(.))) %>% # establecer nombres de columna en minúsculas para trabajar con 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) %>% # eliminar los centroides del país en un radio de 2 km
  cc_cap(buffer = 2000) %>% # eliminar los centroides de las capitales en un radio de 2 km
  cc_inst(buffer = 2000) %>% # eliminar zoológicos y herbarios en un radio de 2 km
  glimpse() # mirar los resultados del flujo de trabajo
```

# Paso 1 - Cargar librerías

Este código cargará las funciones que necesitamos para limpiar los datos.

**CoordinateCleaner** es un paquete de R escrito específicamente para la **limpieza de datos de presencia de GBIF** <https://github.com/ropensci/CoordinateCleaner>

```
library(rgbif)
library(dplyr) # para filtrar y %>%
library(CoordinateCleaner) # para cc_cen, cc_cap, cc_inst
```

## Paso 2 - Importar los Datos

Este código importará esos datos de GBIF a R.

Recuerde que la pipeline (%>%) simplemente pasa los resultados de una función a otra función.

```
# gbif_download_key <- "0071981-210914110416597"  
# primera importación de datos  
gbif_download <- occ_download_get(gbif_download_key, overwrite  
= TRUE) %>%  
  occ_download_import()
```

## Paso 3 - Importar datos y limpiar los nombres de las columnas

Establezca los nombres de las columnas en minúsculas. El punto (".") aquí es un código de pipeline especial que hace referencia al objeto `gbif_download`.

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

```
gbif_download %>%  
  setNames(tolower(names(.))) %>% # establecer nombres de  
  columna en minúsculas para trabajar con CoordinateCleaner
```

## Paso 4 - Filtro establishmentMeans

Aquí usamos `filter` del paquete `dplyr` para remover los registros no naturales. “!” significa **negación** en R. El operador `%in%` comprueba si el valor de la columna está en el vector `c ("MANAGED", "INTRODUCED", "INVASIVE")`.

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

## Paso 5 - Filtrar año (year)

Aquí usamos `filter` para mantener únicamente los registros entre 1955 y 2010. El operador `>=` mayor o igual `<=` menor o igual. El operador `&` combina y confirma que **ambas condiciones se cumplan (TRUE)**.

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

# Paso 6 - Filtrar registros con incertidumbre

Aquí usamos `filter` para mantener solo ciertos registros. El operador `|` combina y verifica que **solamente una condición** sea TRUE. La función `is.na` comprueba si falta el registro. `NA` indica valores faltantes en R.

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

# Paso 7 - filtro con CoordinateCleaner

Aquí usamos **CoordinateCleaner** para remover centroides de país y registros cerca de zoológicos y jardines botánicos.

```
cc_cen(buffer = 2000) %>% # remover centroides de país  
cc_cap(buffer = 2000) %>% # remover centroides de capital  
cc_inst(buffer = 2000) %>% # remover zoológicos y  
herbarios
```

# ENLACES A LOS RECURSOS

[GBIF occurrence search](#) (Búsqueda de ocurrencias en GBIF)

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

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

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

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

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