

# Búsqueda, descarga y limpieza de datos de biodiversidad desde GBIF. Una odisea ecoinformática

Seminario Grupo de Trabajo  
EcoInformática (AEET)

31 Enero 2022





## Contenidos

- ¿Odisea ecoinformática?
- Datos de biodiversidad en GBIF
- Descarga de datos desde R





## Odisea

De *Odisea*, título de un poema homérico.

1. f. Viaje largo, en el que abundan las aventuras adversas y favorables al viajero
2. f. Sucesión de peripecias, por lo general desagradables, que le ocurren a alguien





Conserving  
plant diversity  
for future generations

[www.farmerspride.eu](http://www.farmerspride.eu)  
#EUfarmerspride

- Aim: to establish a network for *in situ* conservation and sustainable use of plant genetic resources in Europe.
- To achieve this: to increase knowledge about the occurrence of natural populations of crop wild relatives (CWR)

# Dataset of occurrences of the priority CWR taxa



List of priority CWR contains 863 taxa – 485 classified at the species level and 378 at the infraspecific level.



GBIF and Genesys databases to collect occurrence data of the selected CWR.

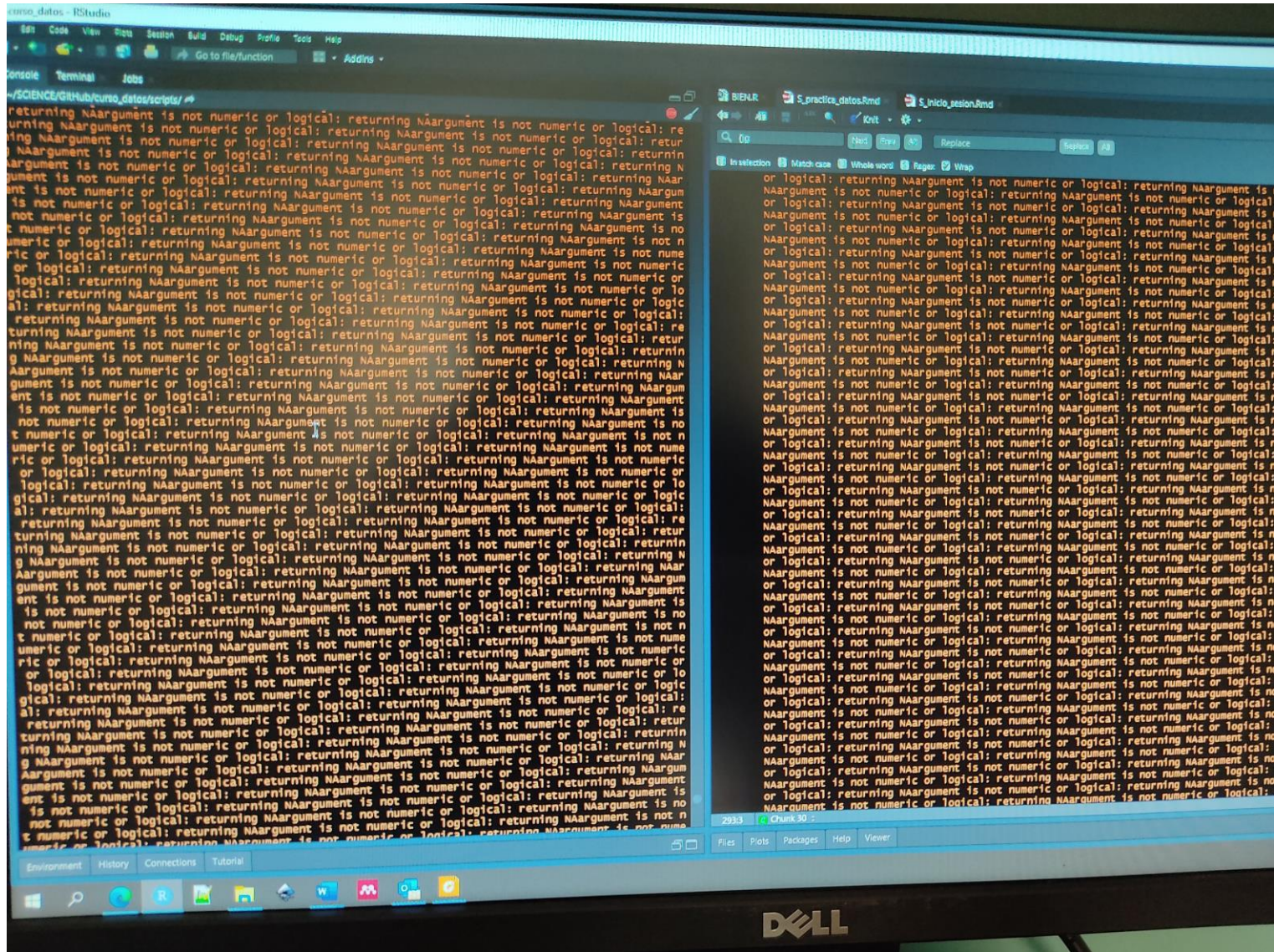


The GBIF database provided 16,534,316 records for 764 taxa.



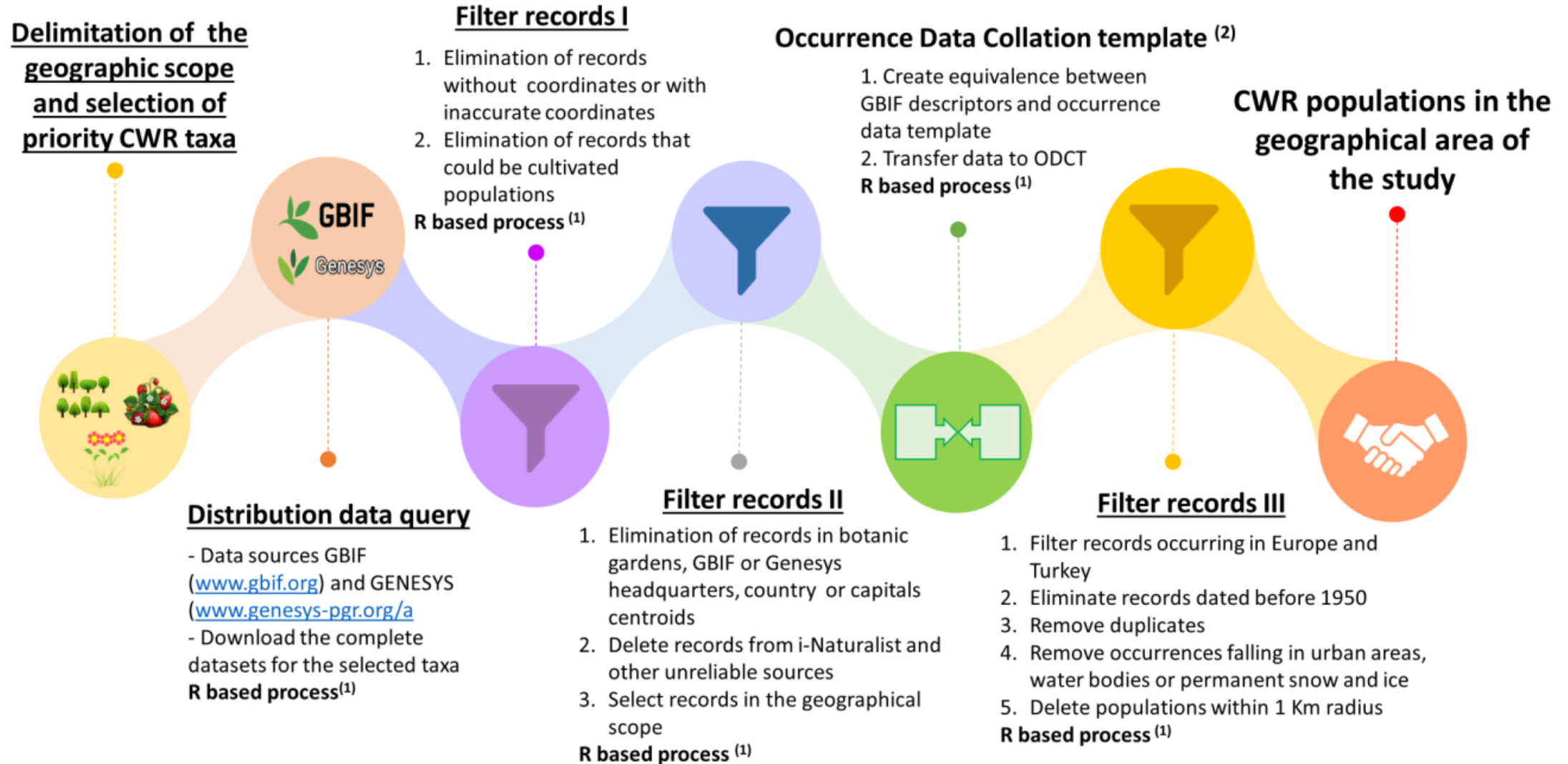
Genesys download resulted in 991,746 accessions for 440 taxa.



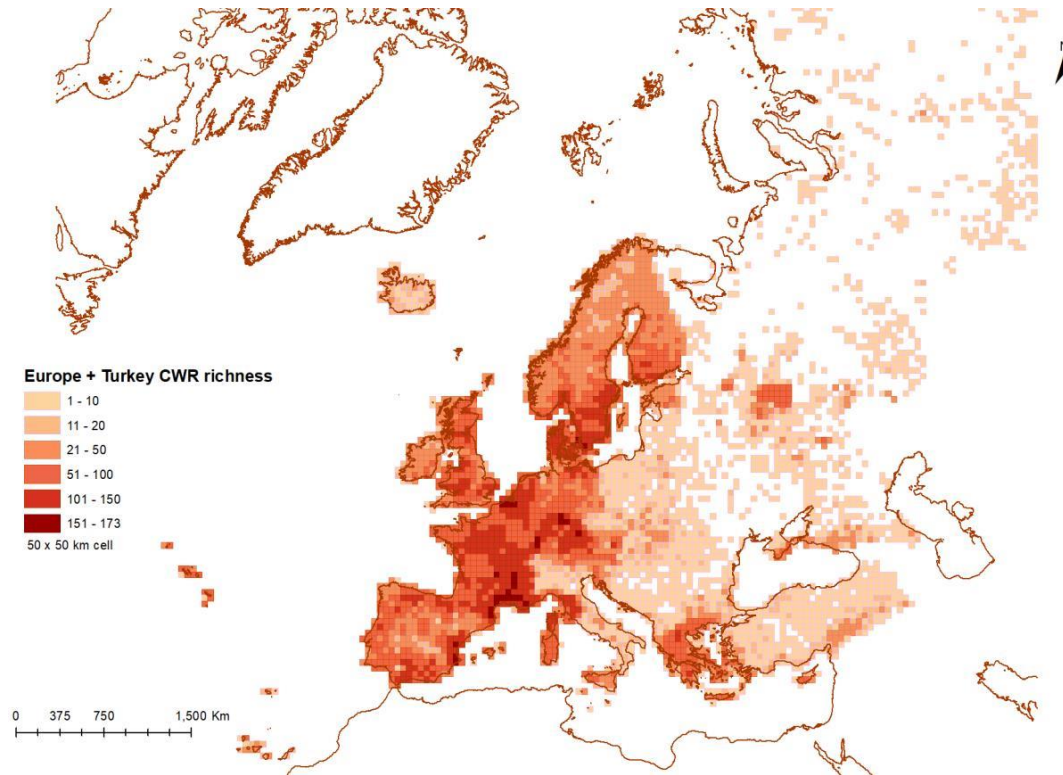




## Steps taken to obtain a high quality dataset of occurrences of the priority CWR taxa.



- 3,094,231 sites for 616 priority taxa distributed across 43 countries
- The largest databases of occurrences for the target CWR generated so far for the entire world and for Europe and Turkey.







Misión GBIF

Acceso libre y gratuito a los datos de biodiversidad

Registros biológicos	Conjuntos de datos	Instituciones que publican	Artículos científicos usando datos
1.921.499.521	65.023	1787	6668



## ¿Qué tipo de datos se encuentran en GBIF?

### DATOS

- Occurrences - Una observación por registro
- Checklist - Un taxón por registro
- Eventos de muestreo – Un evento por registro

**Evidencia de la presencia de una especie (u otro taxón) en un lugar determinado y en una fecha concreta**

### METADATOS

“Datos sobre los datos”. Proporcionan el contexto para los datos, describen su contenido, licencias, datos de contacto, etc. Permite al usuario evaluar la idoneidad de los datos para ser usados

**qué / dónde / cuándo / cómo / quién**

### CITACIÓN

- Descargas de datos de ocurrencia (all contributing datasets)
- Conjuntos de datos individuales (Single datasets)
- Conjuntos de datos derivados (Derived datasets)

**Los usuarios que descargan conjuntos de datos individuales o resultados de búsqueda y los utilizan en la investigación o la política se comprometen a citarlos utilizando un DOI**

## DATOS



Especímenes preservados en colecciones



Especímenes conservados en herbarios



Libretas de campo, Tesis, literatura, informes



Ciencia ciudadana



Imágenes de satélite



Capturas de audio, vídeos, foto-trampeo



Checklist



Evaluaciones de impacto ambiental, inventarios

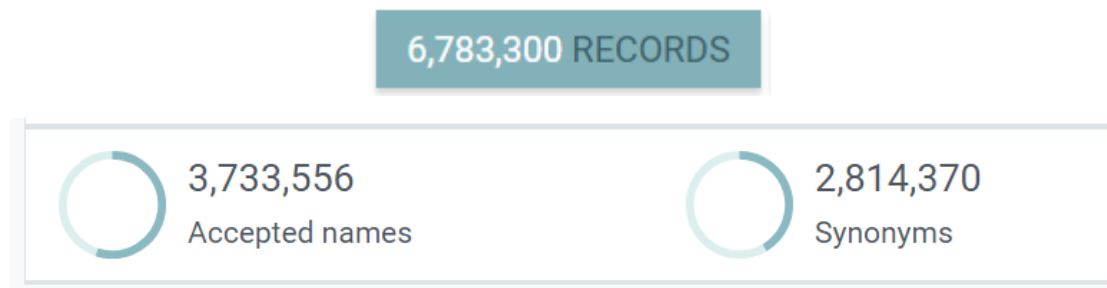


## DATOS

### GBIF Backbone Taxonomy

El árbol taxonómico de GBIF ([GBIF Backbone Taxonomy](https://hosted-datasets.gbif.org/datasets/backbone/)) es una clasificación única y sintética de gestión con el objetivo de cubrir todos los nombres de los que se ocupa GBIF.

Permite a GBIF integrar la información basada en nombres de diferentes recursos como EOL, Genbank o IUCN.



## DATOS

### Darwin Core (DwC)

Los datos en GBIF se organizan en una estructura de tabla utilizando los nombres de los términos de Darwin Core (DwC) como nombres de columna.

DwC es una norma del [Biodiversity Information Standards](#). Organización sin ánimo de lucro dedicada a desarrollar normas de información sobre la biodiversidad.

Mantenido por [Darwin Core Maintenance Group](#).

Incluye un glosario de términos destinado a facilitar el intercambio de información sobre la diversidad biológica proporcionando identificadores, etiquetas y definiciones (<https://dwc.tdwg.org/terms/>).



## DATOS

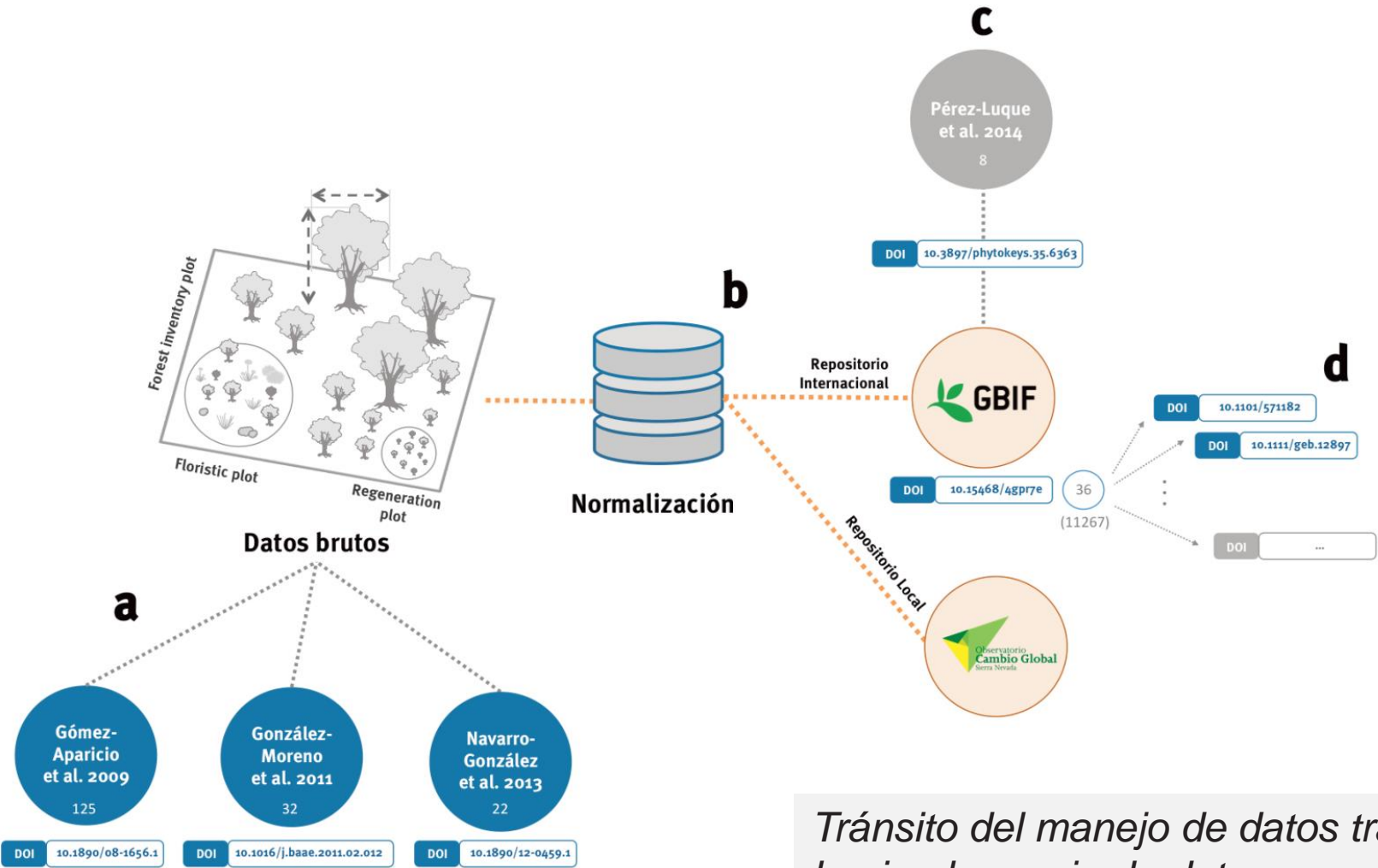
### Campos DwC requeridos

- [occurrenceID](#)
- [basisOfRecord](#)
- [scientificName](#)
- [eventDate](#)

### Campos DwC recomendados

- [taxonRank](#) - to substantiate scientificName
- [kingdom](#) - and other higher taxonomy if possible
- [decimalLatitude](#) & [decimalLongitude](#) & [geodeticDatum](#) - to provide a specific point location
- [countryCode](#)
- [individualCount](#) / [organismQuantity](#) & [organismQuantityType](#) - to record the quantity of a species occurrence

CITACIÓN



*Tránsito del manejo de datos tradicional hacia el manejo de datos reproducible*





# Descarga de datos con rgbif

[https://github.com/CarlosLaraR/teaching\\_R/tree/main/GBIF](https://github.com/CarlosLaraR/teaching_R/tree/main/GBIF)

[Buscar y recuperar datos del Global Biodiversity Information Facility \(GBIF\) utilizando R \(githack.com\)](#)



## Beneficios de la ciencia reproducible

- Trazabilidad
- La utilización de código permite la automatización
- Reducción drástica del riesgo de errores
- Facilita la colaboración.
- Ahorro de tiempo y esfuerzo al reutilizar código en otros proyectos
- La reproducibilidad es un sello de calidad
- La reproducibilidad aumenta el impacto de las publicaciones

# Búsqueda, descarga y limpieza de datos de biodiversidad desde GBIF. Una odisea ecoinformática

Seminario Grupo de Trabajo  
EcoInformática (AEET)

31 Enero 2022

