I Taller GBIF.ES:
Mejora de la calidad de datos de biodiversidad

Cristina Ronquillo

Ayudante investigación

cristinaronquillo@mncn.csic.es

mncn 250 1771 2021 museo nacional de ciencias naturales

CSIC

Consejo Superior de Investigaciones Científicas

GBIF Global Biodiversity Information Facility

Objetivo

Conocer los principales aspectos necesarios para evaluar, filtrar, corregir y validar conjuntos de datos de biodiversidad.

Objetivos

- ✓¿Qué tengo mirar?
- ✓¿Cómo comprobar los datos?
- ✓¿Qué tengo que descartar?
- ✓¿Qué tengo que corregir y cómo puedo hacerlo?

Objetivos

- ✓¿Qué tengo mirar?
- ✓¿Cómo comprobar los datos?
- ✓¿Qué tengo que descartar?
- ✓¿Qué tengo que corregir y cómo puedo hacerlo?

Ausencia de información / Información incompleta / Información incorrecta

Objetivos

✓¿Para qué?

GBIF INTEGRATED PUBLISHING TOOLKIT (IPT)
free and open access to biodiversity data

email | password | login | ENGLISH

Home | About

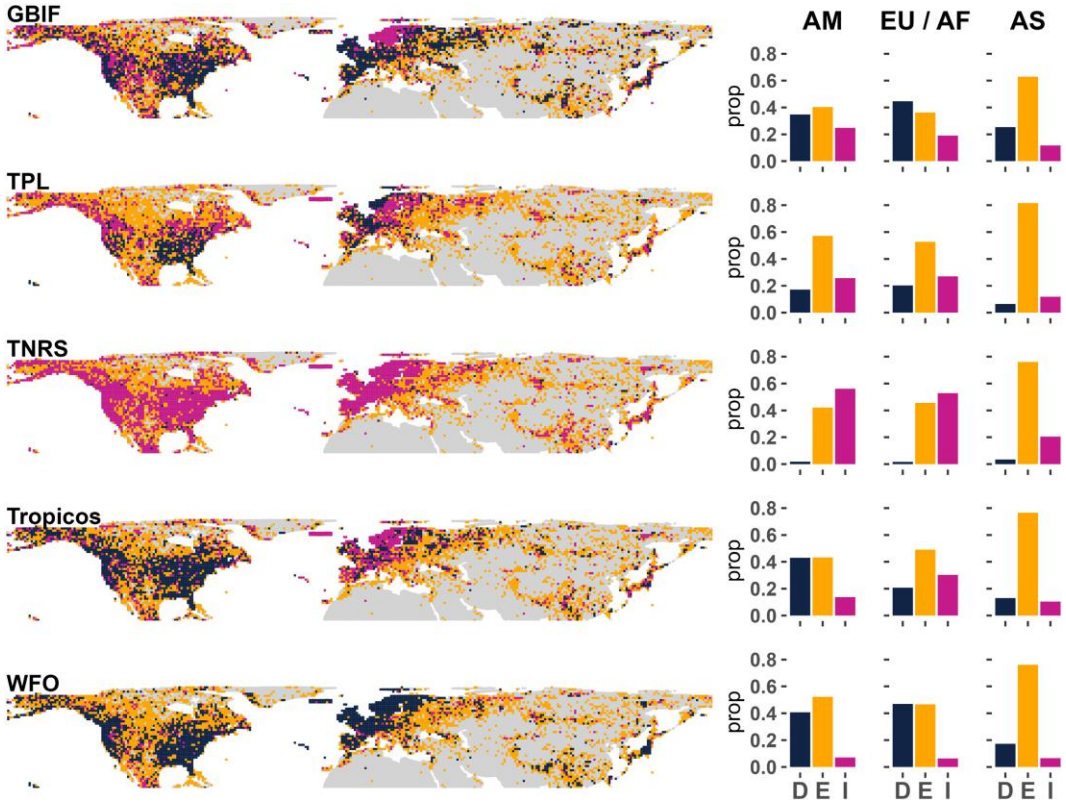
Hosted resources available through this IPT

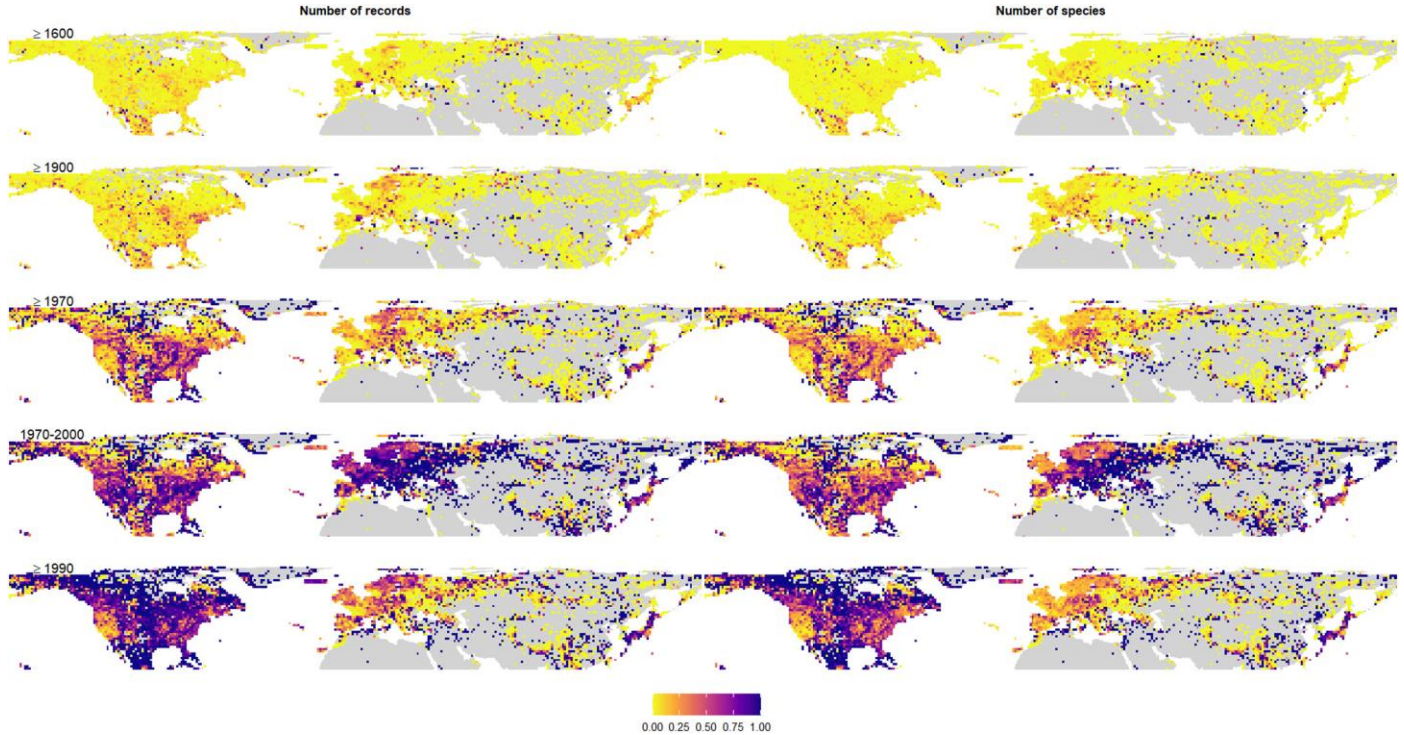
Filter:

Logo	Name	Organization	Type	Subtype	Records	Last modified	Last publication	Next publication
--	Algae (S)	GBIF-Sweden	Occurrence	Specimen	15,953	2021-04-01	2016-01-05	--
--	Artportalen (Swedish Species Observation System)	ArtDatabanken	Occurrence	Observation	80,765,776	2021-04-09	2021-04-09	2021-04-16 15:00:18
--	Axel W. Erikssons African Bird Collection at Vänersborg Museum	GBIF-Sweden	Occurrence	Specimen	1,000	2021-03-30	2020-12-08	--
--	Beetles (LSM)	GBIF-Sweden	Occurrence	--	13,450	2021-03-18	2021-03-18	--
--	Bird Collection of Helsingborg Museums	GBIF-Sweden	Occurrence	Specimen	2,530	2017-09-18	2017-08-16	--

Objetivos

✓¿Para qué?

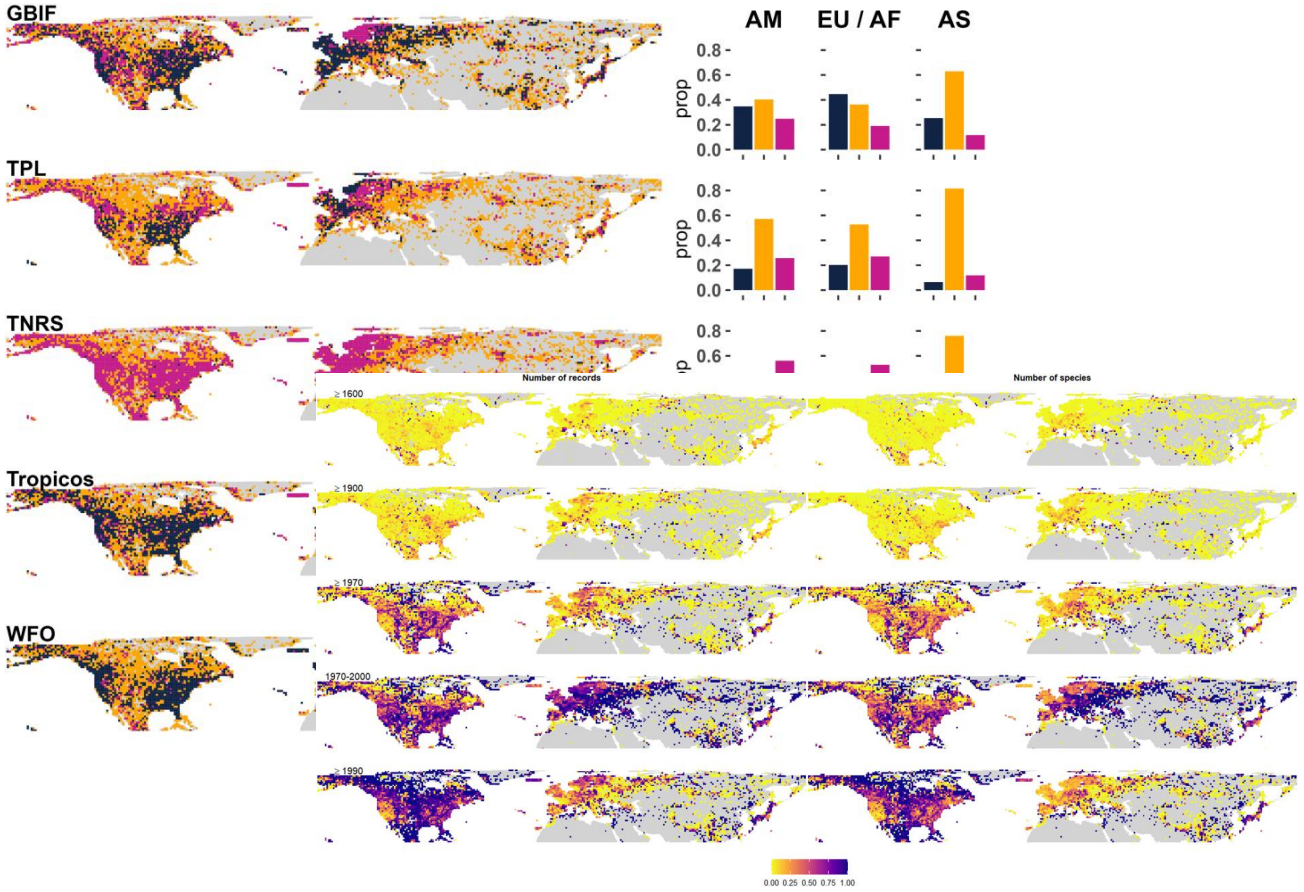


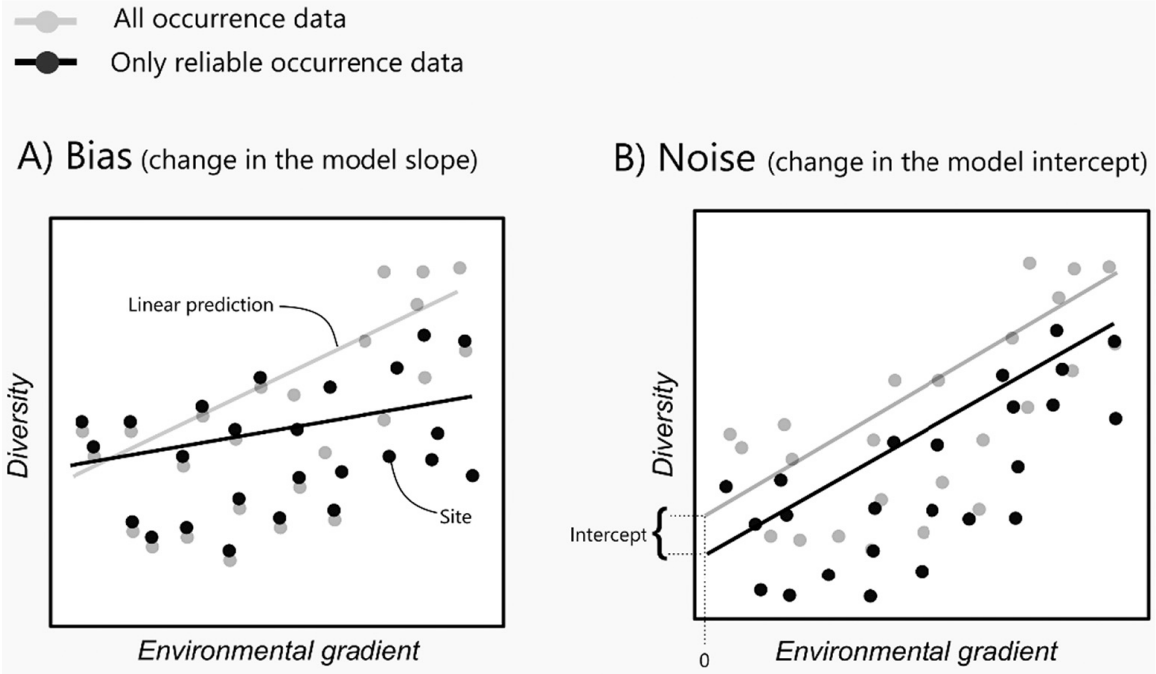


Ronquillo et al. Ecology and Evolution. 2023;13:e10786.

Objetivos

✓¿Para qué?



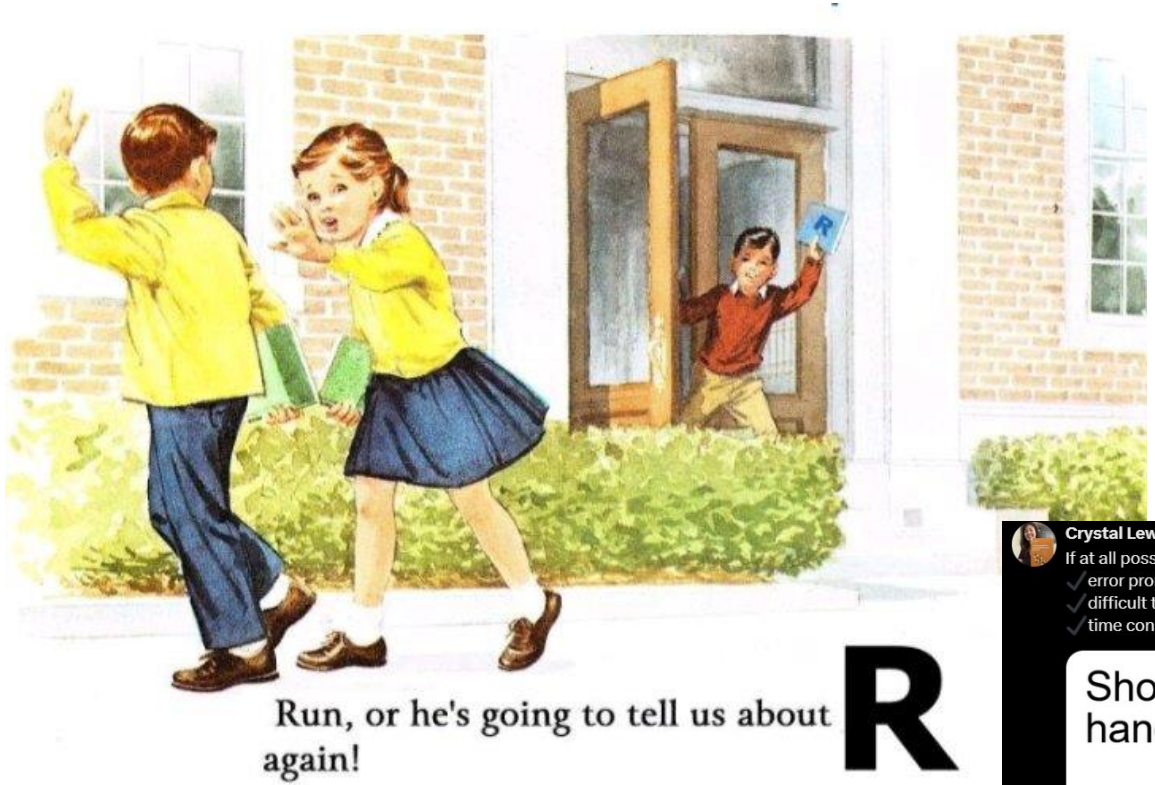


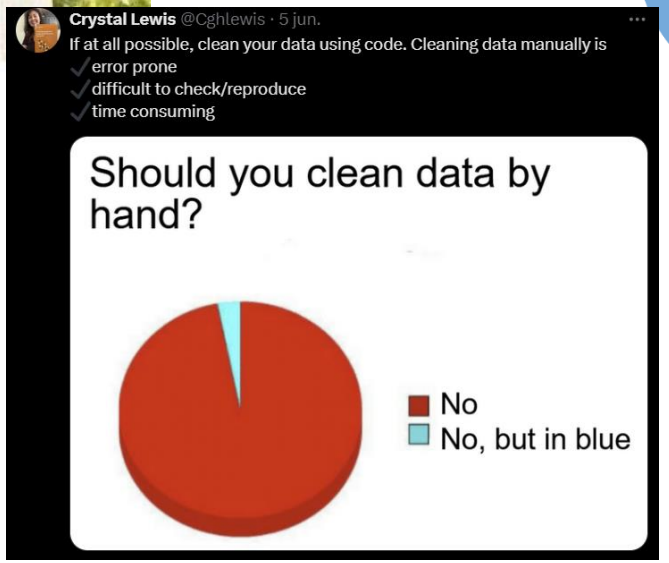
Ronquillo et al. Ecology and Evolution. 2023;13:e10786.

Rodrigues et al. 2022. Ecological Informatics https://doi.org/10.1016/j.ecoinf.2022.101625

Herramientas:

Herramientas:

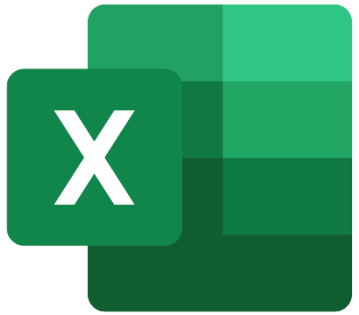




RStudio®

Herramientas:

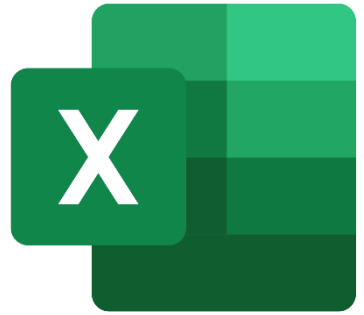
Herramientas:



¿Sé usar Excel?

Herramientas:

¿Sé usar Excel?



- ❑ Abrir un archivo de texto (CSV, TSV)

Herramientas:

¿Sé usar Excel?

- ❑ Abrir un archivo de texto (CSV, TSV)
- ❑ Separadores de campo: tabuladores = ‘\t’

comas ‘,’

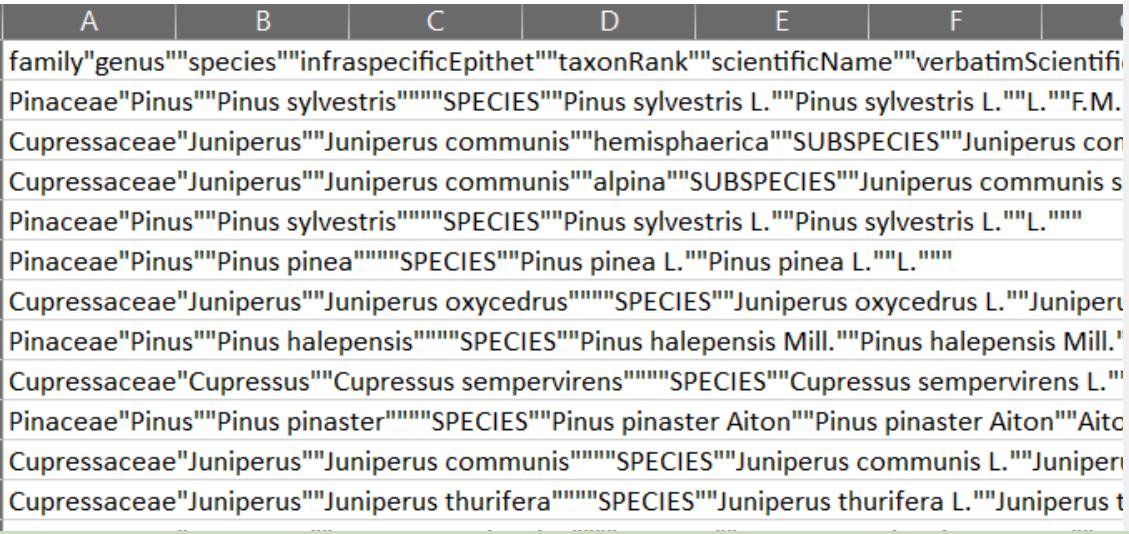
punto y coma ‘;’

A	B	C	D	E	F	G	H	I	J	K
family"genus""species""infraspecificEpithet""taxonRank""scientificName""verbatimScientificName""verbatimScientificNameAuthorship""identifiedBy"										
Pinaceae"Pinus""Pinus sylvestris""""SPECIES""Pinus sylvestris L.""Pinus sylvestris L.""L.""F.M. Vázquez"										
Cupressaceae"Juniperus""Juniperus communis""hemisphaerica""SUBSPECIES""Juniperus communis subsp. hemisphaerica (Jacq. & C.Presl) Nyman""Juni										
Cupressaceae"Juniperus""Juniperus communis""alpina""SUBSPECIES""Juniperus communis subsp. alpina (Suter) Celak.""Juniperus communis L. subsp. al										
Pinaceae"Pinus""Pinus sylvestris""""SPECIES""Pinus sylvestris L.""Pinus sylvestris L.""L."""										
Pinaceae"Pinus""Pinus pinea""""SPECIES""Pinus pinea L.""Pinus pinea L.""L."""										
Cupressaceae"Juniperus""Juniperus oxycedrus""""SPECIES""Juniperus oxycedrus L.""Juniperus oxycedrus L.""L."""										
Pinaceae"Pinus""Pinus halepensis""""SPECIES""Pinus halepensis Mill.""Pinus halepensis Mill.""Mill."""										
Cupressaceae"Cupressus""Cupressus sempervirens""""SPECIES""Cupressus sempervirens L.""Cupressus sempervirens L.""L."""										
Pinaceae"Pinus""Pinus pinaster""""SPECIES""Pinus pinaster Aiton""Pinus pinaster Aiton""Aiton"""										
Cupressaceae"Juniperus""Juniperus communis""""SPECIES""Juniperus communis L.""Juniperus communis L.""L."""										
Cupressaceae"Juniperus""Juniperus thurifera""""SPECIES""Juniperus thurifera L.""Juniperus thurifera L.""L."""										

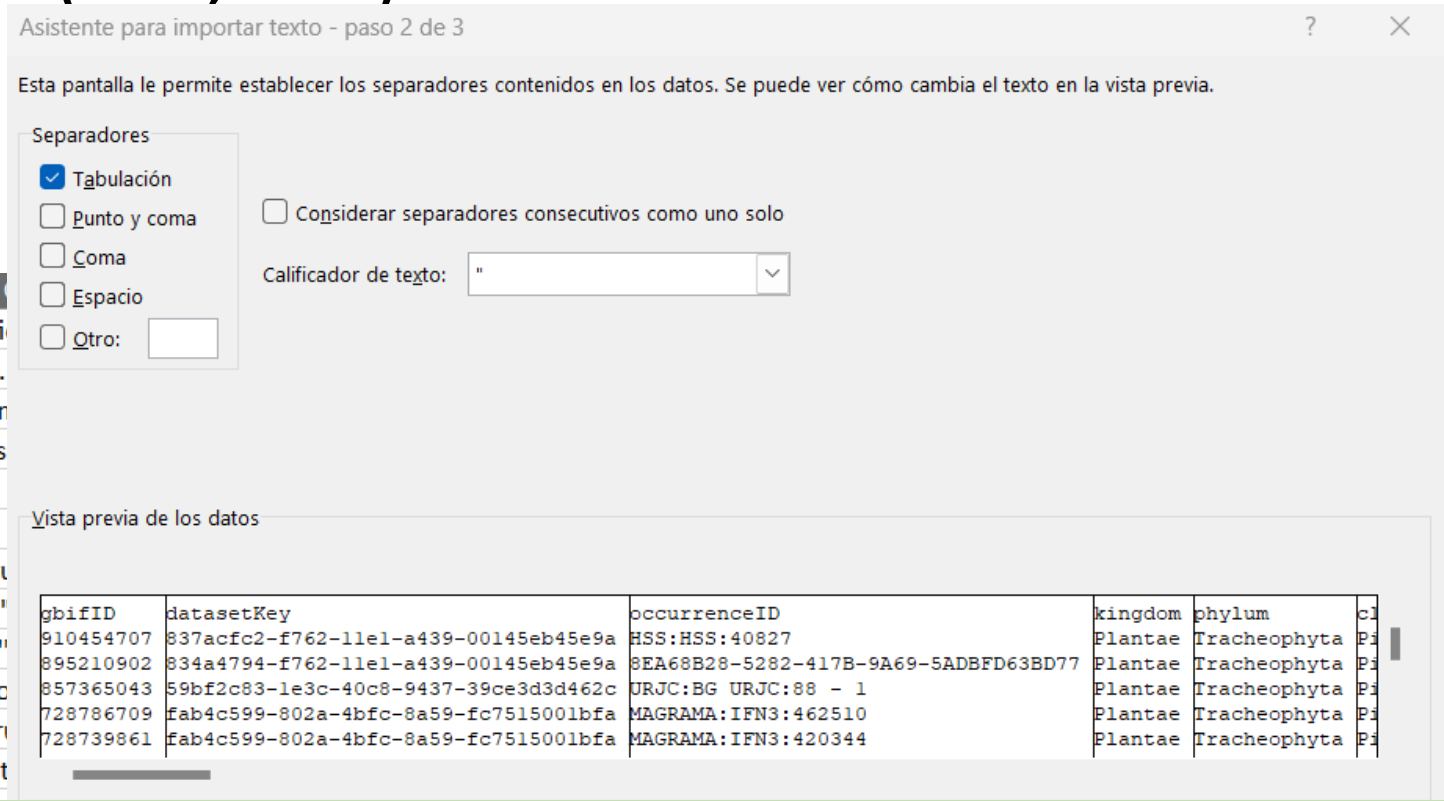
Herramientas:

¿Sé usar Excel?

- ❑ Abrir un archivo de texto (CSV, TSV)
- ❑ Separadores de campo:



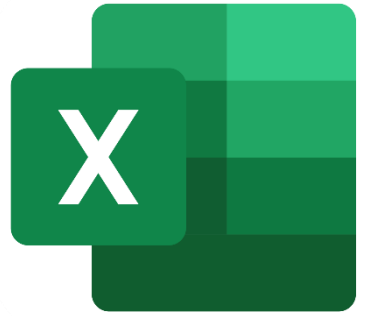
A	B	C	D	E	F
family"genus""species""infraspecificEpithet""taxonRank""scientificName""verbatimScientifi					
Pinaceae"Pinus""Pinus sylvestris""""SPECIES""Pinus sylvestris L.""Pinus sylvestris L.""L.""F.M.					
Cupressaceae"Juniperus""Juniperus communis""hemisphaerica""SUBSPECIES""Juniperus cor					
Cupressaceae"Juniperus""Juniperus communis""alpina""SUBSPECIES""Juniperus communis s					
Pinaceae"Pinus""Pinus sylvestris""""SPECIES""Pinus sylvestris L.""Pinus sylvestris L.""L."""					
Pinaceae"Pinus""Pinus pinea""""SPECIES""Pinus pinea L.""Pinus pinea L.""L."""					
Cupressaceae"Juniperus""Juniperus oxycedrus""""SPECIES""Juniperus oxycedrus L.""Juniperu					
Pinaceae"Pinus""Pinus halepensis""""SPECIES""Pinus halepensis Mill.""Pinus halepensis Mill."					
Cupressaceae"Cupressus""Cupressus sempervirens""""SPECIES""Cupressus sempervirens L."					
Pinaceae"Pinus""Pinus pinaster""""SPECIES""Pinus pinaster Aiton""Pinus pinaster Aiton""Aitc					
Cupressaceae"Juniperus""Juniperus communis""""SPECIES""Juniperus communis L.""Juniperu					
Cupressaceae"Juniperus""Juniperus thurifera""""SPECIES""Juniperus thurifera L.""Juniperus t					



gbifID	datasetKey	occurrenceID	kingdom	phylum	cl
910454707	837acfc2-f762-11e1-a439-00145eb45e9a	HSS:HSS:40827	Plantae	Tracheophyta	Pi
895210902	834a4794-f762-11e1-a439-00145eb45e9a	8EA68B28-5282-417B-9A69-5ADBFD63BD77	Plantae	Tracheophyta	Pi
857365043	59bf2c83-1e3c-40c8-9437-39ce3d3d462c	URJC:BG URJC:88 - 1	Plantae	Tracheophyta	Pi
728786709	fab4c599-802a-4bfc-8a59-fc7515001bfa	MAGRAMA:IFN3:462510	Plantae	Tracheophyta	Pi
728739861	fab4c599-802a-4bfc-8a59-fc7515001bfa	MAGRAMA:IFN3:420344	Plantae	Tracheophyta	Pi

Herramientas:

¿Sé usar Excel?



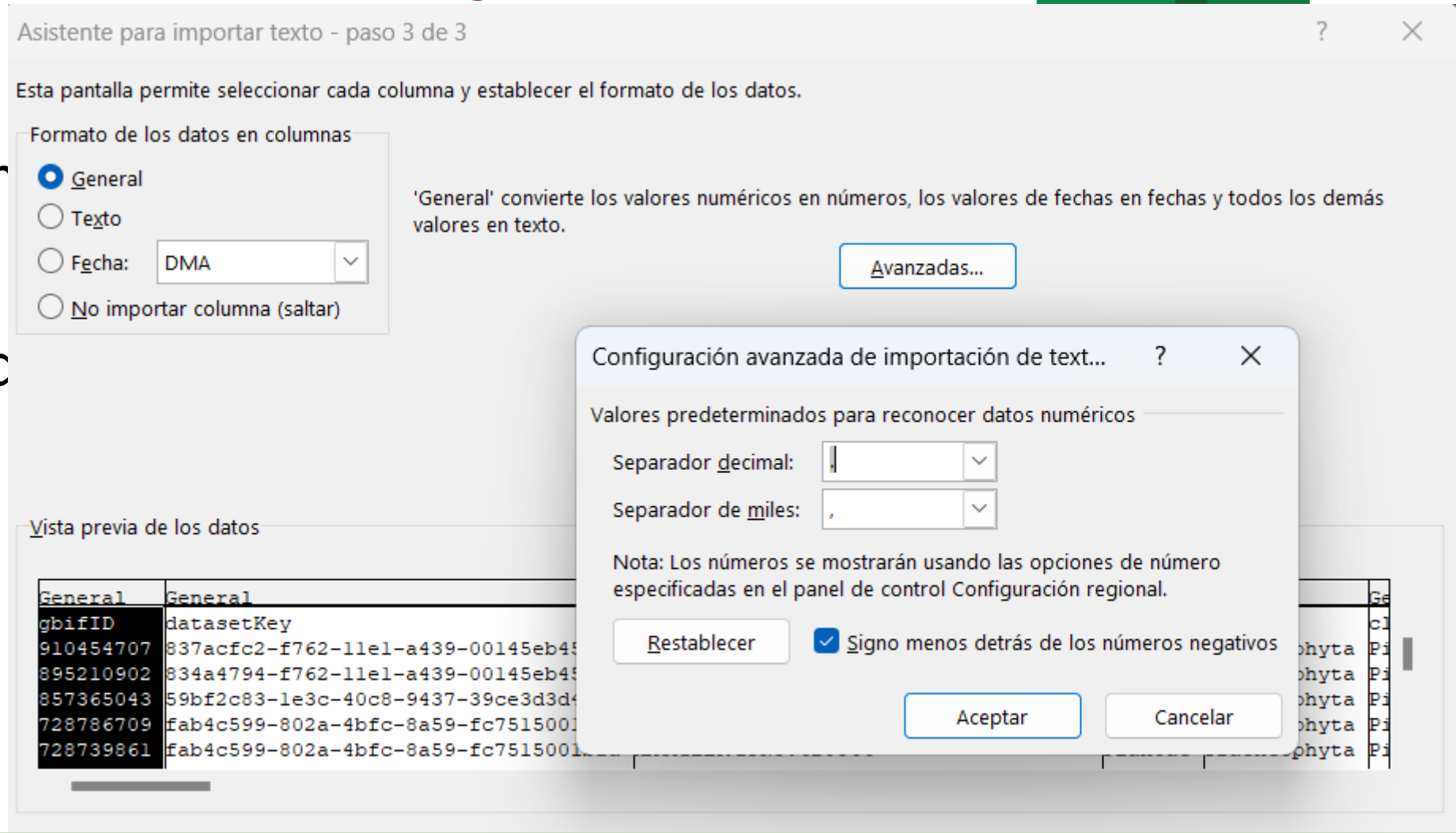
- ❑ Abrir un archivo de texto (CSV, TSV)
- ❑ Separadores de campo
- ❑ Símbolo de decimales

decimalLatitude	decimalLongitude	coordinateUncertainty
403,638	-42,837	25.0
403,644	-42,248	25.0
403,645	-42,131	25.0
403,646	-42,013	25.0
403,661	-3.92	
403,661	-3.92	
403,662	-40,364	25.0
403,667	-39,893	25.0
403,675	-4.31	
403,697	-35,182	25.0
403,697	-35,182	25.0
403,704	-4,071,549	5197.0
403,722	-4,331	25.0
403,724	-43,192	25.0
403,725	-43,074	25.0
403,726	-42,957	25.0
403,726	-42,957	25.0
403,729	-42,721	25.0
403,737	-42,014	25.0
403,737	-4,328,028	16.0
403,797	-32,356	25.0
403,806	-42,854	
403,812	-43,312	25.0
403,812	-43,312	25.0
403,815	-43,076	25.0
403,816	-42,958	25.0
403,818	-42,841	25.0
403,822	-3,598,478	300.0

Herramientas:

¿Sé usar Excel?

- ❑ Abrir un arch
- ❑ Separadores
- ❑ Símbolo de c



Herramientas:

¿Sé usar Excel?

- ❑ Abrir un archivo de texto (CSV, TSV)
- ❑ Separadores de campo
- ❑ Símbolo de decimales
- ❑ Encoding

(Windows, ASCII, UTF-8)

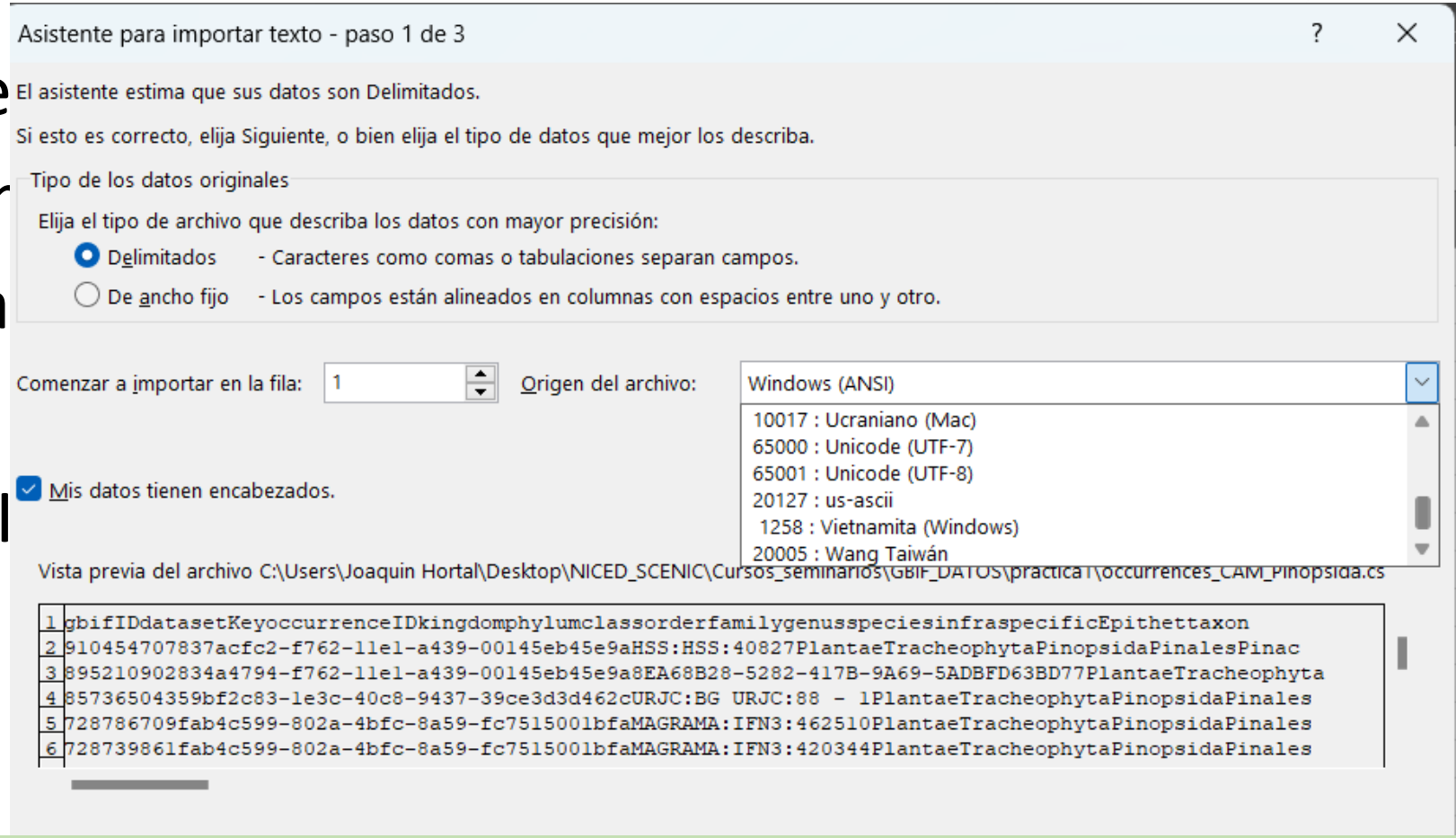
	verbatimScientificName	verbatimScientific	countryCo	locality	st
2	Pinus halepensis Mill.	Mill.	ES	Colmenar de Oreja	M
3	Pinus halepensis Mill.	Mill.	ES	Villamanrique de Tajo, carretera a Belmonte, finca La Encomienda	M
4	Juniperus oxycedrus L.	L.	ES	PelahustÃƒÂ¡n	T
5	Pinus halepensis Mill.	Mill.	ES	Morata de TajuÃƒÂ±a	M
6	Cupressus sempervirens L.		ES		
7	Juniperus communis L.	L.	ES	Cenicientos	M
8	Juniperus communis L.	L.	ES	Cenicientos	M
9	Juniperus oxycedrus L.	L.	ES	Villa del Prado	M
10	Pinus pinaster Aiton	Aiton	ES	Cenicientos	M
11	Juniperus oxycedrus L.	L.	ES	PeÃ±a de Cenicientos	M
12	Juniperus oxycedrus L.	L.	ES	Villa del Prado	M
13	Pinus halepensis Mill.	Mill.	ES	Arganda del Rey	M
14	Juniperus oxycedrus L.	L.	ES	Villa del Prado	M
15	Pinus pinaster Aiton	Aiton	ES		
16	Juniperus communis L.	L.	ES	San MartÃƒÂ-n de Valdeiglesias	M
17	Juniperus communis L.	L.	ES	Navas del Rey	M
18	Juniperus oxycedrus L.	L.	ES	Villamantilla	M
19	Juniperus thurifera L.	L.	ES	Villaviciosa de OdÃƒÂ³n	M
20	Pinus pinea L.	L.	ES	Villaviciosa de OdÃƒÂ³n	M
21	Pinus pinea L.	L.	ES	Villaviciosa de OdÃƒÂ³n	M
22	Pinus pinea L.	L.	ES	San MartÃƒÂ-n de Valdeiglesias	M
23	Pinus pinea L.	L.	ES	San MartÃƒÂ-n de Valdeiglesias	M
24	Juniperus oxycedrus L.	L.	ES	Colmenar del Arroyo	M
25	Pinus pinea L.	L.	ES	San MartÃƒÂ-n de Valdeiglesias	M

Herramientas:

¿Sé usar Excel?

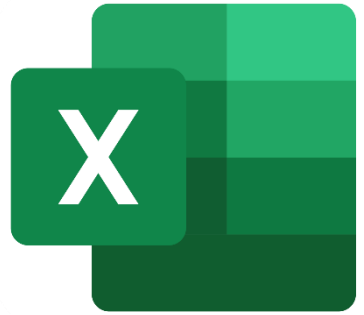
- ❑ Abrir un archivo de
- ❑ Separadores de car
- ❑ Símbolo de decima
- ❑ Encoding

(Windows, ASCII



Herramientas:

¿Sé usar Excel?



- ❑ **AVISO**: Si guardáis un archivo sin atender a su formato la siguiente vez que lo abráis o compartáis con otro ordenador pueden haber lágrimas

- ❑ **SIEMPRE** mantened una copia con la versión original de los datos

- ❑ **SIEMPRE** usar y mantener identificadores únicos (gbifID)

Trabajaremos a dos niveles:

Aprender a validar los datos generados en el proceso de investigación propio o aquellos recolectados por múltiples investigadores en un proyecto y detectar los errores más habituales.

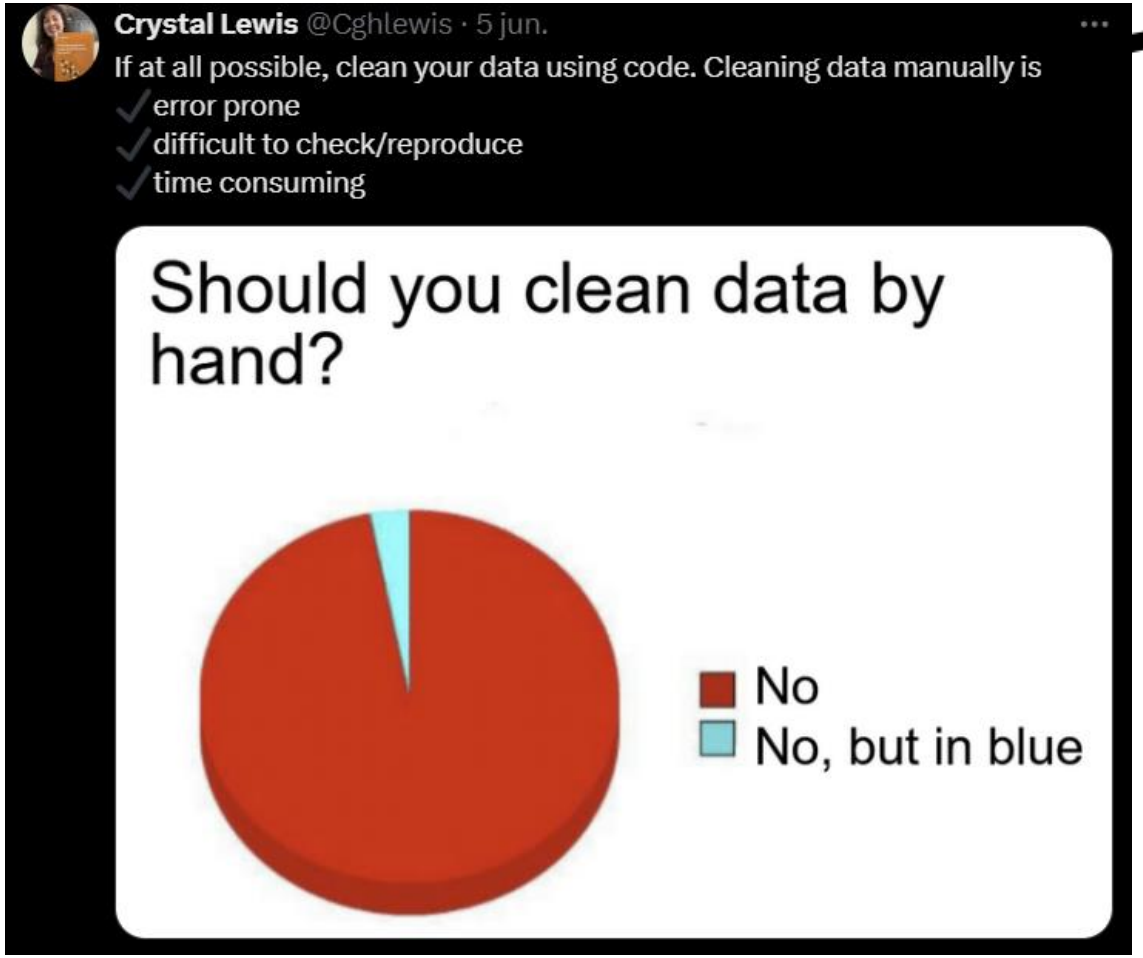
Establecer el tratamiento básico que debemos dar a los datos que **descargamos de repositorios** públicos para poder utilizarlos en nuestra investigación.

Trabajaremos a dos niveles:

Aprender a validar los datos generados en el **proceso de investigación propio** o aquellos recolectados por múltiples investigadores en un proyecto y detectar los errores más habituales.

Establecer el tratamiento básico que debemos dar a los datos que **descargamos de repositorios** públicos para poder utilizarlos en nuestra investigación.

Trabajaremos a dos niveles:

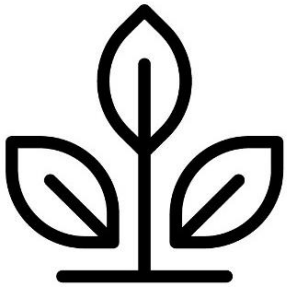


Establecer el tratamiento básico que debemos dar a los datos que **descargamos de repositorios** públicos para poder utilizarlos en nuestra investigación.

¿Qué es un registro biológico?

¿Qué es un registro biológico?

*‘‘Información de que un determinado taxon (‘**qué**’)* *aparece en una localización específica (‘dónde’)* en un momento dado (‘cuándo’) y recogida por alguien (‘quién’)”



¿Qué es un registro biológico?

*'Información de que un determinado taxon ('qué') aparece en una localización específica ('**dónde**')* en un momento dado ('cuándo') y recogida por alguien ('quién')"

¿Qué es un registro biológico?

'Información de que un determinado taxon ('qué') aparece en una localización específica ('dónde') en un momento dado ('**cuándo**') y recogida por alguien ('quién')"

¿Qué es un registro biológico?

'Información de que un determinado taxon ('qué') aparece en una localización específica ('dónde') en un momento dado ('cuándo') y recogida por alguien ('**quién**')"

Bases de datos

BIEN

Atlas of Living Australia
ala.org.au

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

GLOBAL
Building a Global Consortium of Bryophytes and Lichens

TRY
Plant Trait Database

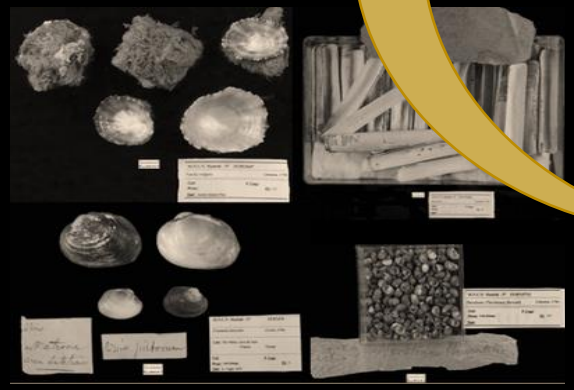
GBIF

Bases de datos

Ciencia Ciudadana

Investigación

Colecciones



Literatura

¿Por dónde empezamos?

Trabajando con registros de presencia de especies

Planteamiento de trabajo

Necesito hacer un modelo de distribución en la península Ibérica para 10 especies de musgos

Necesito analizar los ‘shortfalls’ en Brazil usando todas las especies de termitas del mundo

Necesito analizar la distribución latitudinal de todos los musgos de la región templada del hemisferio norte

Necesito hacer modelos predictivos para saber el nicho de las especies de los ‘drylands’ en esta checklist

Necesito analizar la distribución de 2 especies de *Quercus* en la región mediterránea

Trabajando con registros de presencia de especies

Planteamiento de trabajo

Necesito hacer un modelo de distribución en la península Ibérica para 10 especies de musgos

Necesito analizar los ‘shortfalls’ en Brazil usando tod… …es de termitas del

Necesito analizar la distribución latitudinal de todos los musgos de la región te… del hemi…

‘Que los datos sean buenos’

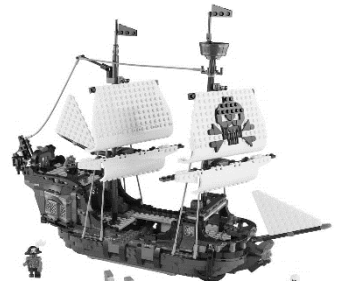
Necesito hacer modelos predictivos para saber el nicho de las especies de los ‘drylands’ en esta checklist

Necesito analizar la distribución de 2 especies de *Quercus* en la región mediterránea

Trabajando con registros de presencia de especies

Planteamiento de trabajo

Objetivo



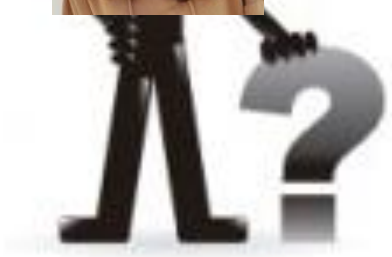
¿Usando alguna taxonomía concreta?

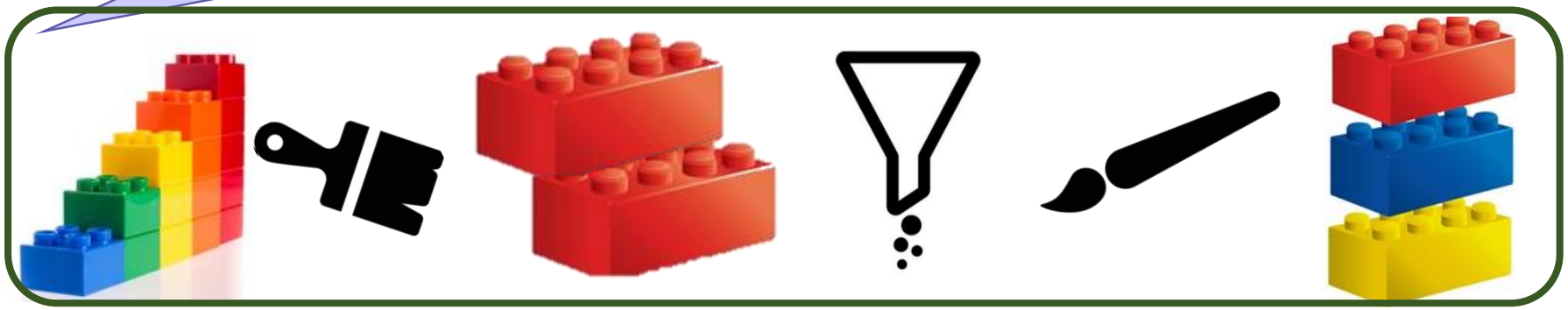
¿En qué periodo de tiempo?

¿Sólo observaciones o **todo**?

¿Con año de colecta y mes o da igual?

¿Sólo en la región nativa o la introducida también?…





Trabajando con registros de presencia de especies

Planteamiento de trabajo

¿Usando alguna taxonomía concreta?

¿En qué periodo de tiempo?

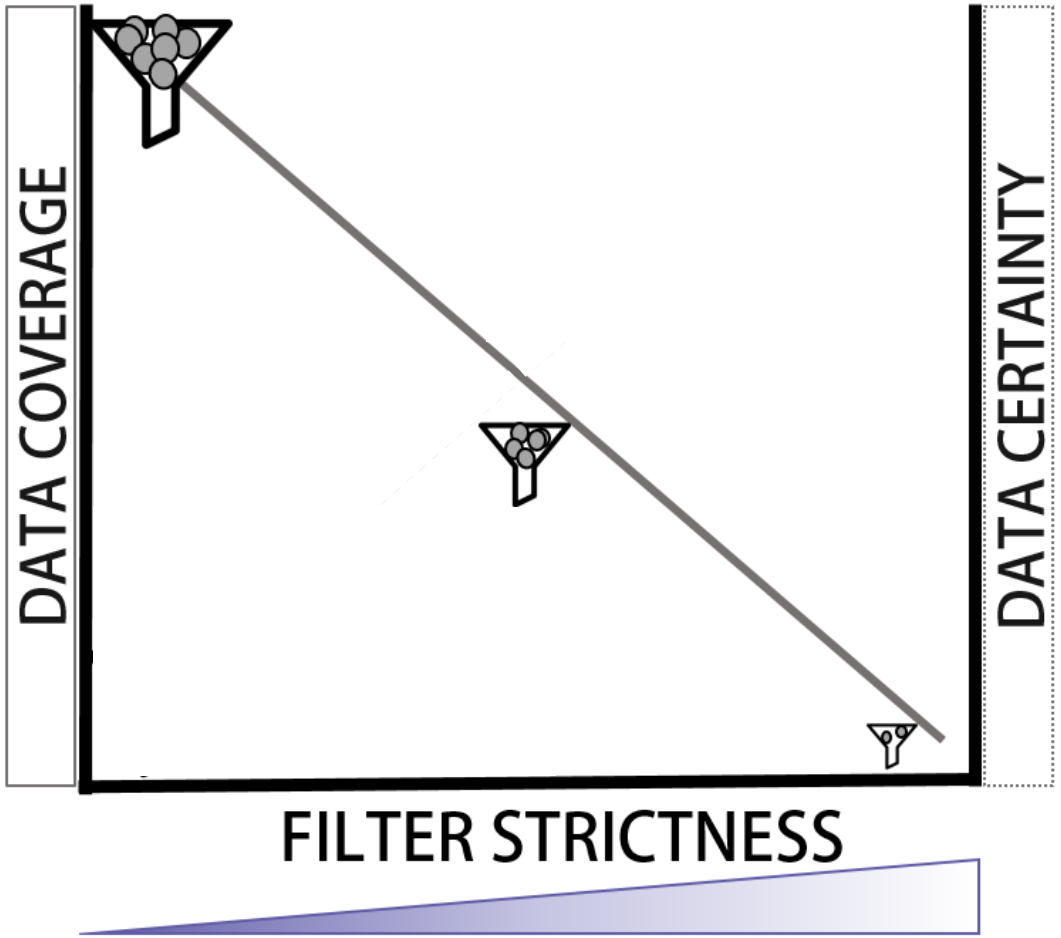
¿Sólo observaciones o **todo**?

¿Con año de colecta y mes o da igual?

¿Sólo en la región nativa o la introducida también?…

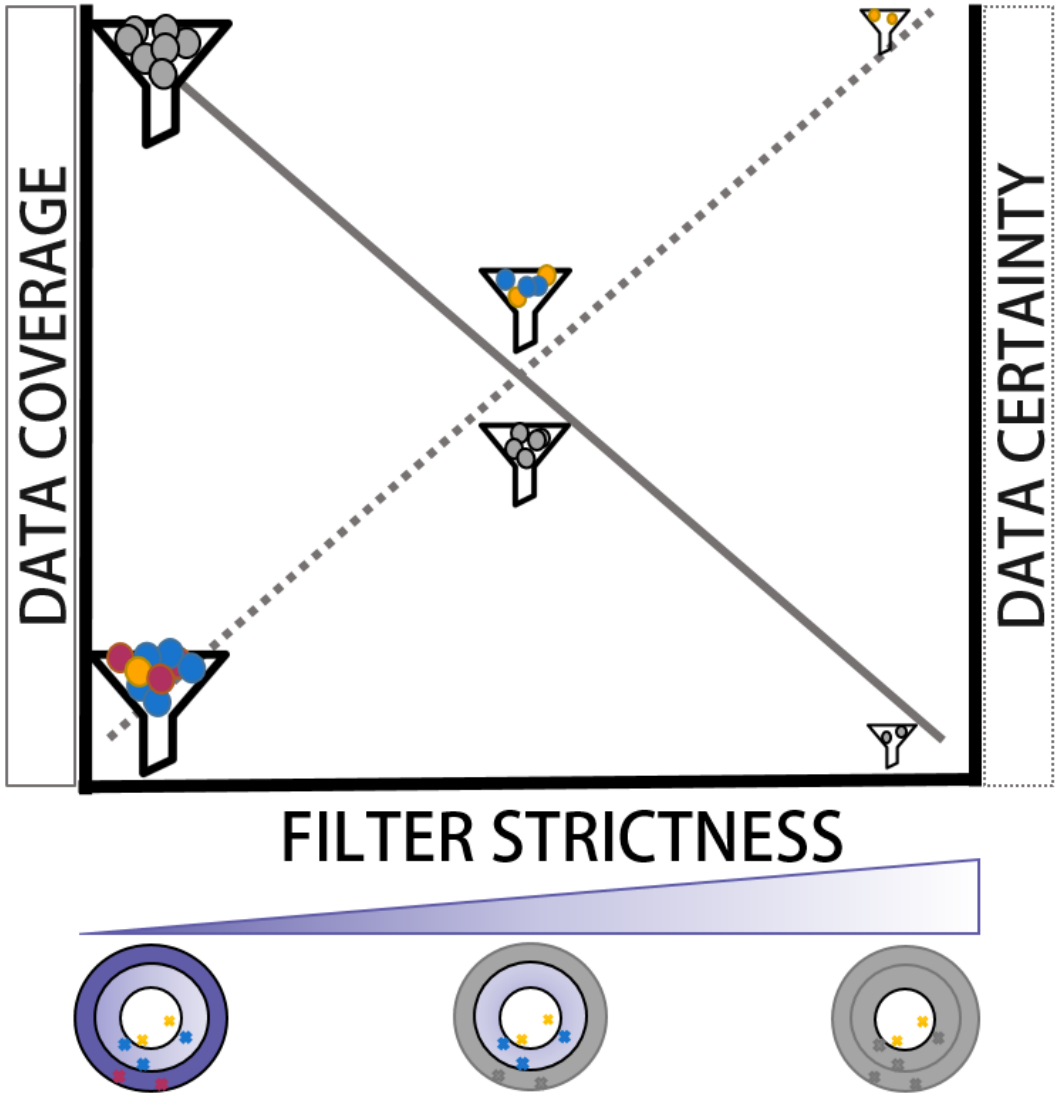
Trabajando con registros de presencia de especies

Número de registros



Trabajando con registros de presencia de especies

Número de registros



Precisión y exactitud

Trabajando con registros de presencia de especies

Trabajando con registros de presencia de especies

Trabajando con registros de presencia de especies

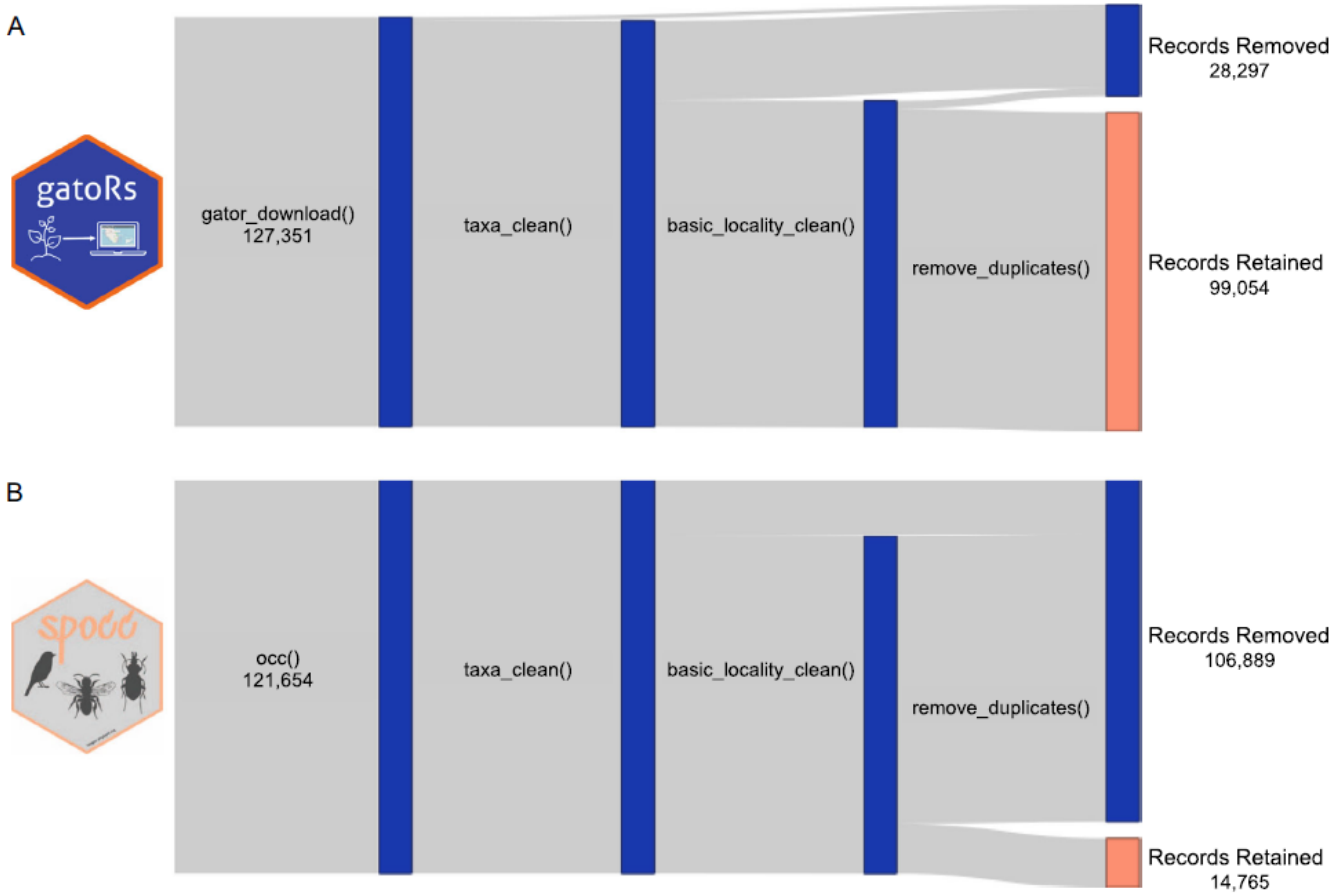


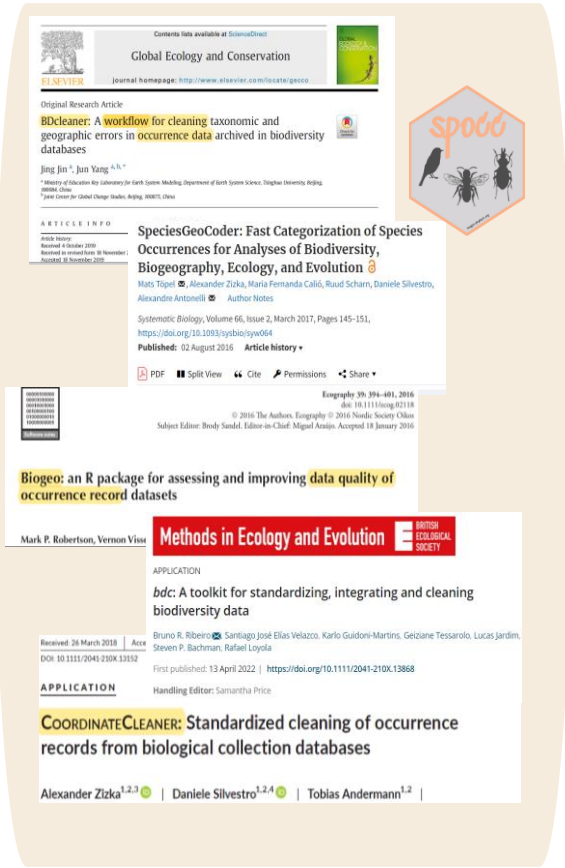
FIGURE 3 Sankey diagrams showing the sum of records returned for all 25 species after each cleaning step when using (A) gators_download() from gatoRs and (B) occ() from spocc with the limit set to 100,000. This Sankey diagram was generated using the networkD3 R package (Allaire et al., 2017) and was inspired by Panter et al. (2020) (see their Figure 3). The number of records after each processing step can be found in Appendix S6. The spocc logo was sourced from https://github.com/ropensci/spocc/blob/master/man/figures/logo.png.

Patten et al 2024. Applications in Plant Sciences
10.1002/aps3.11575

Methods in Ecology and Evolution
BRITISH ECOLOGICAL SOCIETY

OCCUR Shiny application: A user-friendly guide for curating species occurrence records

SCAN ME



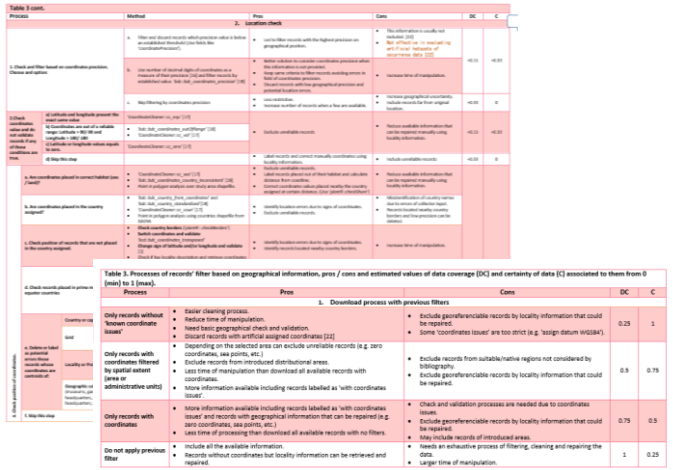


Table 3. Processes of records' filter based on geographical information, pros / cons and estimated values of data coverage (DC) and certainty of data (C) associated to them from 0 (min) to 1 (max).

Process	Pros	Cons	DC	C
1. Download process with previous filters				
Only records without 'known coordinate issues'	• Easier cleaning process. • Reduce time of manipulation. • Need basic geographical check and validation. • Discard records with artificial assigned coordinates [22]	• Exclude georeferenciable records by locality information that could be repaired. • Some 'coordinates issues' are too strict (e.g. 'assign datum WGS84').	0.25	1
Only records with coordinates filtered by spatial extent (area or administrative units)	• Depending on the selected area can exclude unreliable records (e.g. zero coordinates, sea points, etc.) • Exclude records from introduced distributional areas. • Less time of manipulation than download all available records with coordinates. • More information available including records labelled as 'with coordinates issues'	• Exclude records from suitable/native regions not considered by bibliography • Exclude georeferenciable records by locality information that could be repaired.	0.5	0.75
Only records with coordinates	• More information available including records labelled as 'with coordinates issues' and records with geographical information that can be repaired (e.g. zero coordinates, sea points, etc.) • Less time of processing than download all available records with no filters.	• Check and validation processes are needed due to coordinates issues. • Exclude georeferenciable records by locality information that could be repaired. • May include records of introduced areas.	0.75	0.5
Do not apply previous filter	• Include all the available information. • Records without coordinates but locality information can be retrieved and repaired.	• Needs an exhaustive process of filtering, cleaning and repairing the data. • Larger time of manipulation.	1	0.25

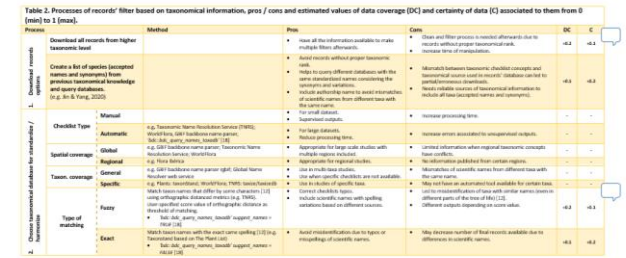


Table 2. Processes of records' filter based on taxonomical information, pros / cons and estimated values of data coverage (DC) and certainty of data (C) associated to them from 0 (min) to 1 (max).

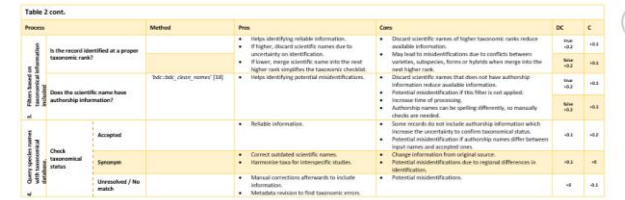
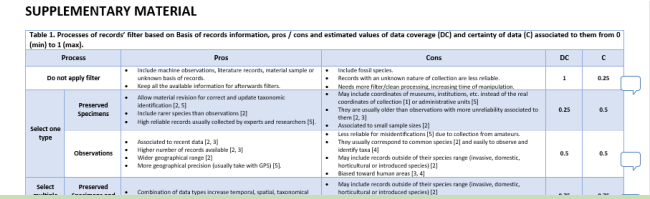


Table 2 cont.



SUPPLEMENTARY MATERIAL

Table 1. Processes of records' filter based on Basis of records information, pros / cons and estimated values of data coverage (DC) and certainty of data (C) associated to them from 0 (min) to 1 (max).

Process		Pros	Cons	DC	C
Do not apply filter		• Include machine observations, literature records, material sample or unknown basis of records. • Keep all the available information for afterwards filters.	• Include fossil species. • Records with an unknown nature of collection are less reliable. • Needs more filter/clean processing, increasing time of manipulation.	1	0.25
Select one type	Preserved Specimens	• Allow material revision for correct and update taxonomic identification [2, 5] • Include rarer species than observations [2] • High reliable records usually collected by experts and researchers [5].	• May include coordinates of museums, institutions, etc. instead of the real coordinates of collection [1] or administrative units [5] • They are usually older than observations with more unreliability associated to them [2, 3] • Associated to small sample sizes [2]	0.25	0.5
	Observations	• Associated to recent data [2, 3] • Higher number of records available [2, 3] • Wider geographical range [2] • More geographical precision (usually take with GPS) [5].	• Less reliable for misidentifications [5] due to collection from amateurs. • They usually correspond to common species [2] and easily to observe and identify taxa [4] • May include records outside of their species range (invasive, domestic, horticultural or introduced species) [2] • Biased toward human areas [3, 4]	0.5	0.5
Select multiple	Preserved Specimens and	• Combination of data types increase temporal, spatial, taxonomical	• May include records outside of their species range (invasive, domestic, horticultural or introduced species) [2]		

SCAN ME

OCCUR app

SOURCE

OCCUR app is a "step by step" guide that goes over 5 different modules to curate biodiversity data records. It was created to facilitate the process of filtering, cleaning and validating occurrence species records from data repositories. This interactive workflow will help the user in the selection of data records between all possibilities depending on their study case, considering their pros and cons. Each module will also display how data certainty and data coverage change when selecting different scenarios of the application of filtering and cleaning rules.

- Basis of Record
- Taxonomic
  - 1. Download records options
  - 2. Choose type of taxonomical source for standardize / harmonize
  - 3. Filters based on taxonomical information included
  - 4. Query species names with taxonomical database
- Geographic
- Temporal
- Duplicates
- Final Report
- References
- About

INSTRUCTIONS

1. Choose a module of the 5 available in the left panel.

Basis Of Record → Taxonomy → Geography → Time → Duplicates

2. Select between filters / steps in left-upper box (there are no previous selections marked).

3. Check the "Trade-off" table that will display with each selection in the right-upper box (left panel).

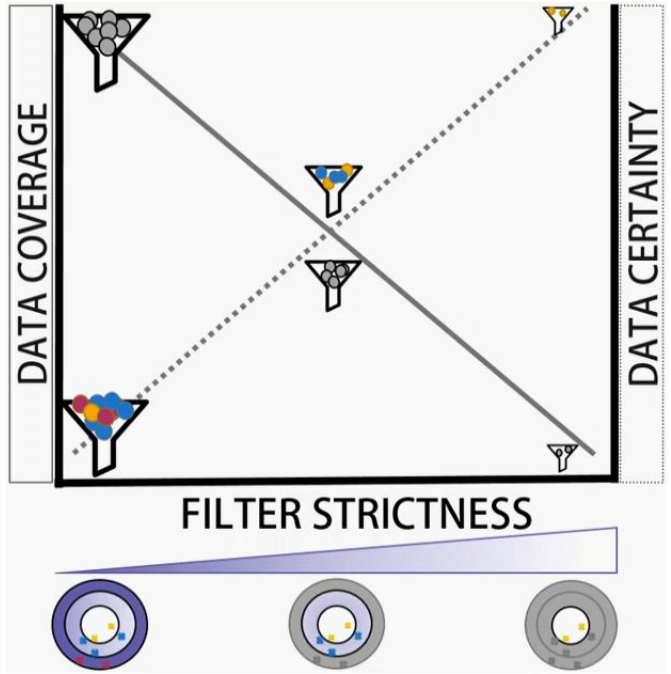
4. Check the "Methods" table that will display with each selection in the right-upper box (middle panel).

5. Check and copy the "R Code" table that will display with each selection in the right-upper box (right panel).

6. See the bibliography associated in the "References" panel.

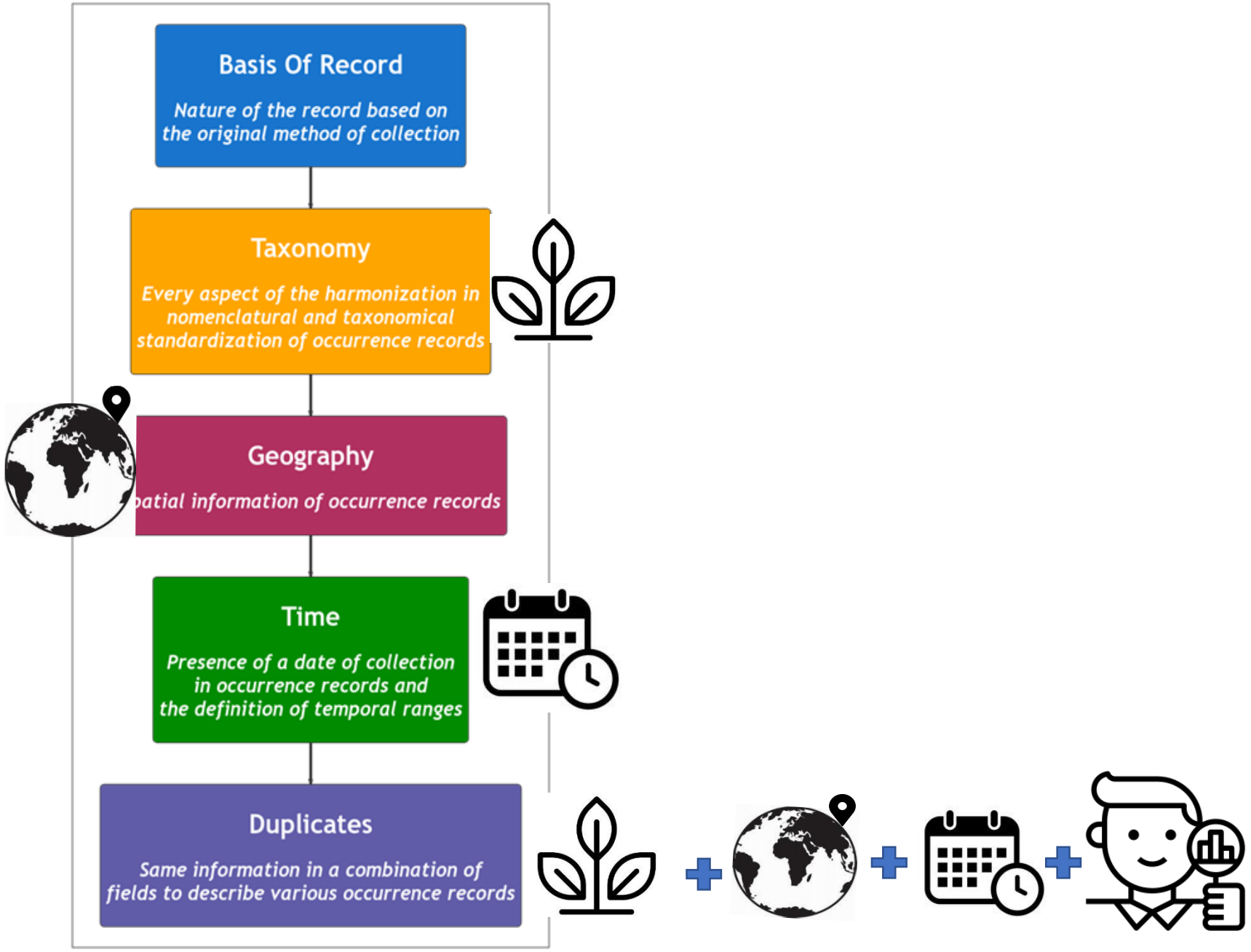
7. Check how certainty and data coverage varies with each selection in the left-bottom panel to make your final selection. Values goes from 0 (minimum certainty or data coverage available) to 1 (maximum certainty or data coverage available).

8. Download the final guide to process data and write the methods section based on the selected steps by module in the "Final report" tab.

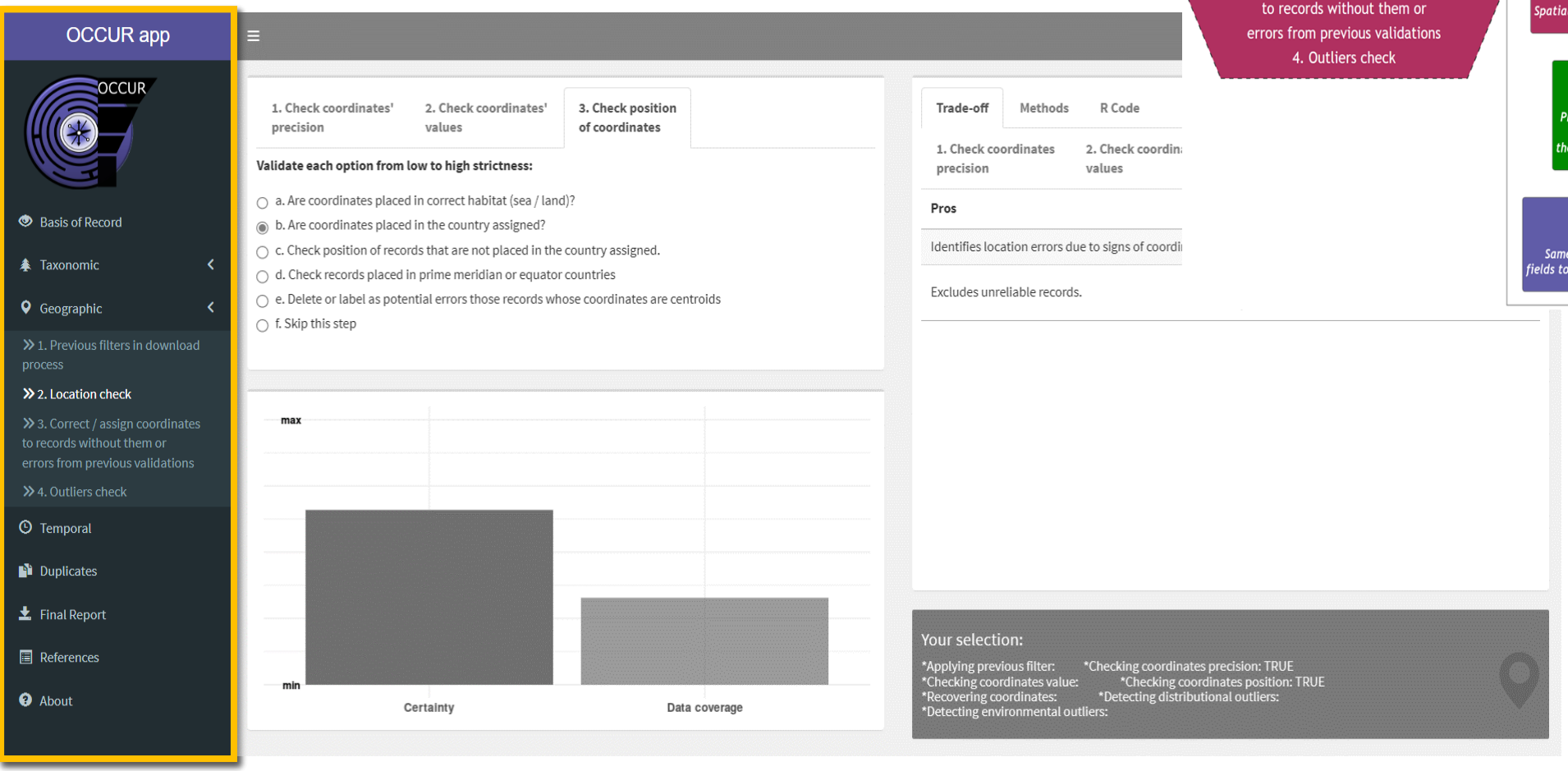


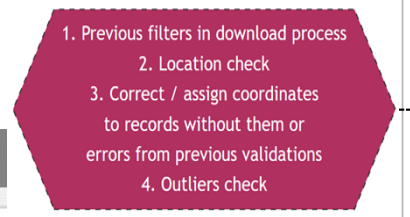
https://ecoinformatic.shinyapps.io/OCCUR/_w_c622ec2c/#shiny-tab-dashboard

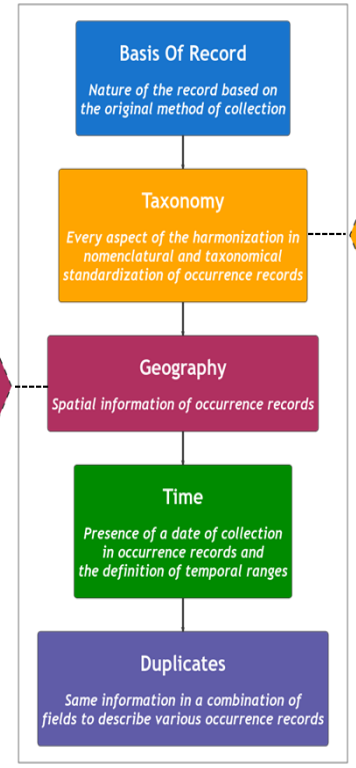
5 módulos de trabajo

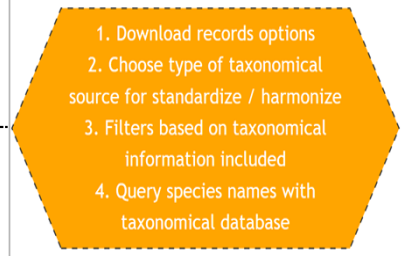


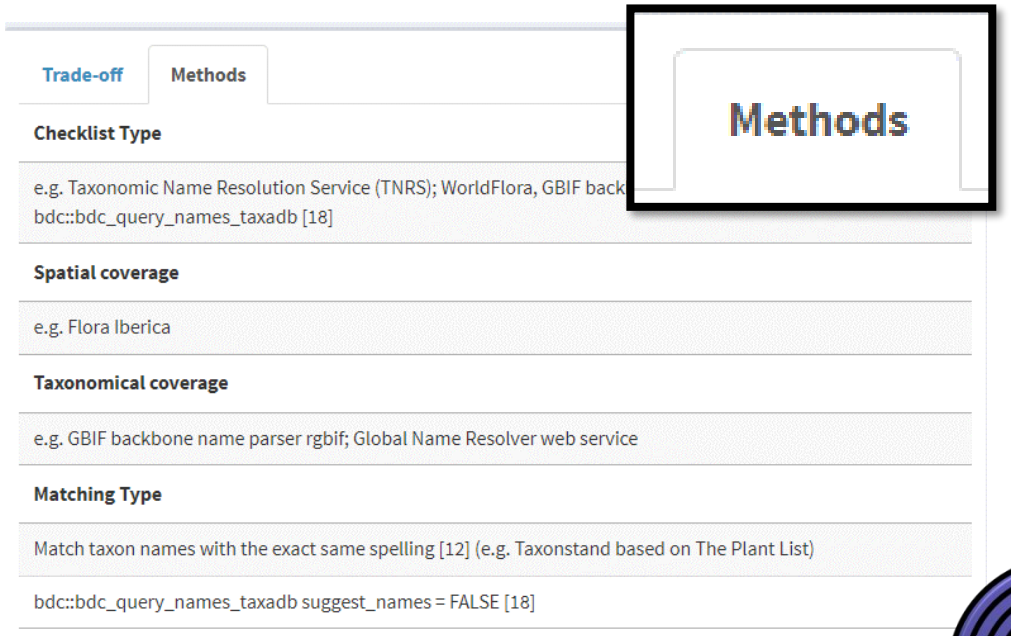
5 módulos de trabajo





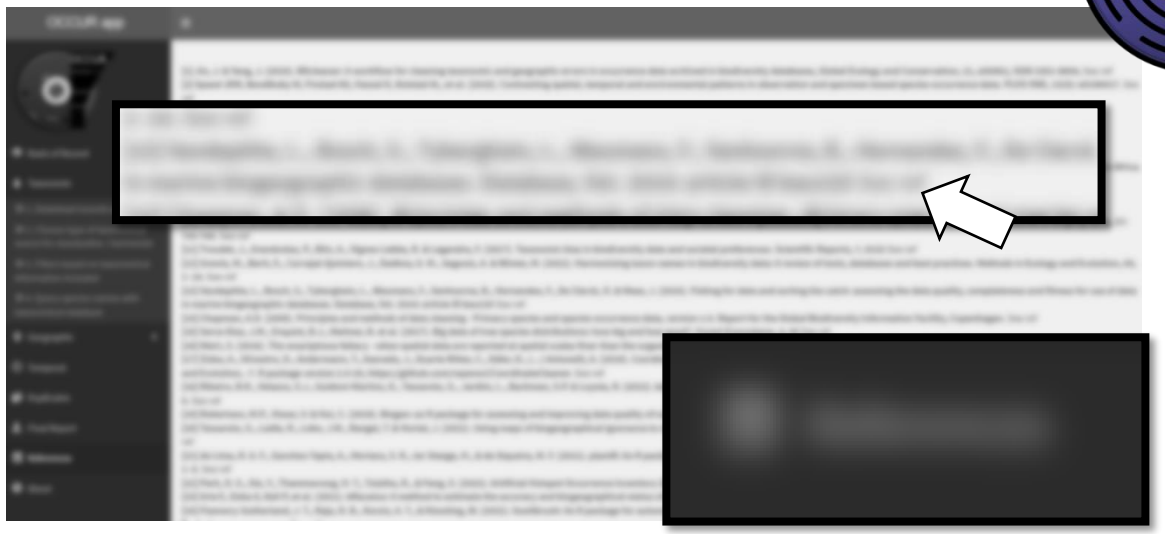


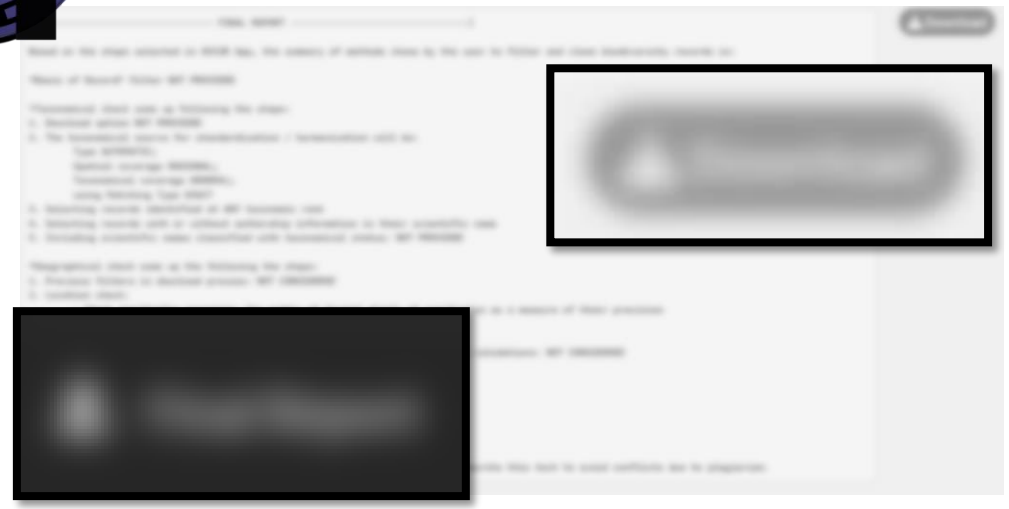


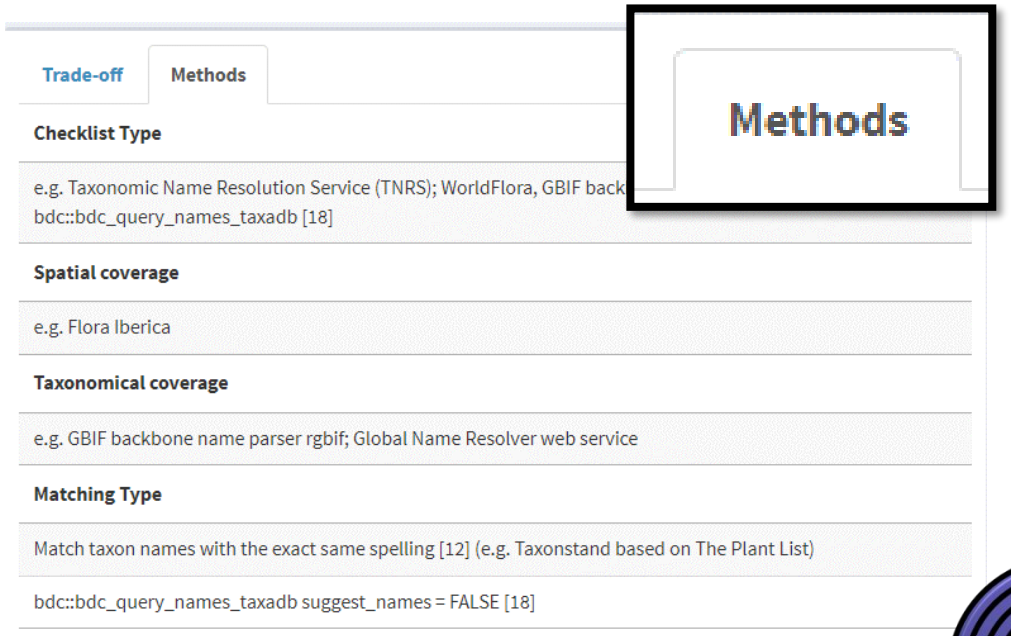


Methods

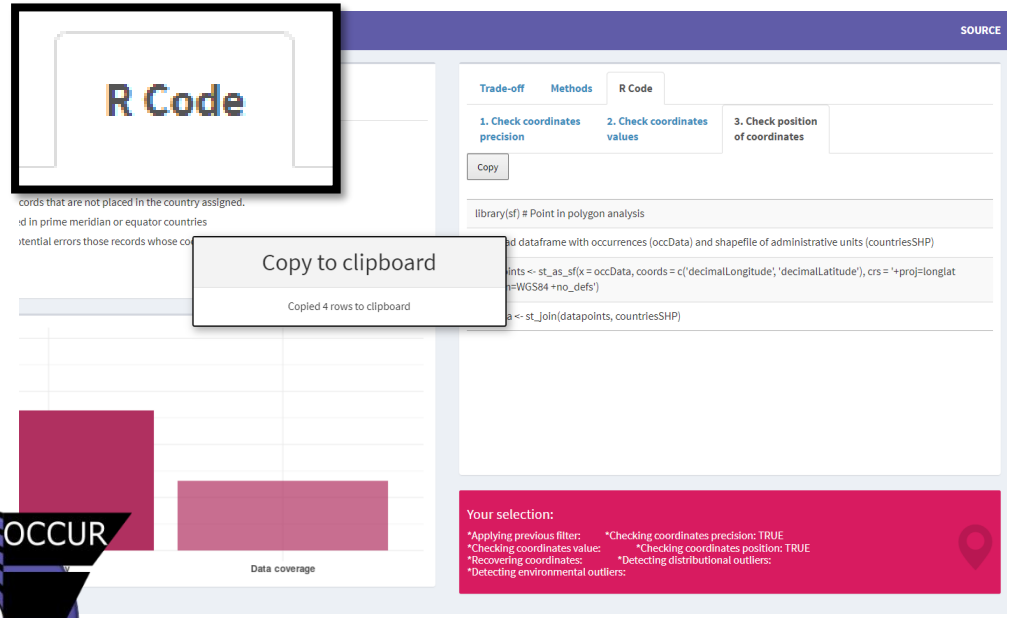
OCCUR





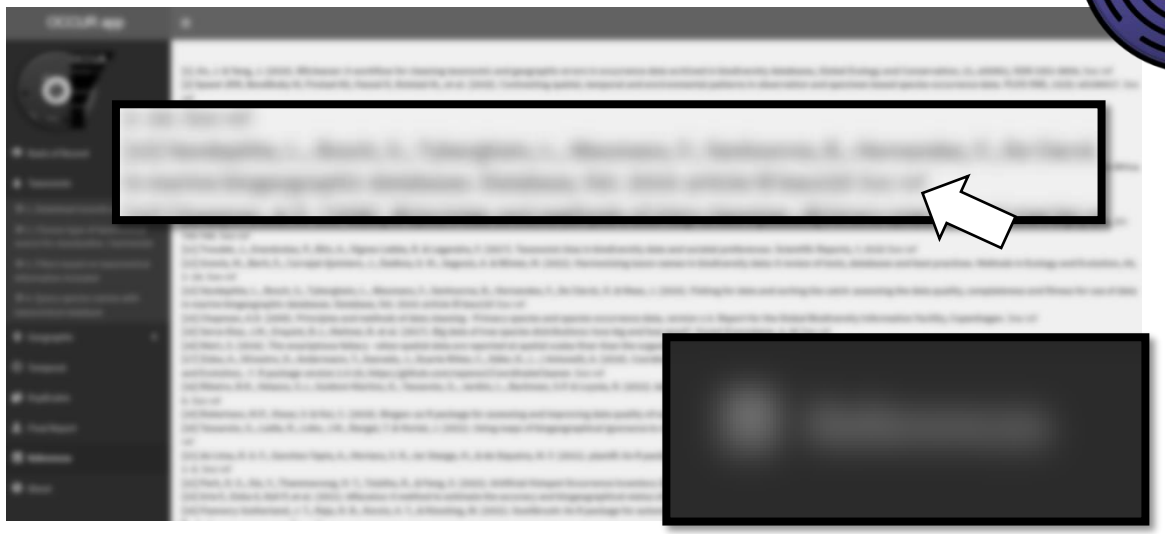


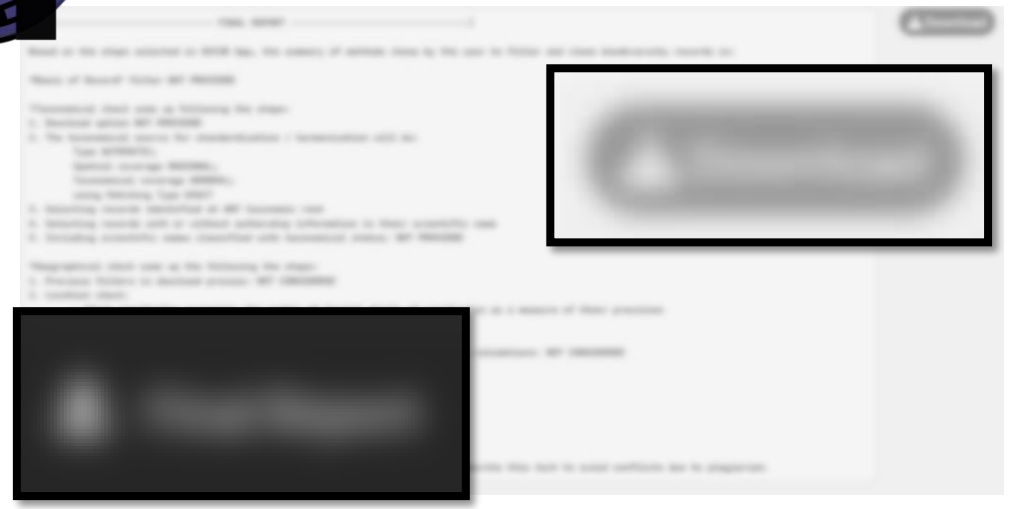
Methods

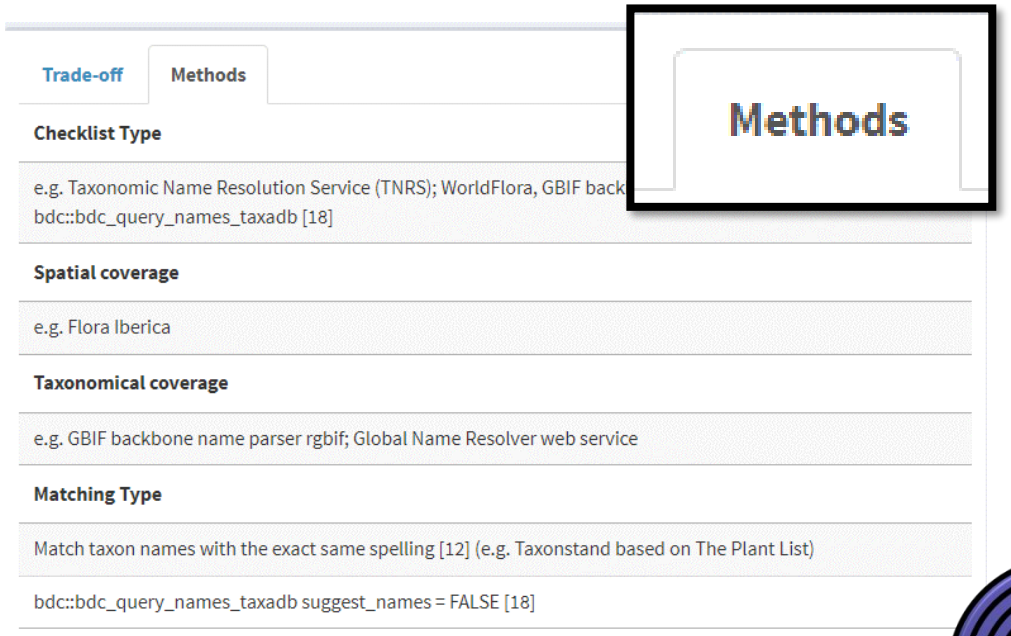


R Code

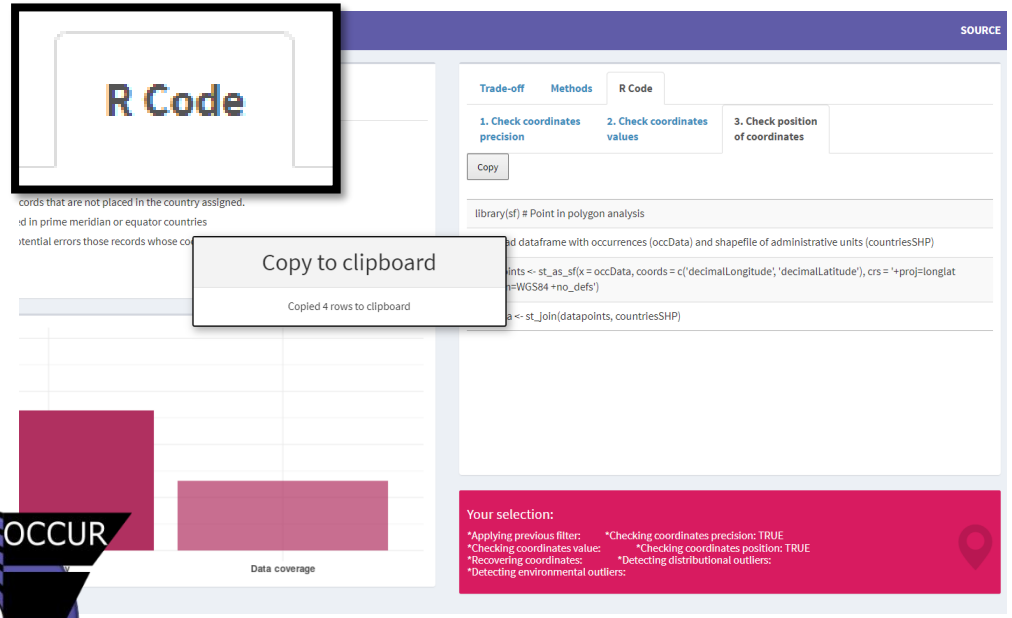
OCCUR







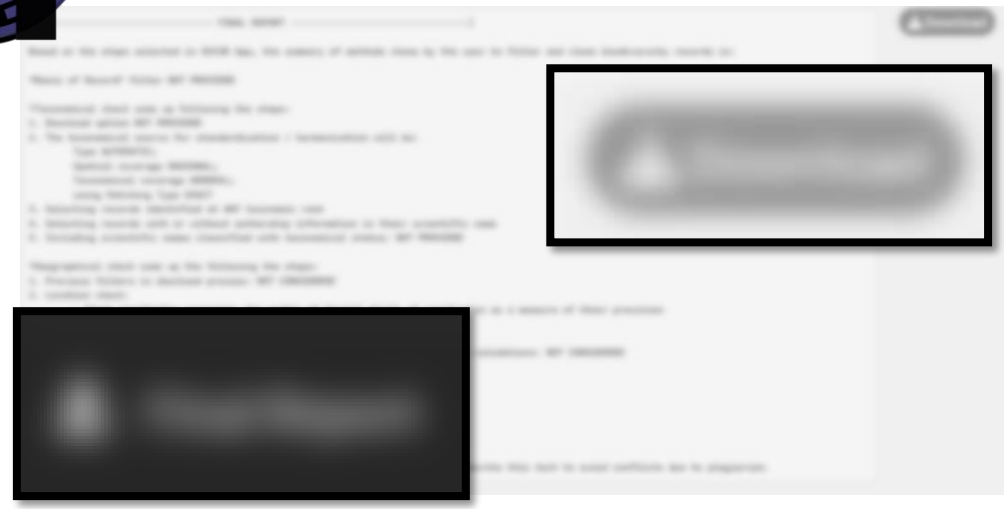
Methods

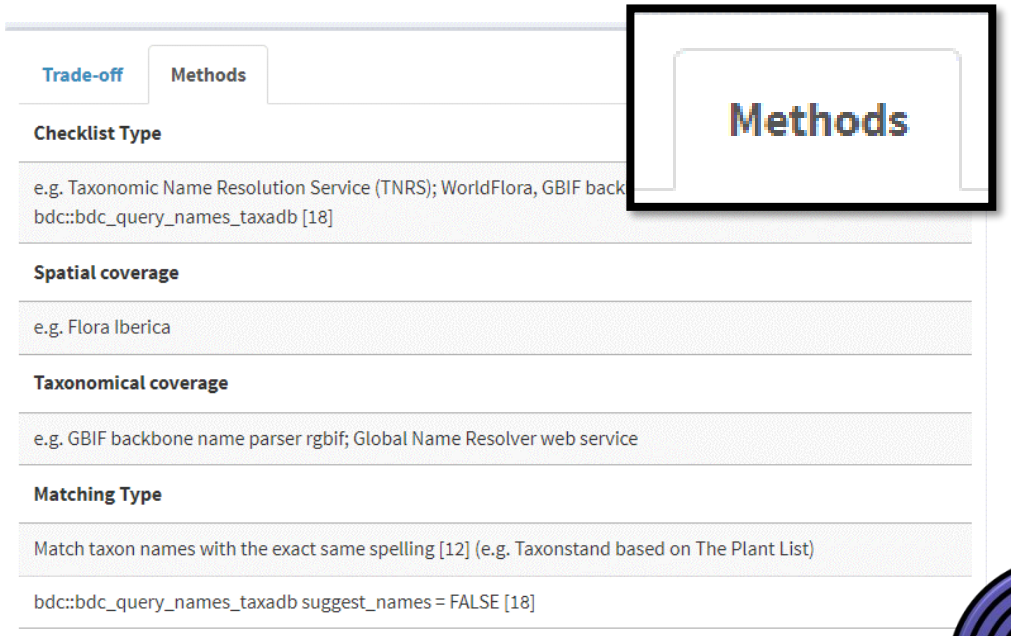


R Code

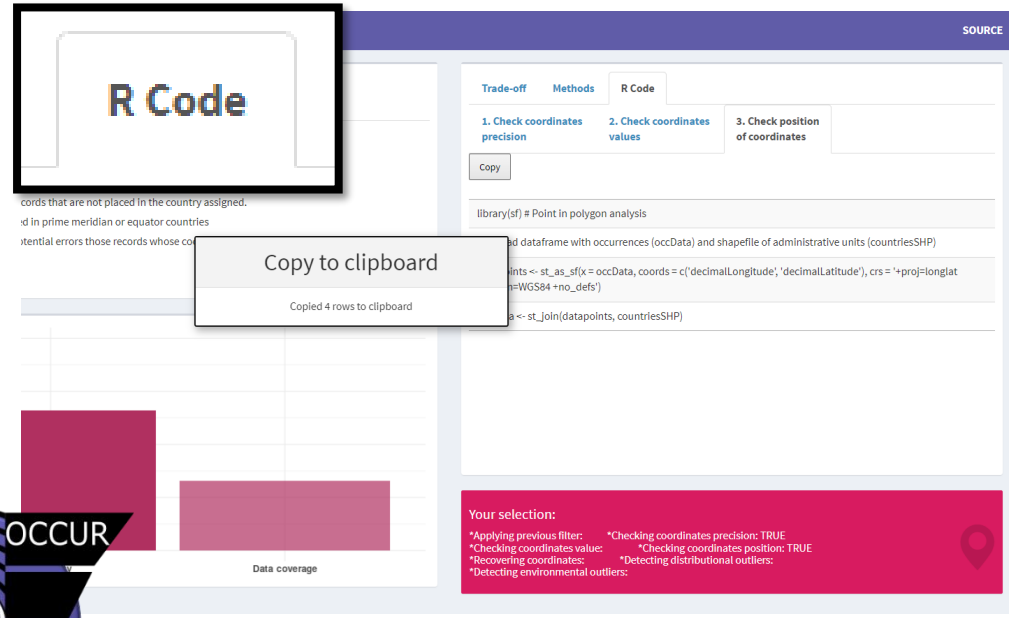
OCCUR

References





Methods



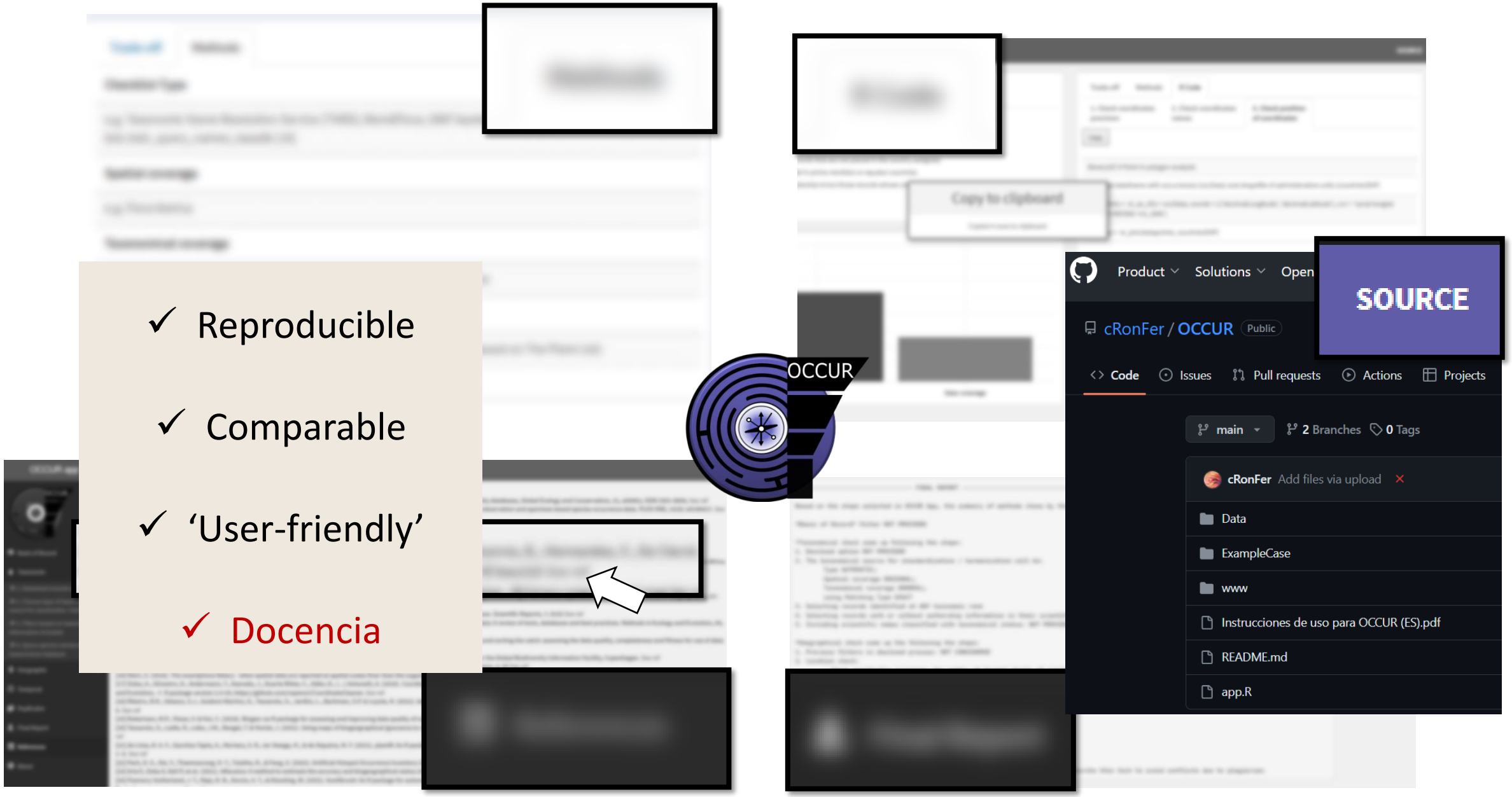
R Code

OCCUR

References

Download

Final Report


✓ Reproducible
✓ Comparable
✓ ‘User-friendly’
✓ Docencia
OCCUR
SOURCE
Product
Solutions
Open
cRonFer / OCCUR
Public
Code
Issues
Pull requests
Actions
Projects
main
2 Branches
0 Tags
cRonFer Add files via upload
Data
ExampleCase
www
Instrucciones de uso para OCCUR (ES).pdf
README.md
app.R

Caso de estudio:

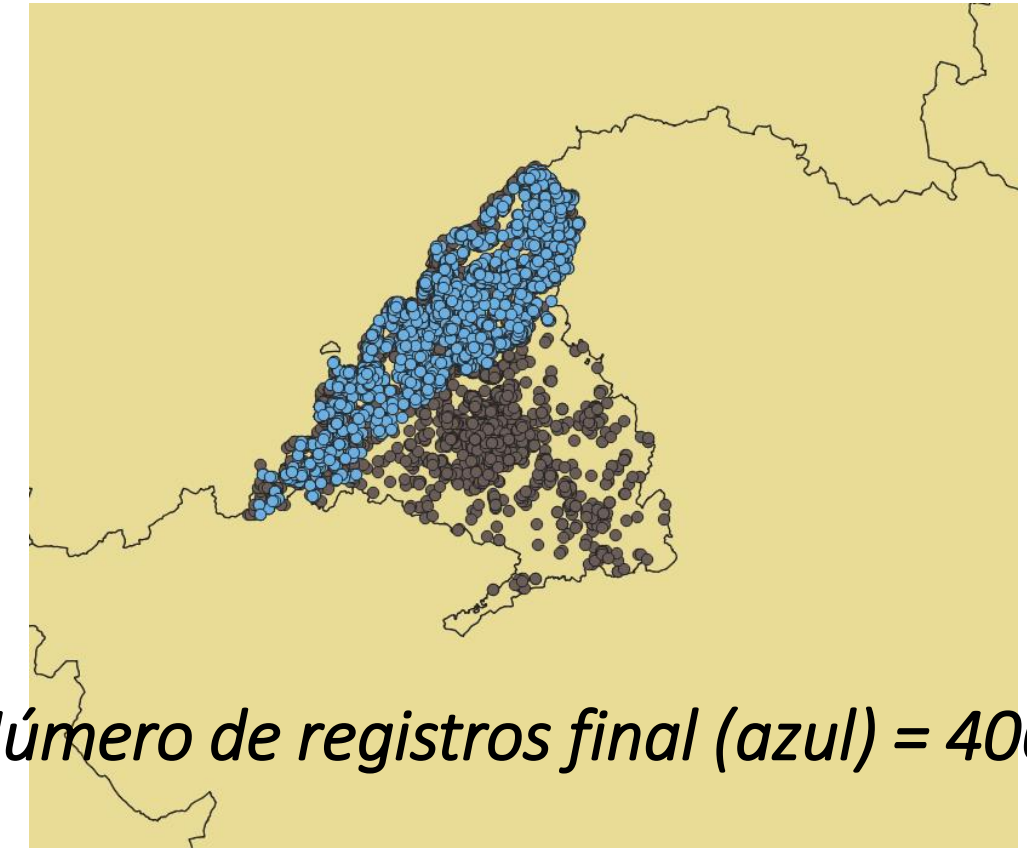
Caso de estudio:

1. *Elige un conjunto de datos de GBIF o usa uno propio*
2. *Elige alguna validaciones durante estos días y aplícalas a tus datos*
3. *Anota el número de registros iniciales y el final*
4. *Crea un mapa donde compares la distribución de registros inicial y tras validarlos*

Caso de estudio:

- Datos de registros de presencia de especies de la clase pinopsida desde 1980
- Descartadas subespecies y variedades o no identificadas a nivel de especie
- Sólo observaciones humanas
- Área Noroeste de la Comunidad de Madrid

Número de registros inicial = 9417



Número de registros final (azul) = 4066