



# *Otras validaciones*

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

*Cristina Ronquillo*

*Ayudante investigación*

*[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)*

mncn 25  museo  
nacional de  
ciencias  
naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

# 1 Taller GBIF.ES: Mejora de la calidad de datos de biodiversidad

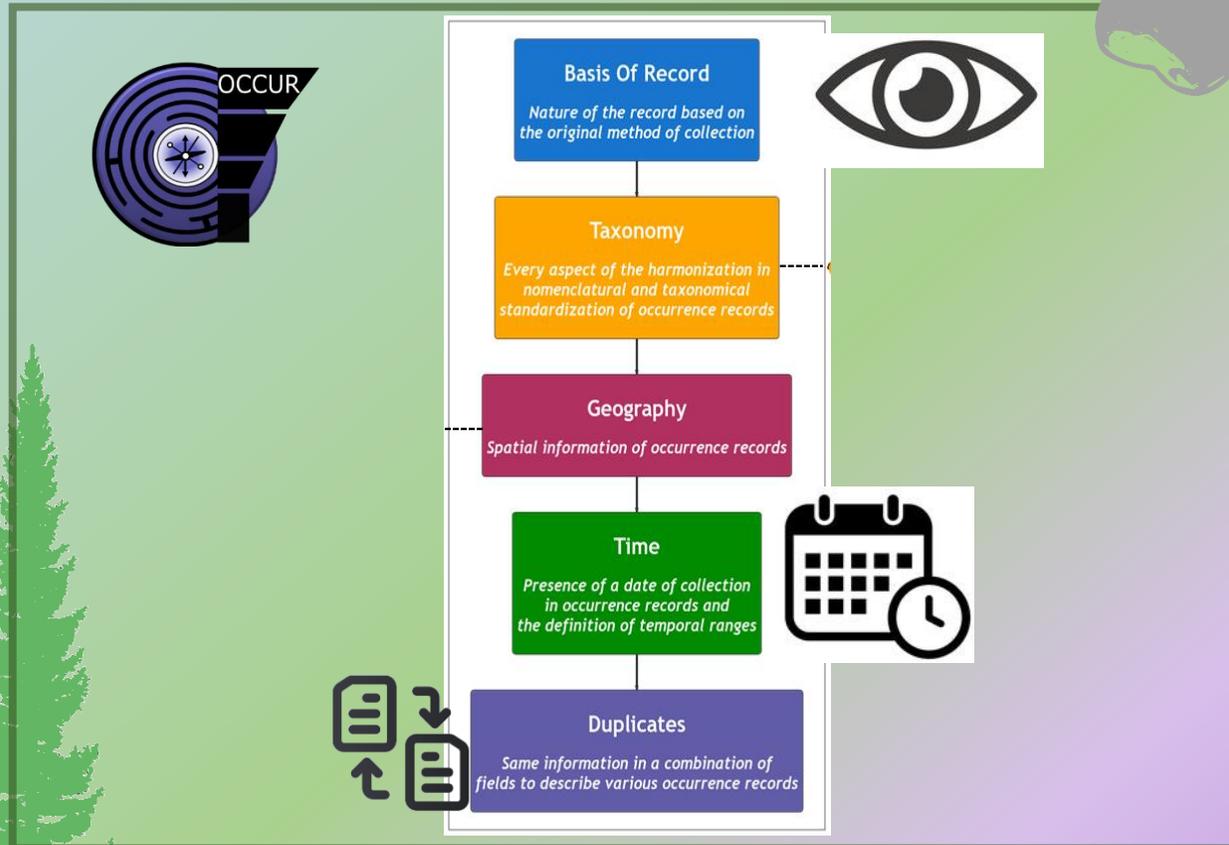
**Cristina Ronquillo**

**Ayudante investigación**

**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

mncn 25 1771  
2021 museo  
nacional de  
ciencias  
naturales

**CSIC**  
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



# 1 Taller GBIF.ES: Mejora de la calidad de datos de biodiversidad

**Cristina Ronquillo**

**Ayudante investigación**

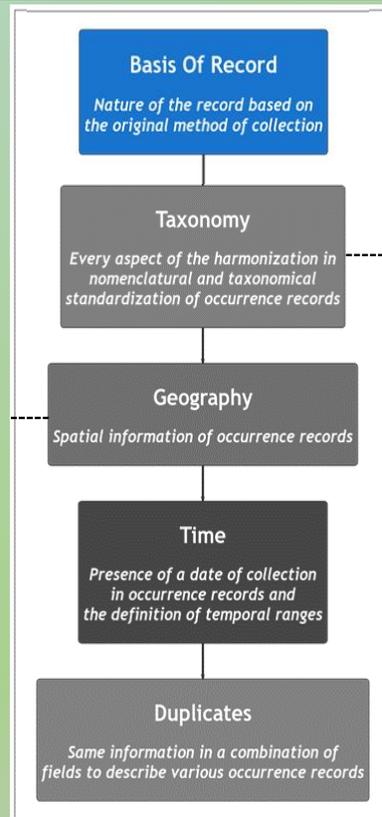
**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

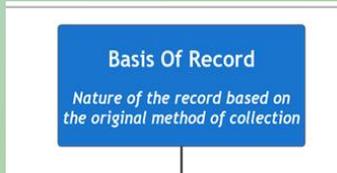
mncn 25 1771  
2021 museo  
nacional de  
ciencias  
naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS





## Objetivos

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



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

**Cristina Ronquillo**

**Ayudante investigación**

**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

mncn 25 1771  
2021 museo  
nacional de  
ciencias  
naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



# Módulo 1: DwC::basisOfRecord

Home

Basis of Record

Taxonomic <

Geographic <

Temporal

Duplicates

Final Report

References

Basis of record describes how records were originally collected.

Choose filter:

- a. No filter
- b. Preserved specimens
- c. Observations
- d. Preserved specimens + Observations

Category	Value (approx. %)
Certainty	75
Data coverage	75

Trade-off

Methods

Pros	Cons
Associated to recent data [2][3]	Less reliable for misidentifications [5] due to collection from amateurs.
Higher number of records available [2][3]	They usually correspond to common species [2] and easily to observe and identify taxa [4]
Wider geographical range [2]	May include records outside of their species range (invasive, domestic, horticultural or introduced species) [2]
More geographical precision (usually take with GPS) [5].	Biased toward human areas [3][4]

Your final selection:

Observations



# Módulo 1: DwC::basisOfRecord

*'(Human) observations'*

*An output of a human observation process.*



*'Preserved specimens'*

*A specimen that has been preserved.*





# Módulo 1: DwC::basisOfRecord

*'(Human) observations'*  
An output of a human observation process.



- **LivingSpecimen** (A specimen that is alive)
- **FossilSpecimen** (A preserved specimen that is a fossil)
- **MaterialCitation** (A reference to or citation of one, a part of, or multiple specimens in scholarly publications)
- **MachineObservation** (An output of a machine observation process)

*'Preserved specimens'*  
A specimen that has been preserved.



## Metadatos básicos

Todos los campos de metadatos de esta página son obligatorios. Tenga en cuenta que para cada contacto debe proporcionar al menos un apellido, un cargo o una organización.

Título del Recurso [Plantas medicinales endémicas de Colombia](#)

### Metadatos Básicos

Documente todas las propiedades obligatorias en la sección de Metadatos Básicos y luego continúe documentando los metadatos en las demás secciones que apliquen a su recurso. Entre más metadatos proporcione, mayor será la oportunidad de que el recurso sea encontrado, reutilizado por otros investigadores y citado.

Título\*  
Plantas medicinales endémicas de Colombia

Organización publicadora\*  
Test Organisation Jenkins

Tipo\*  
Lista de chequeo

Idioma de los Metadatos\*  
Español

Frecuencia de actualización\*  
Desconocido

Subtipo  
Inventario Temático

Idioma de los Datos\*  
Español

Licencia de los datos\*  
Creative Commons Attribution (CC-BY) 4.0

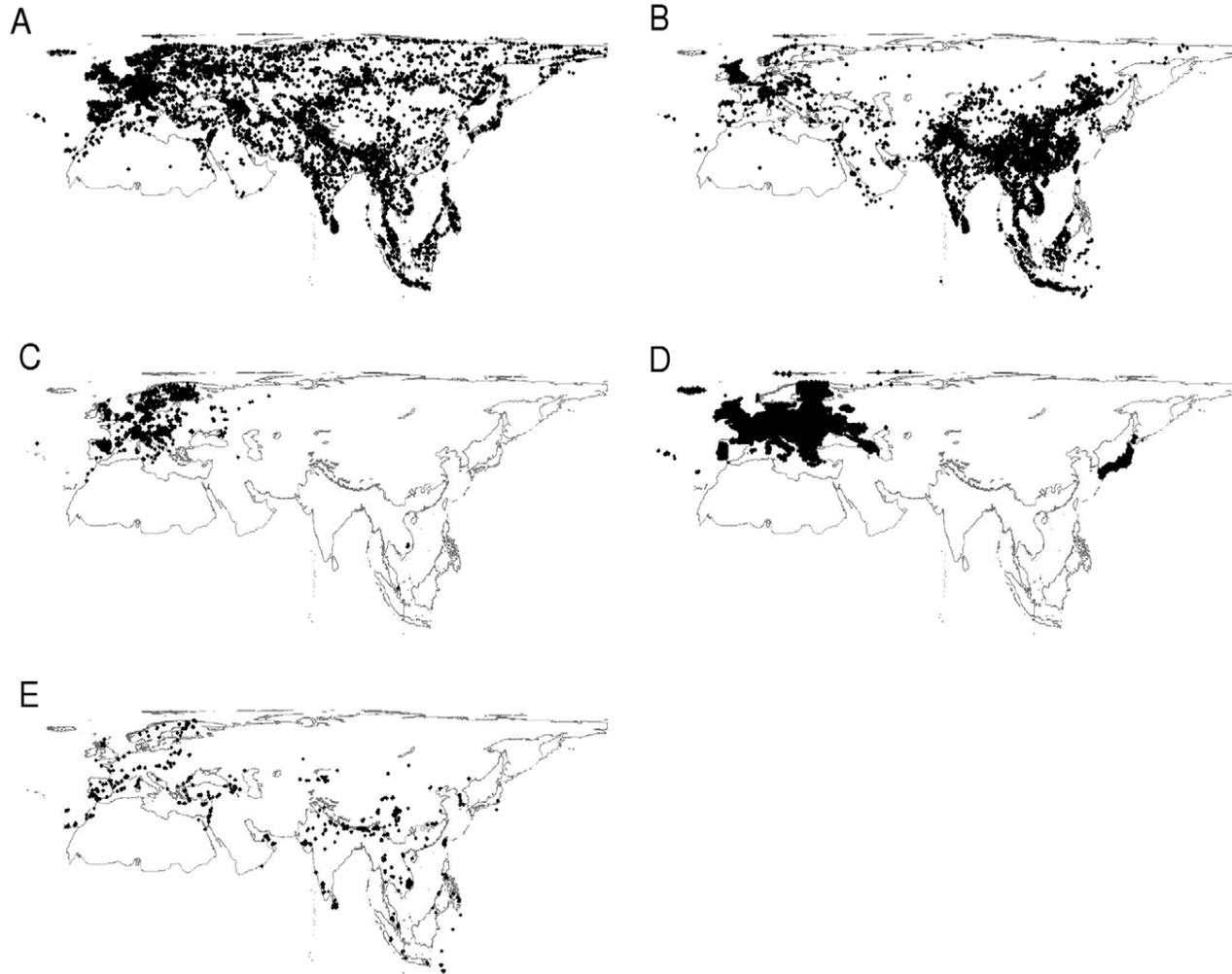
This work is licensed under a [Creative Commons Attribution \(CC-BY\) 4.0 License](#).

### Sección

- [Metadatos Básicos](#)
- [Cobertura Geográfica](#)
- [Cobertura Taxonómica](#)
- [Cobertura Temporal](#)
- [Palabras Clave](#)
- [Partes Asociadas](#)
- [Datos del Proyecto](#)
- [Métodos de Muestreo](#)
- [Referencias](#)
- [Datos de la Colección](#)
- [Enlaces Externos](#)
- [Metadatos Adicionales](#)



# Módulo 1: DwC::basisOfRecord



Boakes, et al (2010) PLoS Biol 8(6): e1000385.  
10.1371/journal.pbio.1000385



# Módulo 1: DwC::basisOfRecord

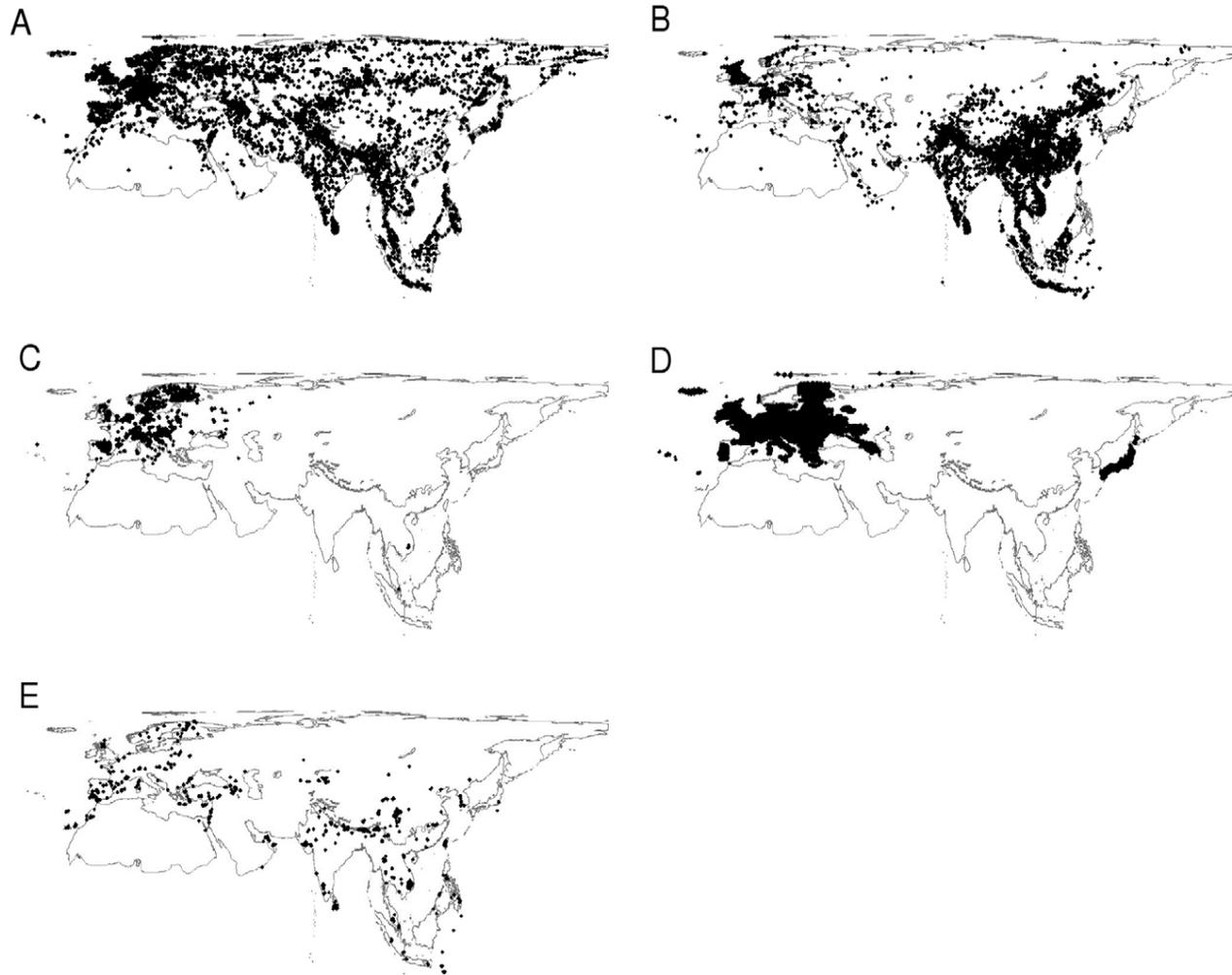
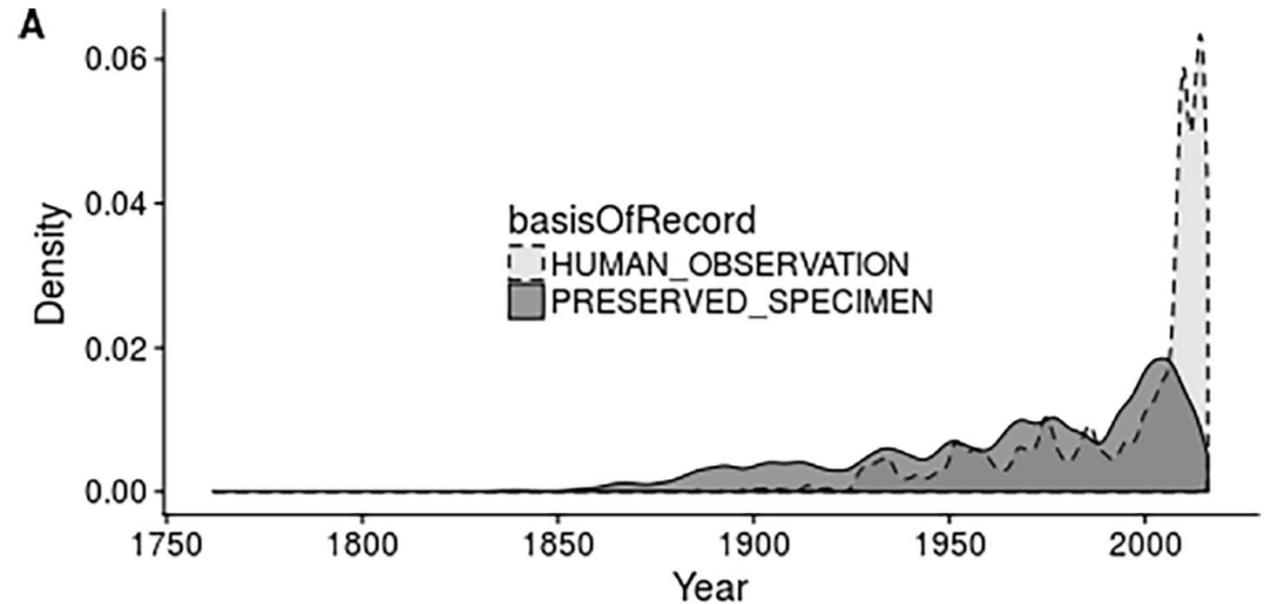
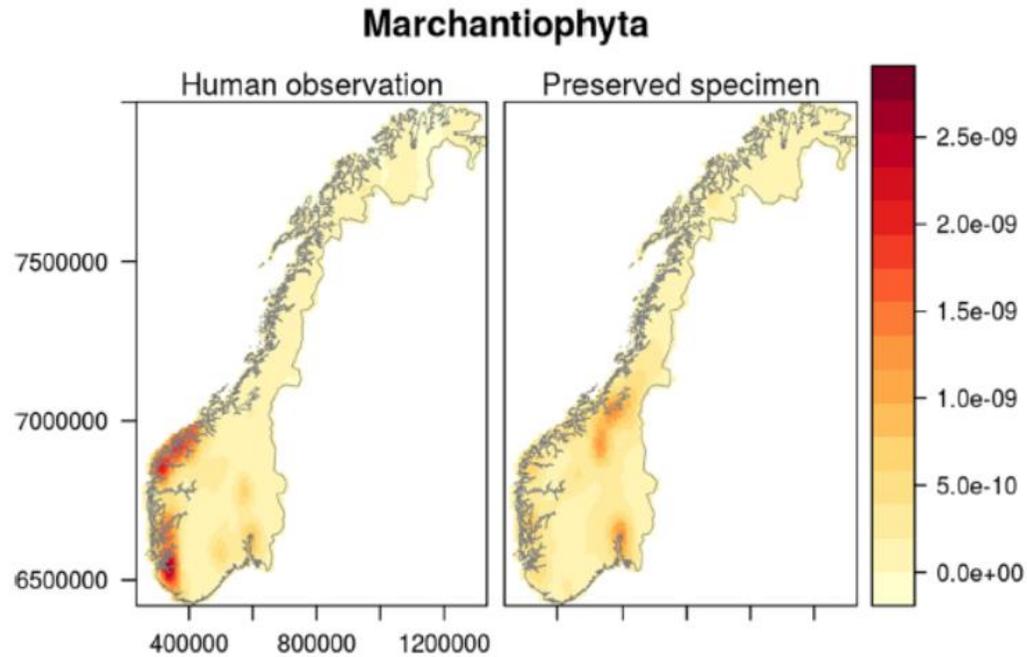


Figure 2. The spatial distribution of records from different sources. A) museums, B) literature, C) ringing, D) atlas, and E) website trip reports.

Boakes, et al (2010) PLoS Biol 8(6): e1000385.  
10.1371/journal.pbio.1000385



# Módulo 1: DwC::basisOfRecord



Speed et al (2018) PLoS ONE 13(4).  
<https://doi.org/10.1371/journal.pone.0196417>



# Módulo 1: DwC::basisOfRecord

*'(Human) observations'*



JUEGO DE DATOS DE REGISTROS BIOLÓGICOS | REGISTRADO

## Epiphytic mosses from the northwest Iberian quadrant (Spain)

Publicado por [Museo Nacional de Ciencias Naturales \(CSIC\)](#)  
G. Medina N • Ronquillo Ferrero C

CONJUNTO DE DATOS ESTADÍSTICAS ACTIVIDAD DESCARGA **1479 REGISTROS** 68 CITAS

PORQUE ERES UN CONTACTO CONFIABLE

HISTÓRICO IMPORTANDO REGISTROS

The dataset comprises a subset of the information recorded in Medina, N.G.; Mazimpaka, V.; Hortal, J. & Lara (2015). Catálogo de los briófitos epífitos que crecen en bosques de quercíneas del cuadrante noroccidental ibérico', F. Bol. Soc. Esp. Briol. 44-45: 1-30. It contains 1479 occurrences of 78 epiphytic mosses species recorded from the northwest Iberian sector. The main purpose of the study design was to contribute to reduce the knowledge gaps on the distribution of Iberian epiphytic mosses. All the occurrences were obtained by a systematic survey.

Fecha de publicación: 20 de marzo de 2020  
Última modificación de metadatos: 19 de julio de 2021  
Alojado por: GBIF-Spain  
Licencia: CC BY-NC 4.0  
[Cómo citar](#) DOI: 10.15470/rqv6jb

1479 Registros de presencia    100% Con coincidencia de taxón    100% Con coordenadas    100% Con año

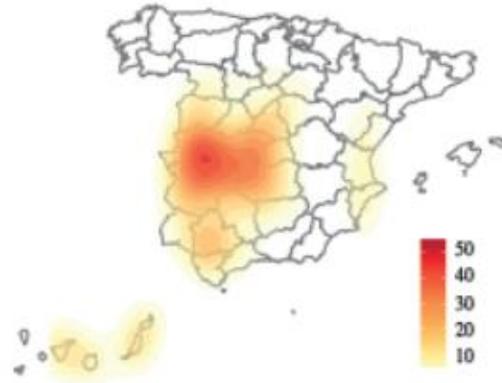
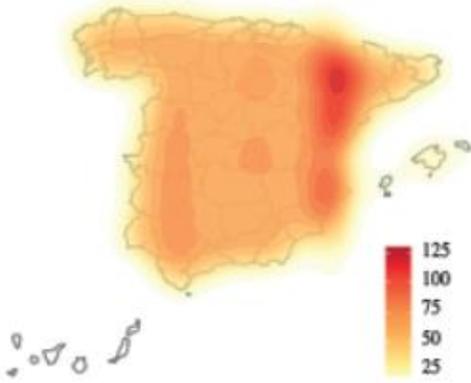
1479 REGISTROS GEOREFERENCIADOS



*Passer domesticus*

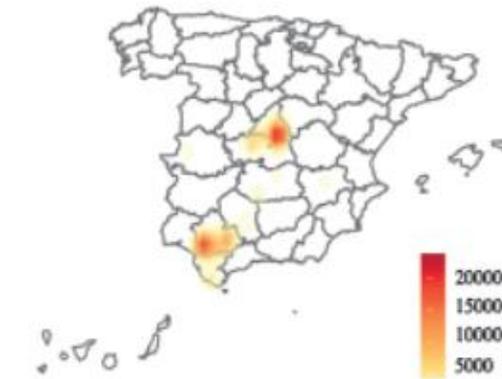
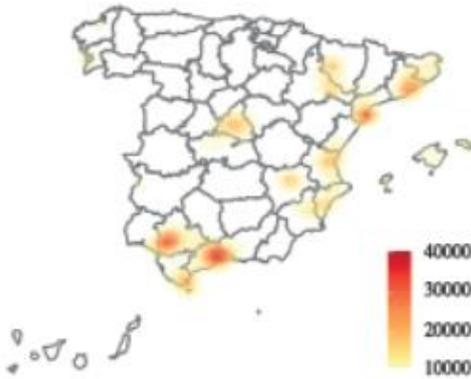
*Passer hispaniolensis*

IEET



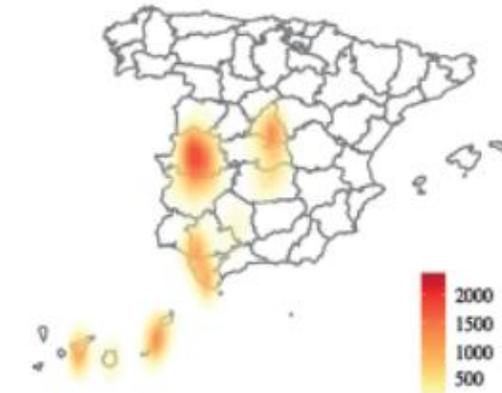
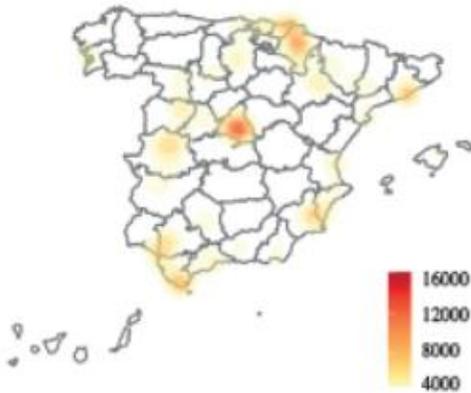
**Inventario Español de Especies Terrestres [celdas UTM]**

SEO



**Registros de la Sociedad Española de Ornitología (investigadores + anilladores amateurs)**

EBIRD



**Proyecto de Ciencia Ciudadana**

Galván et al. 2021

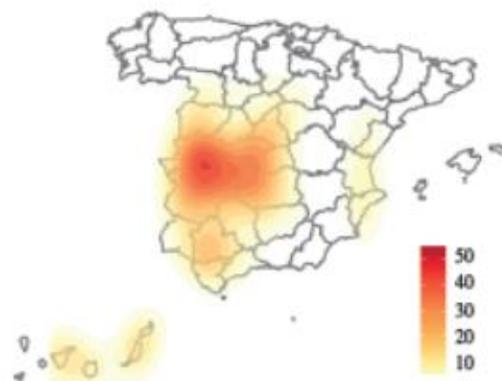
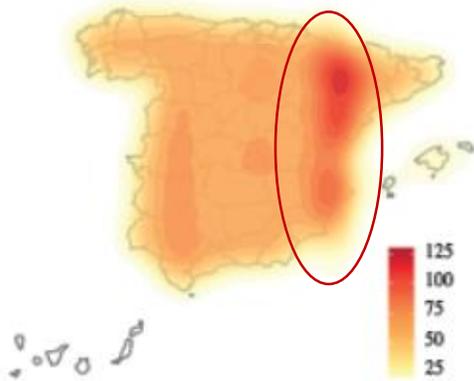
<https://doi.org/10.13157/arla.69.1.2022.ra6>



*Passer domesticus*

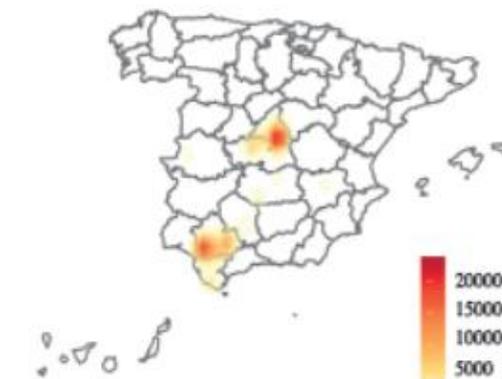
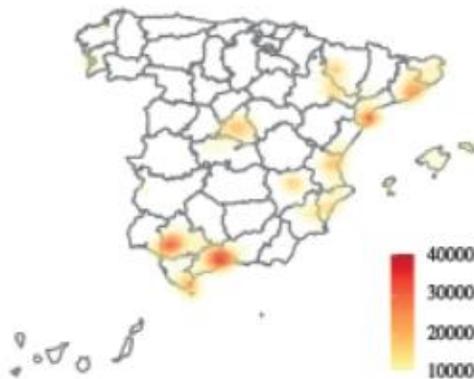
*Passer hispaniolensis*

IEET



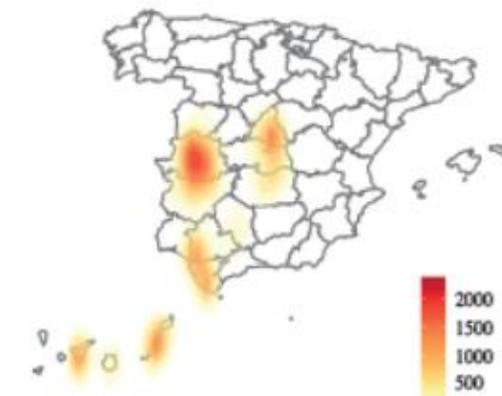
Inventario Español de Especies Terrestres [celdas UTM]

SEO



Registros de la Sociedad Española de Ornitología (investigadores + anilladores amateurs)

EBIRD



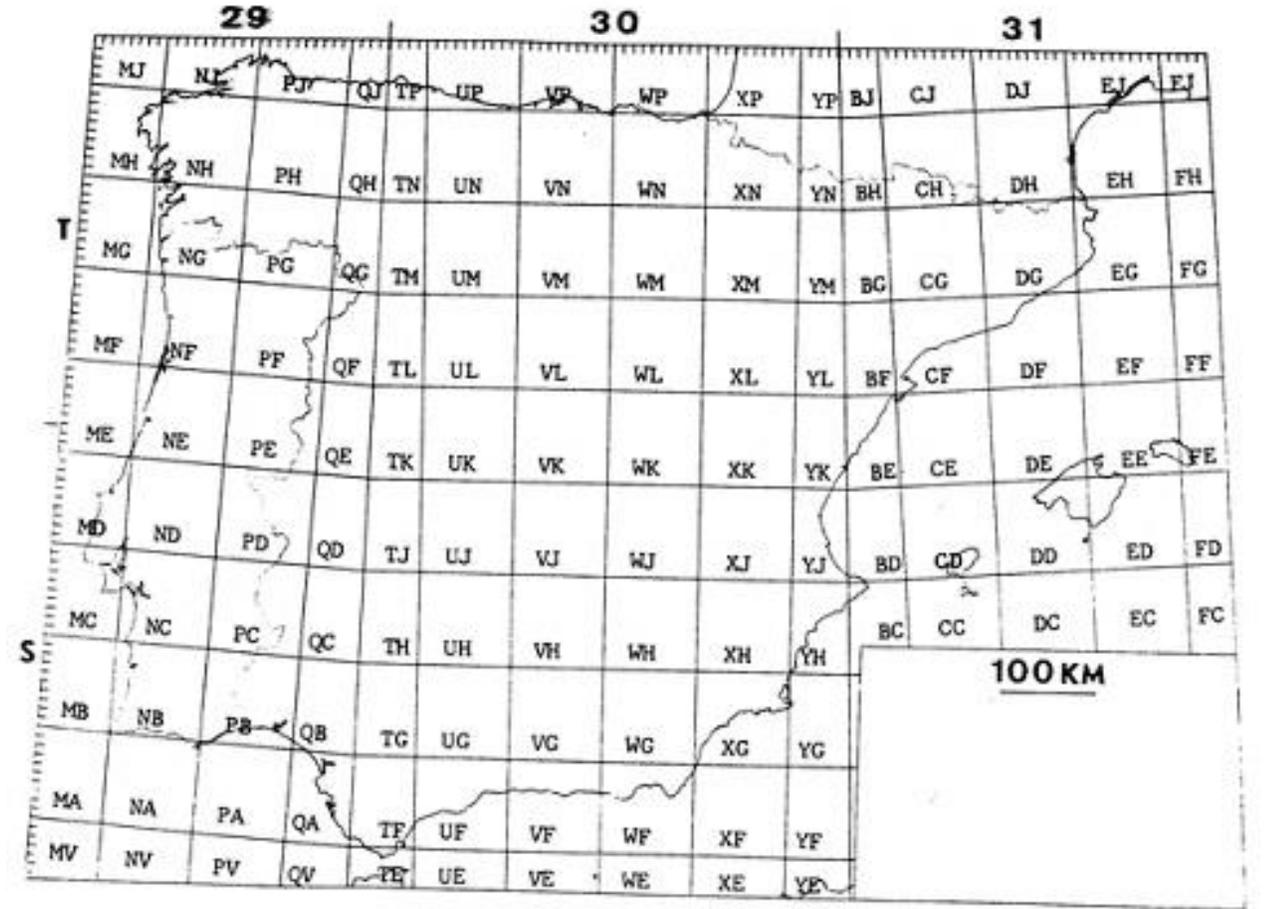
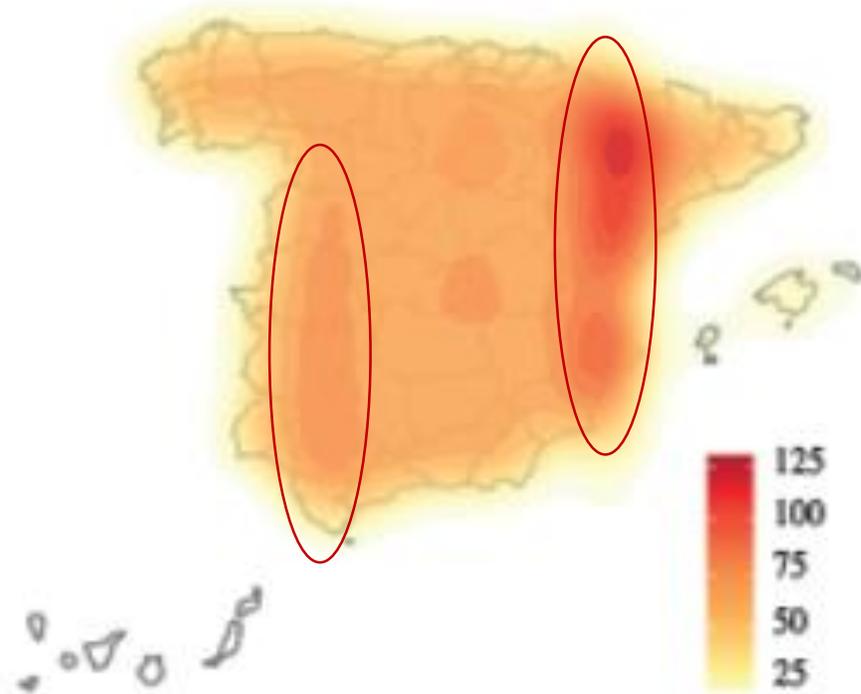
Proyecto de Ciencia Ciudadana

Galván et al. 2021  
<https://doi.org/10.13157/arla.69.1.2022.ra6>



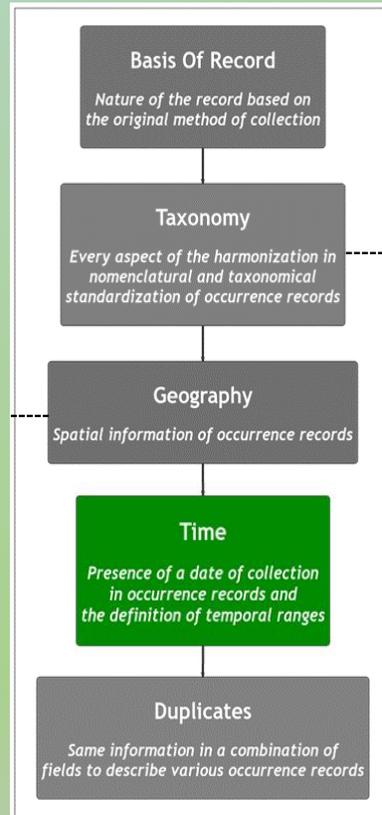
# *Passer domesticus*

IEET



Galván et al. 2021  
<https://doi.org/10.13157/arla.69.1.2022.ra6>

# 1 Taller GBIF.ES: Mejora de la calidad de datos de biodiversidad



**Cristina Ronquillo**

**Ayudante investigación**

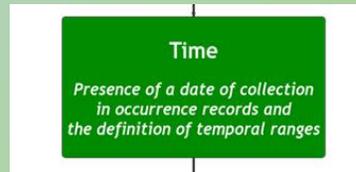
**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

mncn 25 1771 2021 museo nacional de ciencias naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



## Objetivos

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



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

**Cristina Ronquillo**

**Ayudante investigación**

**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

mncn 25 1771  
2021 museo  
nacional de  
ciencias  
naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



# Módulo 4: Tiempo

Home

Basis of Record

Taxonomic <

Geographic <

**Temporal**

Duplicates

Final Report

References

**Filters associated with 'when' was data recorded.**

Choose between:

a. Do not apply temporal filter

b. Apply temporal filter

Set temporal range of time

TRUE

FALSE

Choose level of temporal information

a. Complete date of collection

b. At least year of collection

Trade-off **Methods**

Range Level

Pros	Cons
Highest precision of record<92>s temporal coverage.	Many records are not linked with the day and/or month of collection.
Help to detect duplicate records by complete date of collection.	Format date can include DD/MM or MM/DD generating unreliable information.
Allow seasonal-temporal analyses of records.	

Category	Value
Certainty	High
Data coverage	Medium

**Your final selection:**

Data within temporal range  
Using Year of collection



# Módulo 4: Tiempo

¿Tenemos año de colecta?

¿Hay información del día y mes de recogida del dato?

Por favor ingrese los metadatos de los periodos de tiempo cubiertos por el recurso. Primero, seleccione el tipo de cobertura temporal, luego diligencie los campos que aparecerán.

Tipo de Cobertura Temporal [\[Eliminar esta\(e\) cobertura temporal\]](#)

Rango de Fechas

Fecha Inicial  Fecha Final

[Añadir nuevo\(a\) cobertura temporal](#)

Guardar Cancelar

## PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13  
 20130227 2013.02.27 27.02.13 27-02-13  
 27.2.13 2013.II.27. 27/2-13 2013.158904109  
 MMXIII-II-XXVII MMXIII <sup>LVII</sup>/<sub>CCCLXV</sub> 1330300800  
 ((3+3)×(111+1)-1)×3/3-1/3<sup>3</sup> 2013  
 10/1101/1101 02/27/20/13 <sup>2 3 4</sup>/<sub>5 6 7 8</sub> 



Ed Hansberry - MVP  @ehansalytics · 11 ago.

Optimist: The glass is 1/2 full

Pessimist: The glass is 1/2 empty

Excel: The glass is January 2



# Módulo 4: Tiempo





# Módulo 4: Tiempo

## Nivel de información en la fecha de colecta

Especie	Fecha	Año	Mes
Ursus arctos	19/11/2011	2011	11
Ursus arctos	11/2011	2011	11
Ursus arctos	11/2011	2011	11
Ursus arctos	2011	2011	



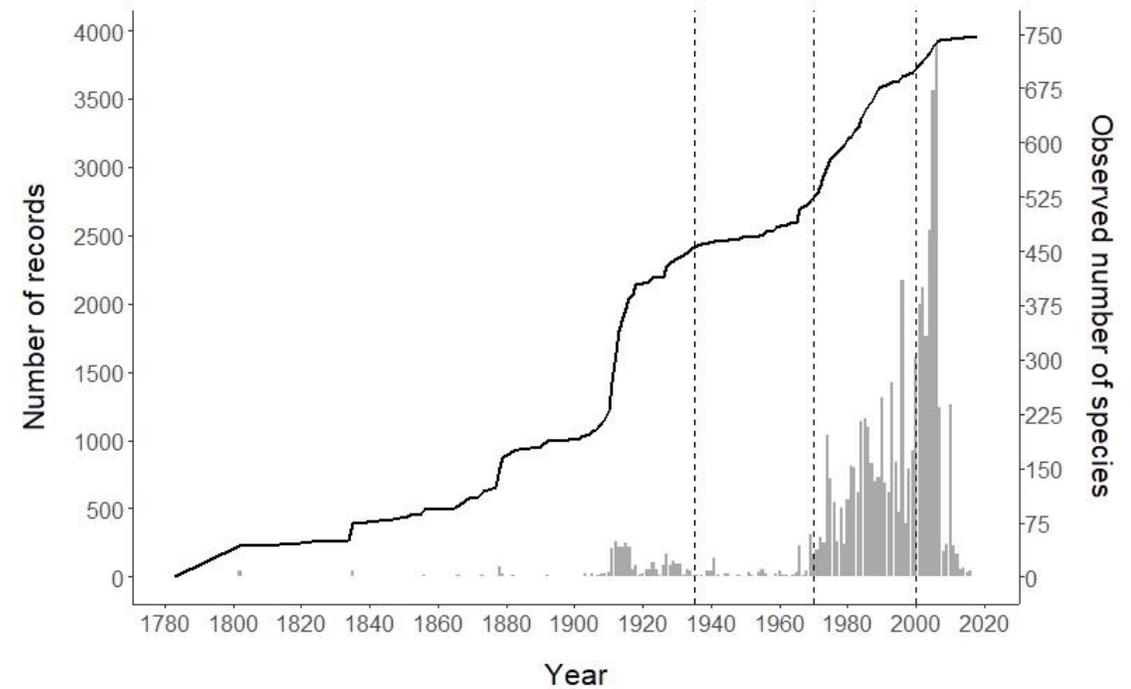


# Módulo 4: Tiempo

## Nivel de información en la fecha de colecta

Especie	Fecha	Año	Mes
Ursus arctos	19/11/2011	2011	11
Ursus arctos	11/2011	2011	11
Ursus arctos	11/2011	2011	11
Ursus arctos	2011	2011	

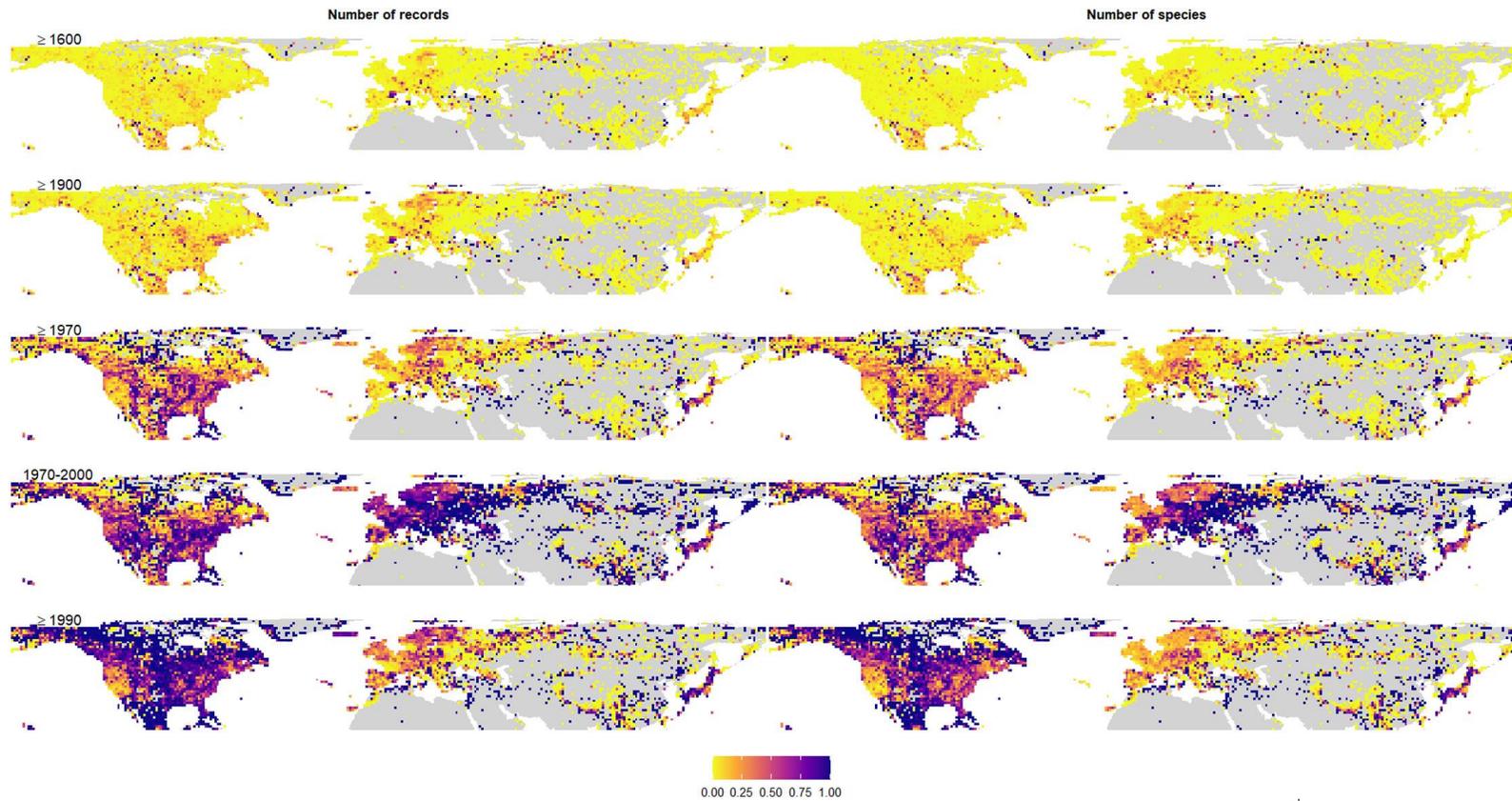
## Rango temporal



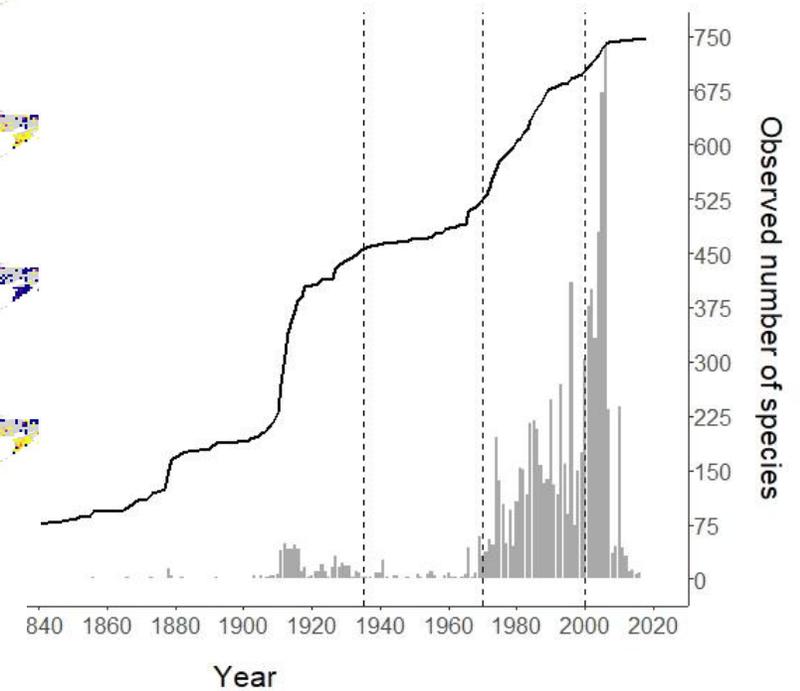
Ronquillo et al. (2020). BDJ



# Módulo 4: Tiempo



## Rango temporal

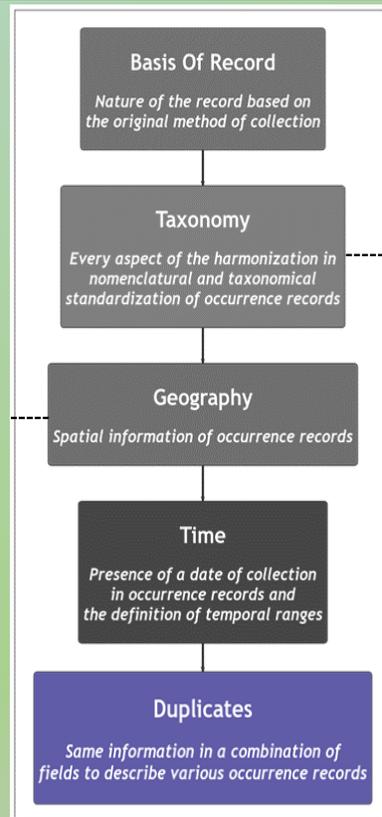


**FIGURE 5** Geographical distribution of the proportions of records (left) and observed species (right) of mosses discarded after filtering by different temporal ranges, at 100 km cell width resolution (NSIDC EASE-Grid 2.0 Global projection).

Ronquillo et al. (2020). BDI

Ronquillo et al. (2023). Ecol&Evol

# 1 Taller GBIF.ES: Mejora de la calidad de datos de biodiversidad



**Cristina Ronquillo**

**Ayudante investigación**

**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

mncn 25 1771  
2021 museo  
nacional de  
ciencias  
naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

**Duplicates**  
*Same information in a combination of  
fields to describe various occurrence records*

## Objetivos

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

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

**Cristina Ronquillo**

**Ayudante investigación**

**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

mncn 25 1771  
2021 museo  
nacional de  
ciencias  
naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



# Módulo 5: Duplicados

- Home
- Basis of Record
- Taxonomic
- Geographic
- Temporal
- Duplicates**
- Final Report
- References

Select the combination of fields to detect duplicate records besides species name

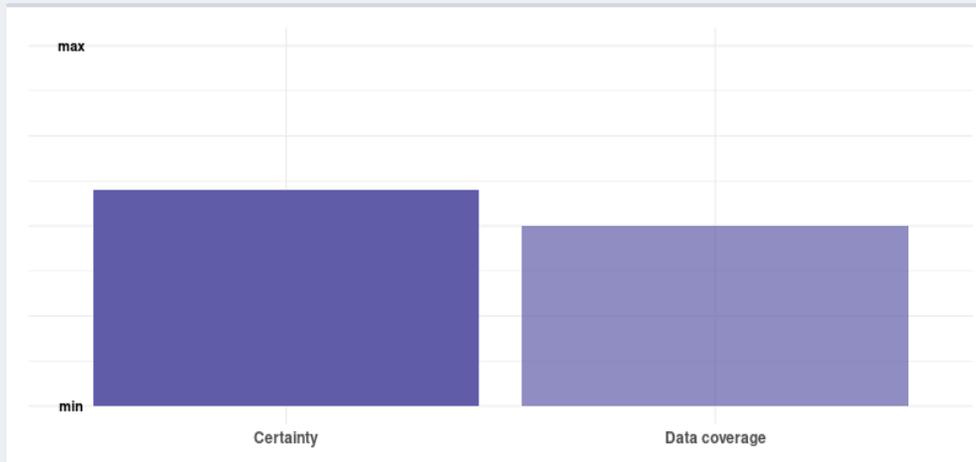
**a. Include position information as:**

- Geographic grid cell
- Coordinates + Buffer or rounded coordinates
- Coordinates (latitude and longitude)

**b. Include time of collection as:**

- Date - dd/mm/yyyy
- Year
- Discard temporal information for detecting duplicates

**c. Include recorder's name**



**Trade-off** | **Methods**

---

a. Cell: <91>biogeo::duplicatesexclude<92> [19]

Create a grid of the study area adapted to cell size needed [11]

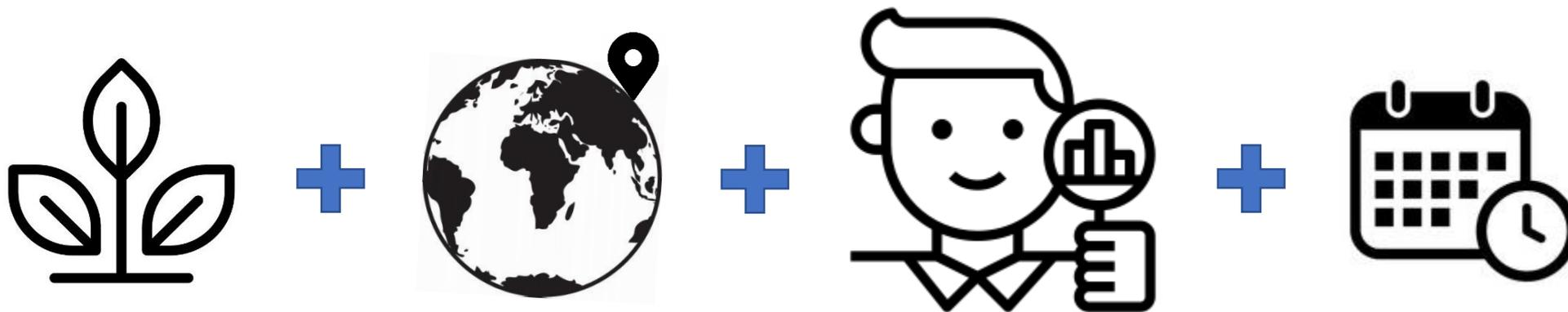
---

b. Add date to <91>CoordinateCleaner::cc\_dupl<92> [17]

**Your final selection:**  
Identify duplicates as the combination of:  
Species Cell Date



# Módulo 5: Duplicados



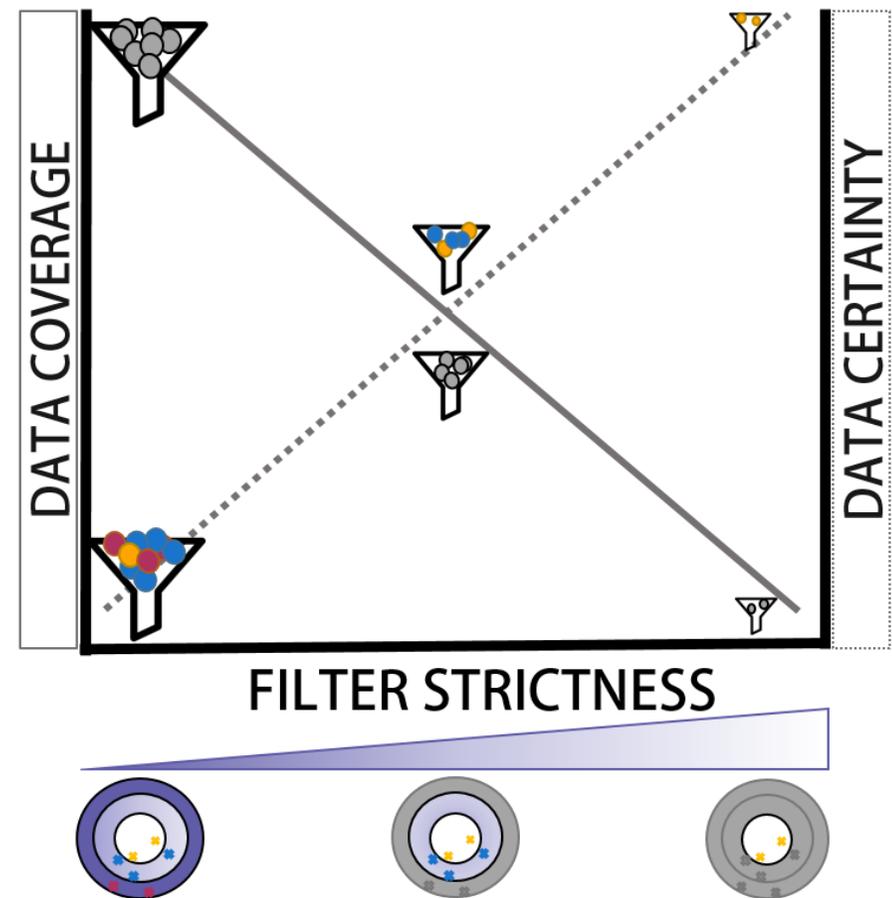


# Módulo 5: Duplicados

## Pregunta:

¿En qué escenario tendremos más registros disponibles?

- A) Si detecto duplicados por localización + fecha + recolector
- B) Si detecto duplicados por localización + fecha





# Módulo 5: Duplicados

## Pregunta:

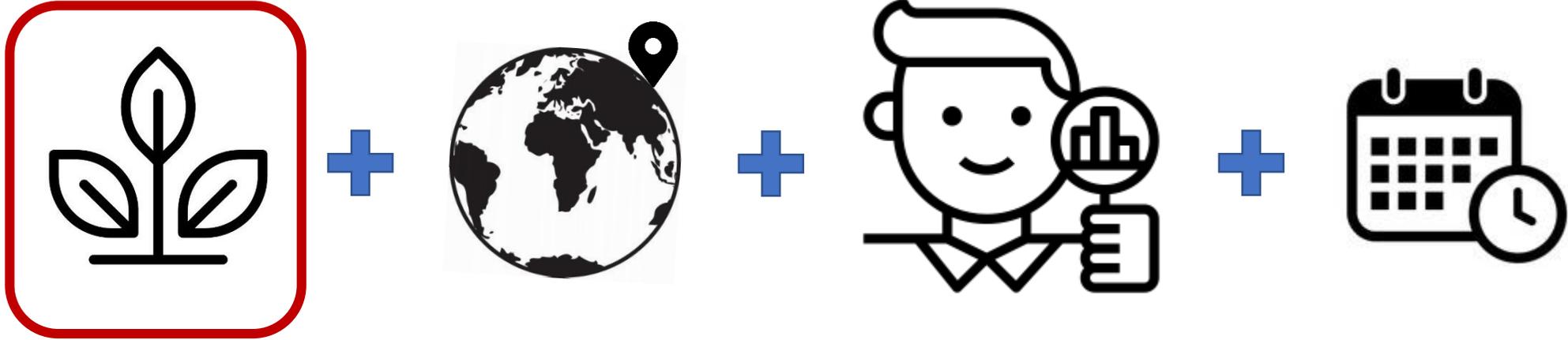
¿En qué escenario tendremos más registros disponibles?

- A) Si detecto duplicados por localización + fecha + recolector
- B) Si detecto duplicados por localización + fecha





# Módulo 5: Duplicados



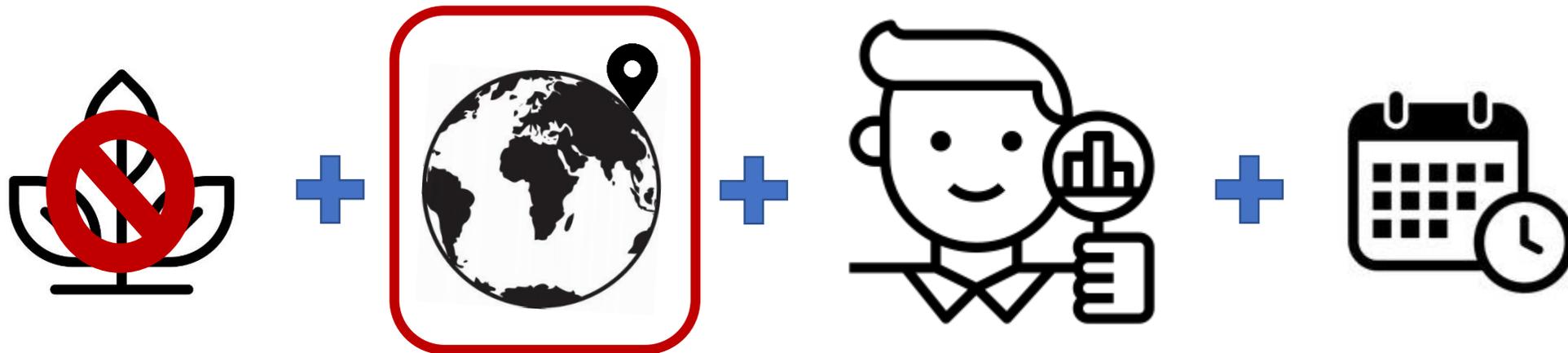
*Domain Schizophrenia* – This involves searching for fields used for purposes for which they may not have been intended. This often happens where a certainty field has not been included in the database and question marks, uncertainties such as cf., aff. are added in the same field as the species epithet, or comments added (Table 2). The nature of this ‘error’ may also depend on database design.

Family	Genus	Species
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

Table 2. Examples of Domain schizophrenia (from Chapman 2005a).

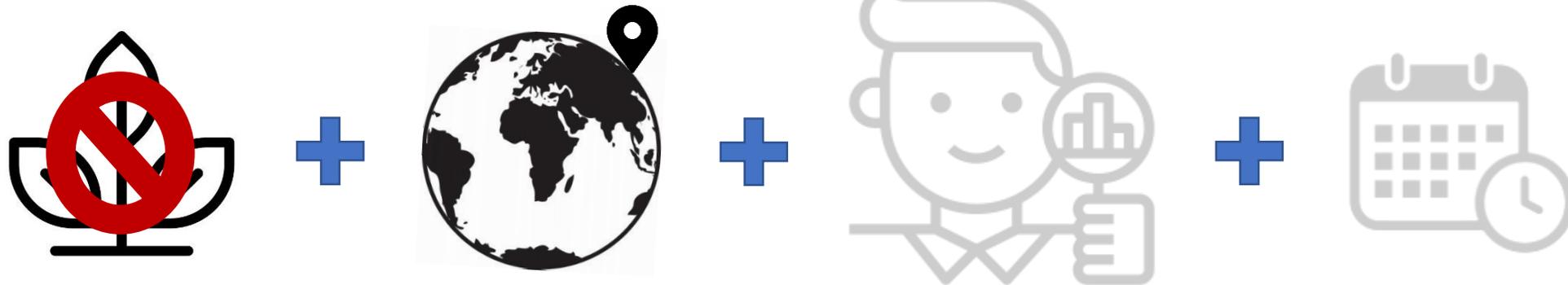


# Módulo 5: Duplicados

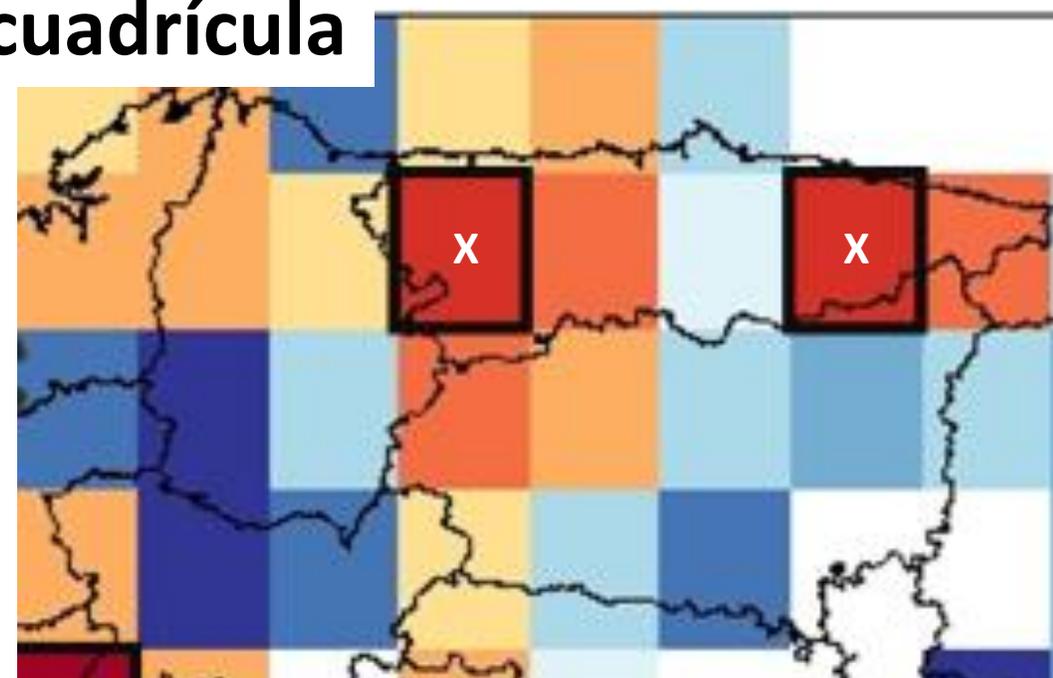
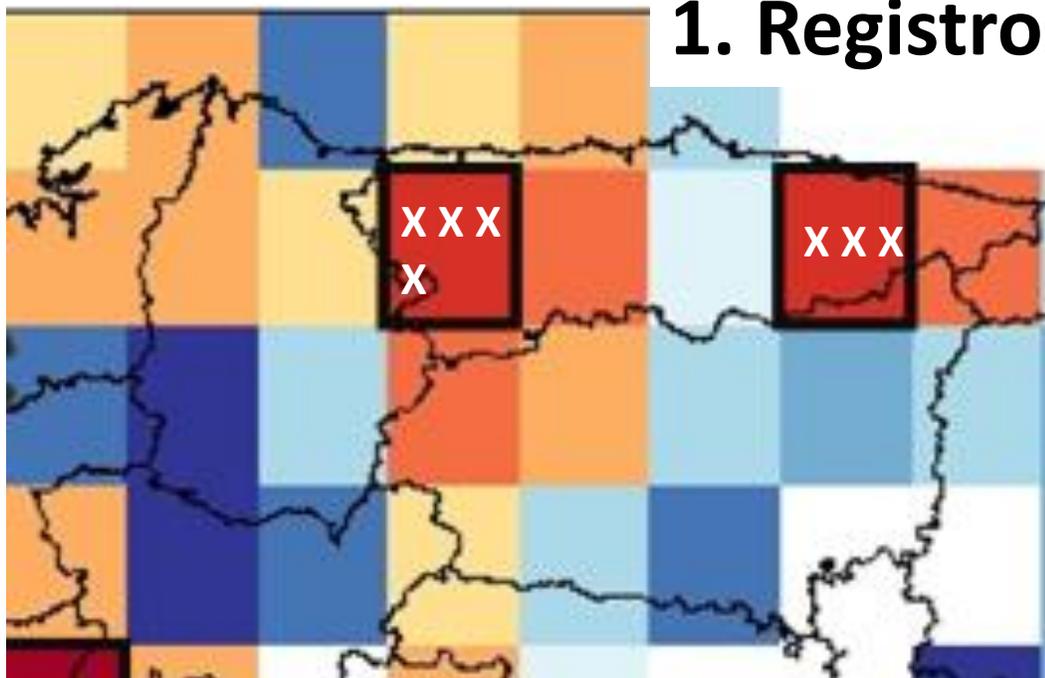




# Módulo 5: Duplicados

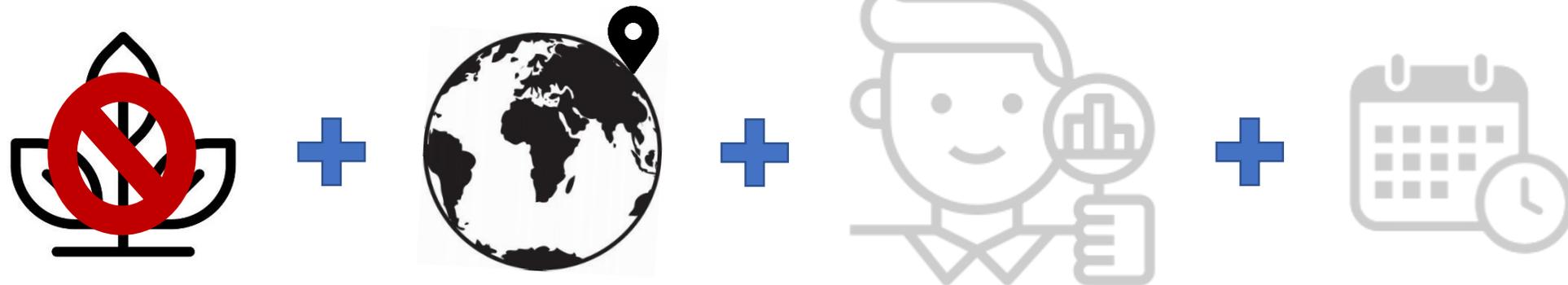


## 1. Registros por cuadrícula

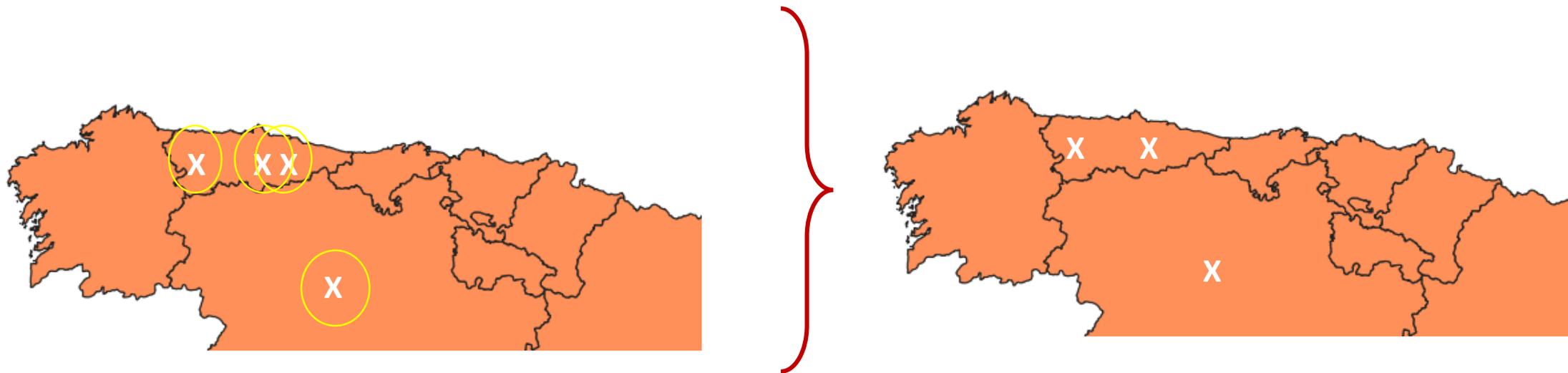




# Módulo 5: Duplicados

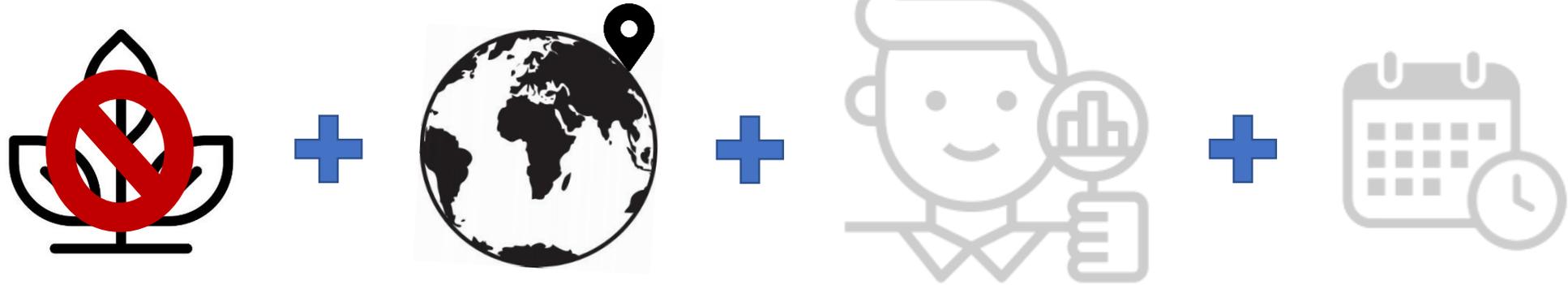


## 2. 'Buffers'





# Módulo 5: Duplicados



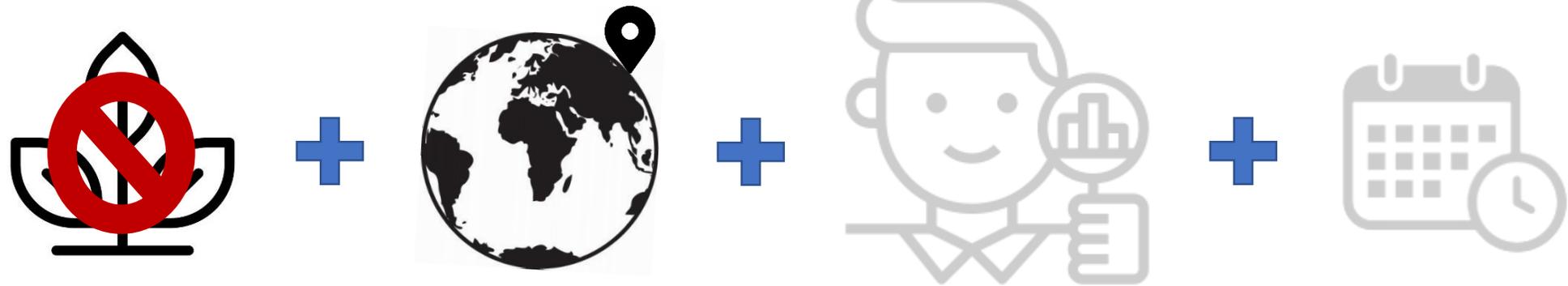
## 3. Valores de las coordenadas

Especie	Localidad	Latitud	Longitud
Ursus arctos	Narnia	43.1533	142.5
Ursus arctos	Narnia	43.1533	142.5
Ursus arctos	Narnia	43.15	142.5

} 2



# Módulo 5: Duplicados



## 4. Valores de las coordenadas redondeados

Especie	Localidad	Latitud	Longitud	Lat Round	Long Round
Ursus arctos	Narnia	23.1533	142.5	23.15	142.5
Ursus arctos	Narnia	23.1533	142.5	23.15	142.5
Ursus arctos	Narnia	23.15	142.5	23.15	142.5

} 1



# Módulo 5: Duplicados



**Fecha de colección**



# Módulo 5: Duplicados



## Fecha de colección

Especie	Localidad	Latitud	Longitud	Fecha
Ursus arctos	Narnia	43.1533	142.5	19/11/2011
Ursus arctos	Narnia	43.1533	142.5	11/2011
Ursus arctos	Narnia	43.15	142.5	11/2011
Ursus arctos	Narnia	43.15	142.5	2011



4



# Módulo 5: Duplicados



## Fecha de colección

Especie	Localidad	Latitud	Longitud	Fecha	Año	Mes
Ursus arctos	Narnia	43.1566	142.5	19/11/2011	2011	11
Ursus arctos	Narnia	43.1566	142.5	11/2011	2011	11
Ursus arctos	Narnia	43.15	142.5	11/2011	2011	11
Ursus arctos	Narnia	43.15	142.5	2011	2011	



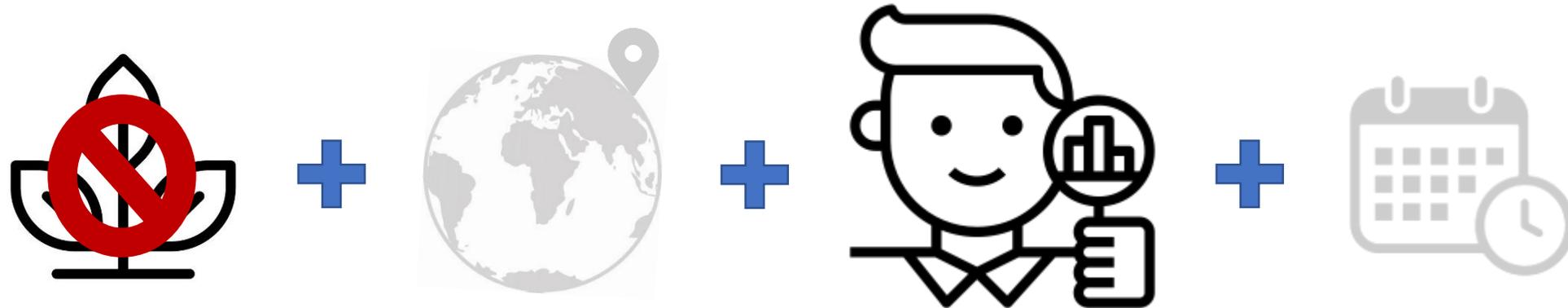
# Módulo 5: Duplicados



**Nombre de los colectores**



# Módulo 5: Duplicados



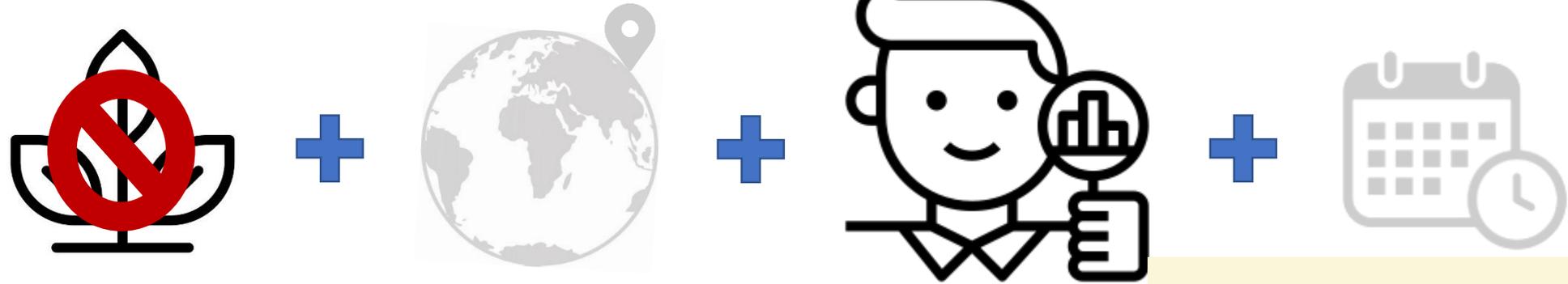
## Nombre de los colectores

Especie	Localidad	Latitud	Longitud	Colector
Ursus arctos	Narnia	23.1566	142.5	M.R.
Ursus arctos	Narnia	23.1566	142.5	M. Ramos
Ursus arctos	Narnia	23.15	142.5	M.R.

} 2



# Módulo 5: Duplicados



## Nombre de los colectores

Especie	Localidad	Latitud	Longitud	Colector
Ursus arctos	Narnia	23.1566	142.5	M.R.
Ursus arctos	Narnia	23.1566	142.5	M. Ram
Ursus arctos	Narnia	23.15	142.5	M.R.

### CHIHUAHUA SYNDROME

MESSY, UNVALIDATED HAND-ENTERED DATA

DOG DATABASE	
Breeds ↑	
BOXER	118,726
BULLDOG	109,215
CHE HUD HUA	2
CHE-HUA-HUA	16
CHE-WAWA	3
CHEE-HUA-HUA	14
CHEEHUAHUA	25
CHEHUAHUA	1032
CHEHAHA	458
CHEHAWA	1
CHIHABA	67
CHIHUA HUA	65,421
CHIHUAWA	17
CHIWAWA	13,098
CHIEHUAVA	362





# Módulo 5: Duplicados



## Nombre de los colectores Ausente!

No. of records | recordedBy

9 | Travassos, L.

1 | Travassos, L.

2 | Travassos; L.

2 | Arag<sup>o</sup> | Jesus | Silva

1 | D.M. Takiya | M. Monn<sup>E</sup>

1 | Gon<sup>A</sup>alves, Pinto | L. Silveira



# Módulo 5: Duplicados



Methods in Ecology and Evolution



APPLICATION | Open Access |

## COORDINATECLEANER: Standardized cleaning of occurrence records from biological collection databases

Alexander Zizka , Daniele Silvestro, Tobias Andermann, Josué Azevedo, Camila Duarte Ritter, Daniel Edler, Harith Farooq, Andrei Herdean, María Ariza, Ruud Scharn ... [See all authors](#)

First published: 20 January 2019 | <https://doi.org/10.1111/2041-210X.13152> | Citations: 494





# Módulo 5: Duplicados



```
> dataDuplic1b <- cc_dup1 (data,  
+                           lon = "decimalLon",  
+                           lat = "decimalLat")  
Testing duplicates  
Removed 55530 records.
```

The screenshot shows the RStudio interface. The 'Data' pane on the left lists several datasets:

Dataset	Observations	Variables
data	87750	53
dataDuplic1b	32220	51
dataDuplic2b	42757	51
dataDuplic3b	53226	51
dataDuplic4b	60490	51
dataDuplic5b	22696	53

On the right, a diagram illustrates the relationship between 'DATA COVERAGE' (y-axis) and 'DATA CERTAINTY' (x-axis). A diagonal line represents the trade-off. A funnel icon is positioned at the intersection of the line. Below the diagram, a 'FILTER STRICTNESS' slider is shown with three circular icons representing different filter settings.

Ya sabéis lo necesario para evaluar calidad de registros..





... Hora de ponerlo en práctica con vuestros casos



# 1 Taller GBIF.ES: Mejora de la calidad de datos de biodiversidad



**GBIF**

Global Biodiversity  
Information Facility

**Cristina Ronquillo**

**Ayudante investigación**

**[cristinaronquillo@mncn.csic.es](mailto:cristinaronquillo@mncn.csic.es)**

mncn 25  museo  
nacional de  
ciencias  
naturales



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS