Explorando el ciclo de vida de los datos de biodiversidad y ambientales: desde la recopilación hasta la publicación

Antonio J. Pérez-Luque
Instituto de Ciencias Forestales (CIFOR) | INIA-CSIC (Madrid)

Marzo 04-06, 2025

Calidad de los Datos de Biodiversidad y Ambientales

Antonio J. Pérez-Luque
Instituto de Ciencias Forestales (CIFOR) | INIA-CSIC (Madrid)

2025-03-06

Conceptos de calidad y limpieza de datos científicos

Principios generales de la calidad de datos

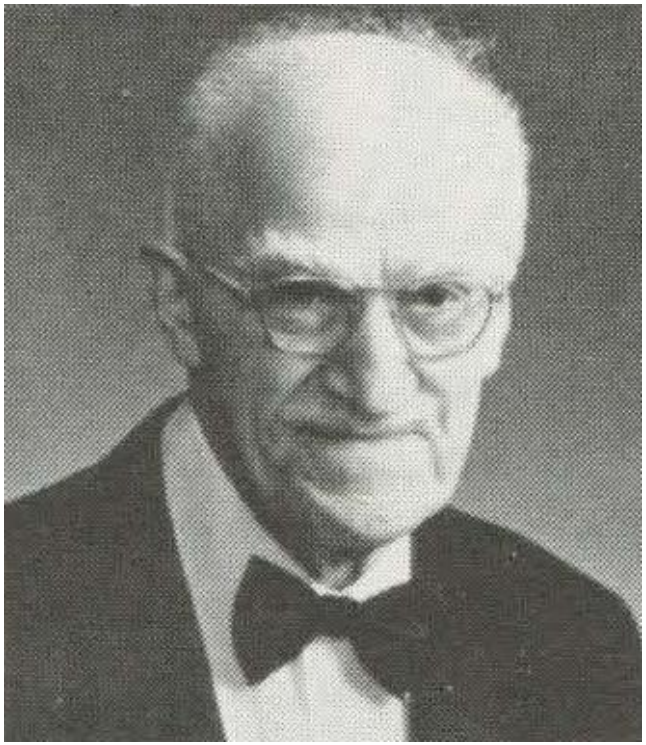
Evaluación de la calidad de los datos: datos espaciales y temporales; datos taxonómicos

Calidad de los datos durante la digitalización y el almacenamiento

Calidad de datos

Fitness for use

Los datos son de alta calidad si son adecuados para su uso previsto en operaciones, toma de decisiones y planificación



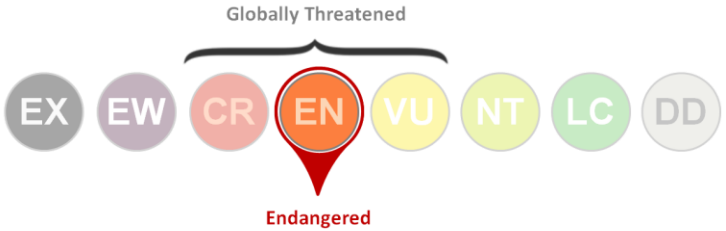
Joseph M. Juran

Calidad de datos

Fitness for use

Registro de presencia de esta especie (0.15° precisión)

Penelope albipennis

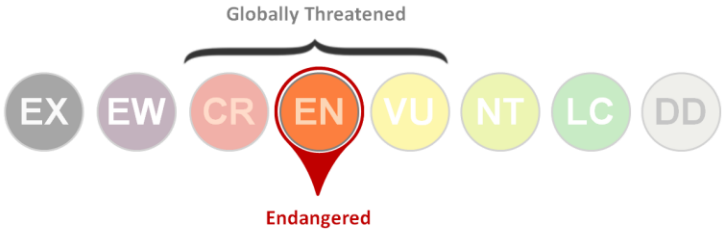


https://www.gbif.org/occurrence/2596301932

Calidad de datos

Fitness for use

Penelope albipennis



https://www.gbif.org/occurrence/2596301932

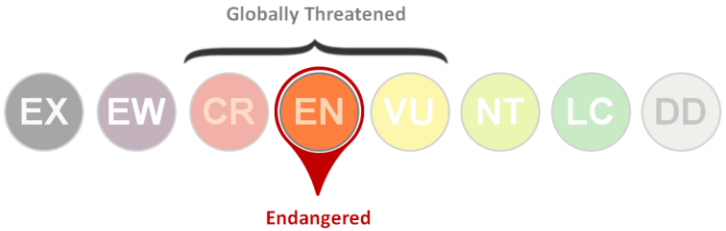
1. Prepare un listado de aves amenazadas de Péru

Calidad de datos

Fitness for use

Refugio de Vida Silvestre Laquipampa

Penelope albipennis

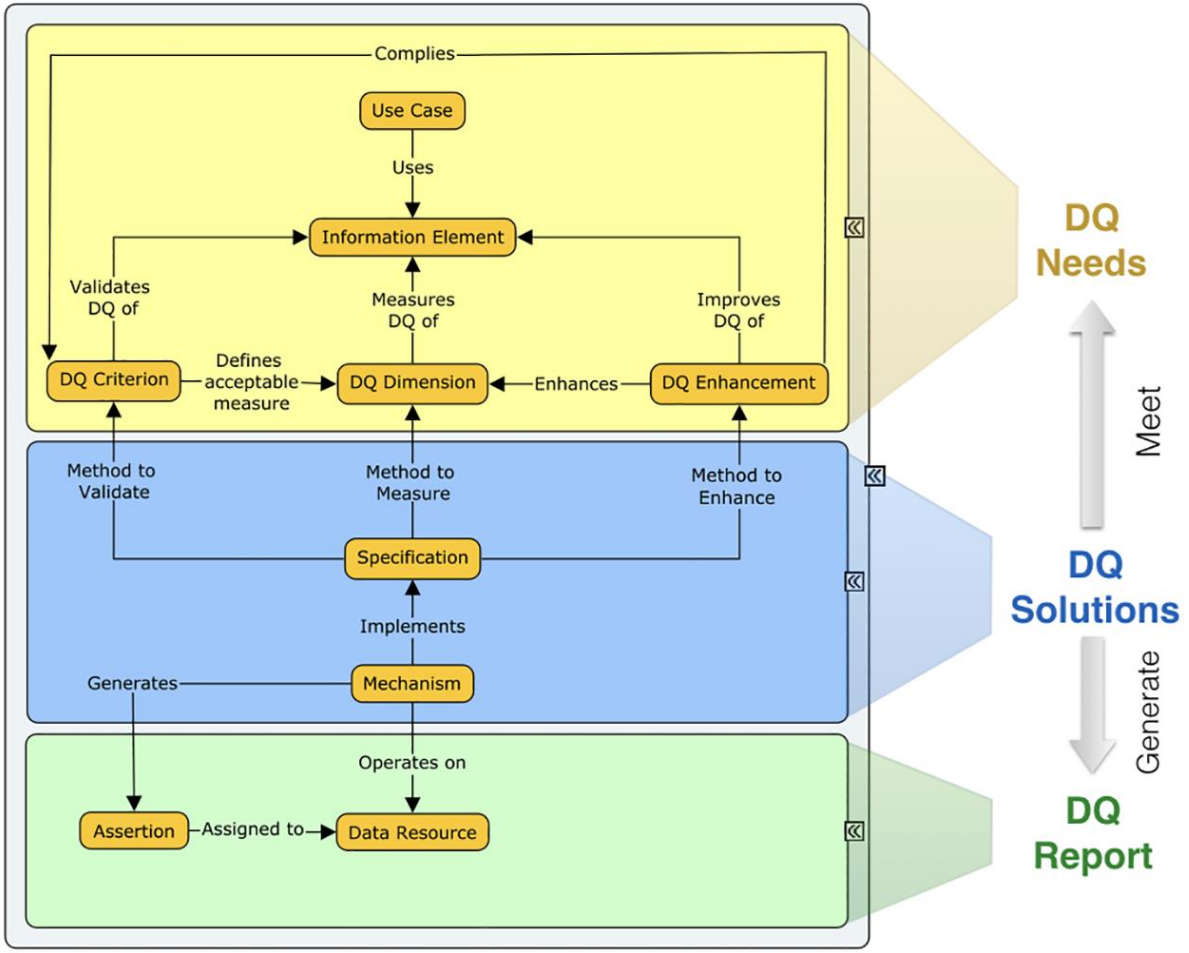


https://www.gbif.org/occurrence/2596301932

Prepare un listado de aves amenazadas del Espacio Natural: Refugio de Vida Silvestre Laquipampa

Calidad de datos

La evaluación y la gestión de la calidad de los datos no pueden llevarse a cabo si no hemos establecido claramente las necesidades de calidad desde el punto de vista del **usuario de los datos**.



Veiga et al. 2017 https://doi.org/10.1371/journal.pone.0178731

https://tdwg.github.io/bdq/tg1/site/index.html

Calidad de datos: Problemas

Redundancia: sinónimos, no estándares
Ecuador; *República del Ecuador*

Registros duplicados

Propagación del error: creación de nuevos datos a partir de datos existentes incorrectos.

Valor faltante: valor ausente. *Ausencia latitud, longitud*

Valor incorrecto: valor no representa los hechos.
Quercus pyrenaica - Quercus pirenaica

Valor no atomizado: más de un valor en un campo atomizado.
Locality = "Robledal de Cáñar, Cáñar, Granada, España"

Campos utilizados erróneamente. *lifeStage = "Male"*

Valores inconsistentes: incongruencia en la representación de un hecho
Presencia de una especie extinta e.g. *Lagostomus crassus (*Perú) (UICN

Algunas de las causas de la falta de calidad de datos

Idioma: España, Spain

Formato: Ecuador, ECUADOR, ecuador, EC

Mapeo incorrecto

(mi base de datos → estándar)
county, country

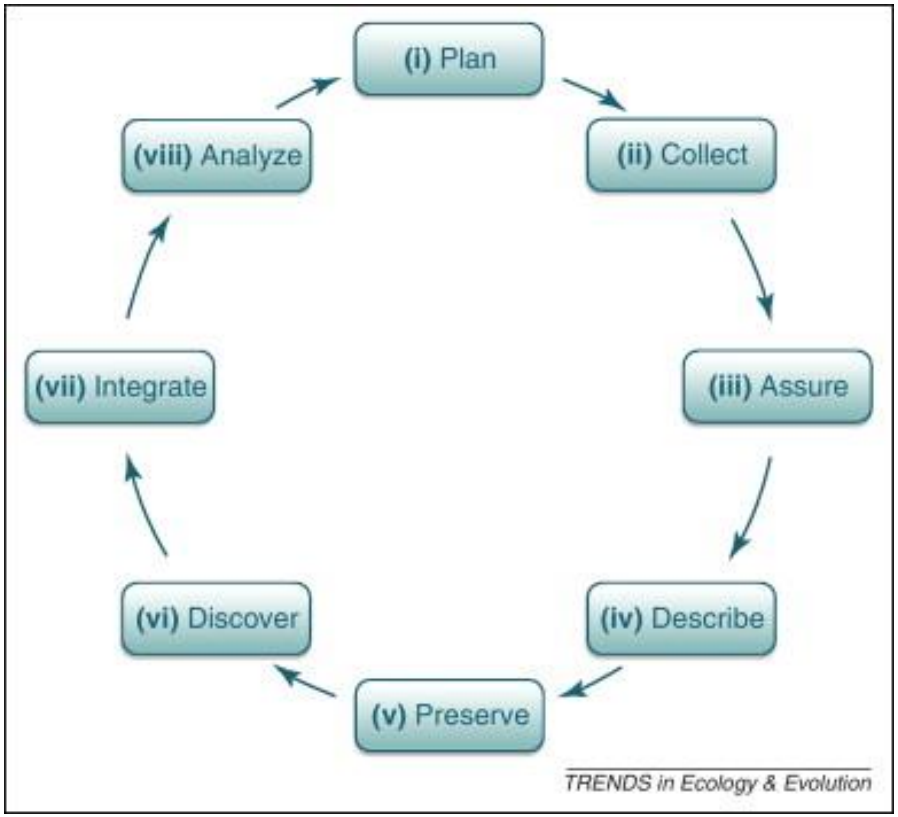
Varias palabras mismo concepto

Misma palabra varios conceptos

Calidad de datos: Soluciones

Pérdida de calidad de los datos …
¿Cuando pierden calidad los datos?

e.g. Datos Primarios de Biodiversidad
- Durante la recolección
- Durante la digitalización
- Durante la documentación
- Durante el almacenamiento y conservación
- Durante el análisis y la manipulación
- En el momento de la presentación
- En el uso que se les dé

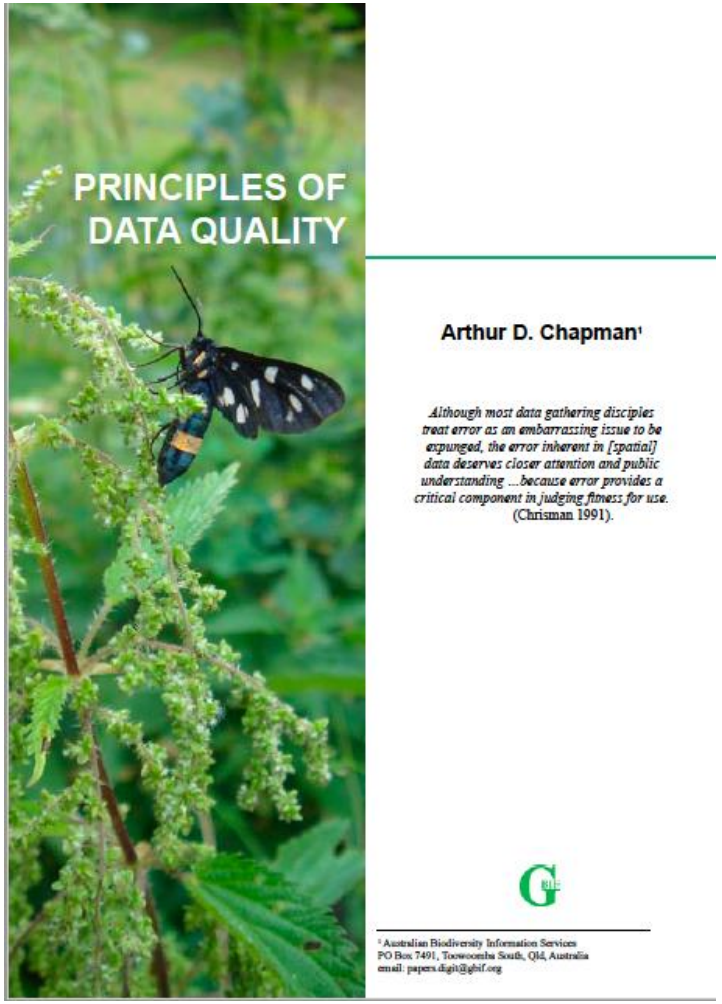


Calidad de datos: Soluciones

COSTO DE LA CORRECCIÓN DE ERRORES

Calidad de datos de Biodiversidad: Principios

Chapman, A. D. 2005.
Principles of Data Quality, version 1.0.
Report for the Global Biodiversity Information Facility, Copenhagen.
http://www.gbif.org/orc/?doc_id=1229



Calidad de datos de Biodiversidad

Prevenir es mejor que curar

La prevención de errores nada tiene que hacer con los datos que ya existen en la base de datos. En estos casos, la validación y la corrección serán muy importantes en el proceso hacia la calidad.

Detectar las causas nos ayudará a prevenirlas

Corregir los datos y no hacer nada para prevenir los errores significa que los errores seguirán apareciendo sistemáticamente y no los reduciremos nunca.

Calidad de datos de Biodiversidad

Prevenir:

Evitar que se presenten errores previo a la creación de los datos

Detectar y Limpiar:

Detectar errores en el conjunto de datos y corregirlos

Detectar y Recomendar

Detectar errores en el conjunto de datos y generar recomendaciones de limpieza

Calidad de datos: Principios

Es muy importante que como institución/proyecto exista:

Una **visión** con respecto a la calidad de sus datos

Una **política** para implementar esta visión

Una **estrategia** para su implementación

No llevar nunca esta labor a cabo sin planificación ni sin coordinación.

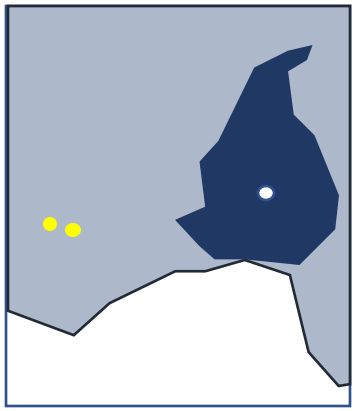
Data Management Plan

Calidad de datos de Biodiversidad: Principios

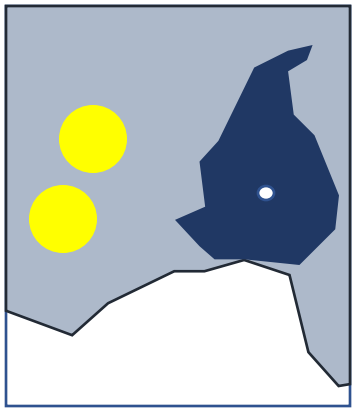
Exactitud

Exactitud y precisión

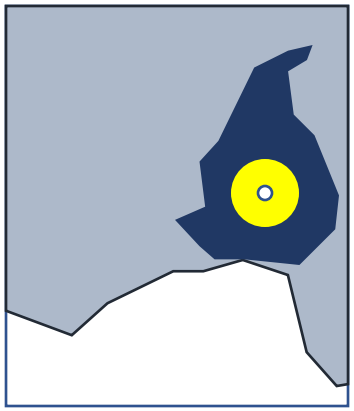
Localización real
Localización proporcionada por el Colector



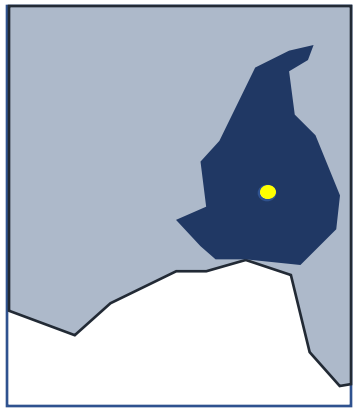
Alta Precisión,
Baja Exactitud



Baja Precisión,
Baja Exactitud



Baja Precisión,
Alta Exactitud



Alta Precisión,
Alta Exactitud

Modificado de Chapman (2005)

Calidad de datos de Biodiversidad: Principios

Consistencia

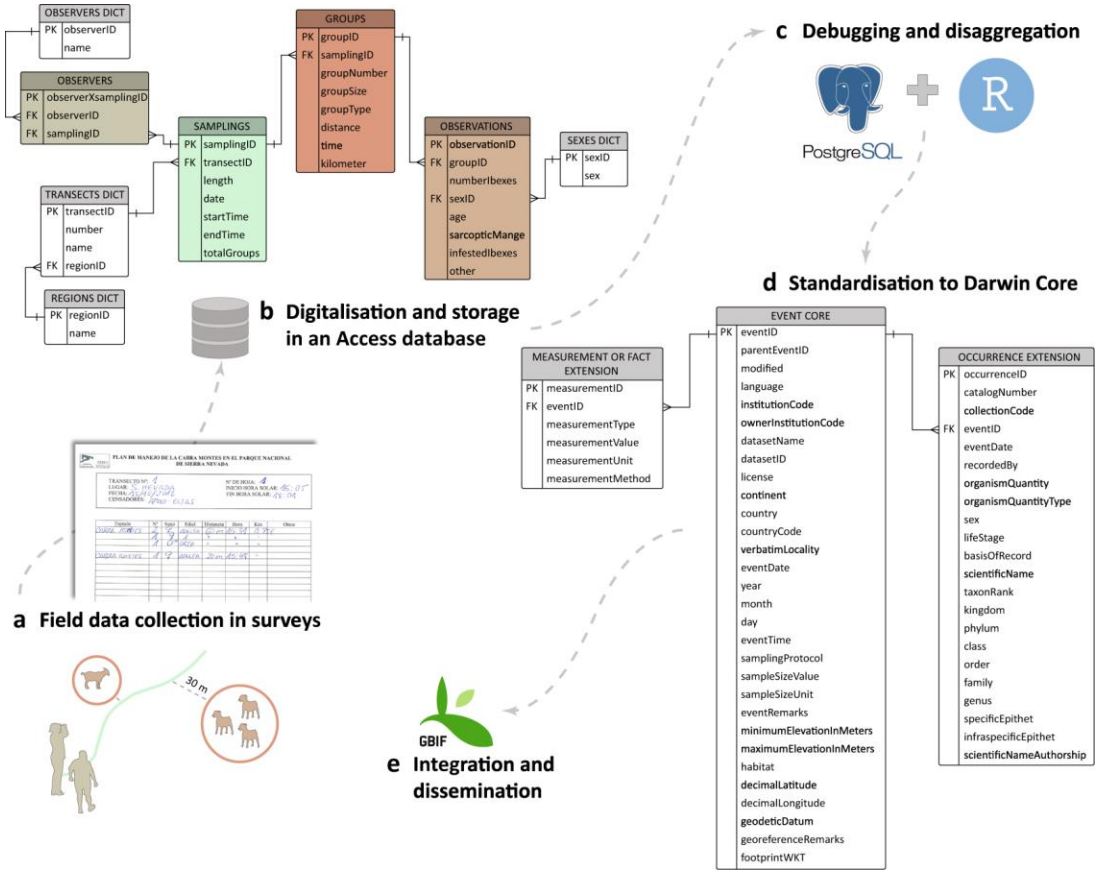
- Usar reglas, convenciones, estándares
- Representar la información relativa a un atributo de forma consistente, *i.e.* siempre de la misma forma

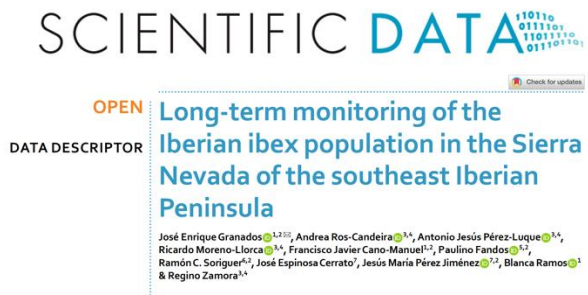
Género	Especie	Infraespecie
Quercus	ilex	subsp. ballota

Género	Especie	Rango Infraespecífico	Infraespecie
Quercus	ilex	subsp.	ballota

Calidad de datos de Biodiversidad: Principios

Conservación de datos originales





https://www.gbif.org/es/dataset/5fe81b76-6d2f-4f09-a9e3-301e8e1eb918

Granados, J.E., Ros-Candeira, A., Pérez-Luque, A.J. *et al.* *Sci Data* (2020)
https://doi.org/10.1038/s41597-020-0544-1

Calidad de datos de Biodiversidad: Principios

Transparencia (Documentación)

Documentar todos los procesos realizados sobre el conjunto de datos

Creación de la vista (sql) de DWC-A de borreguiles

Versión de la BD

- Versión: BD_borreguiles_v5.9_20141103.mdb
- path: `./BorreguilesDP/1_bd/BD_borreguiles_v5.9_20141103.mdb`
- La llamaremos *BD_obsnev*

Para esta consulta consideramos una ocurrencia: *un taxon es observado en un plot y anotado en una fecha concreta*

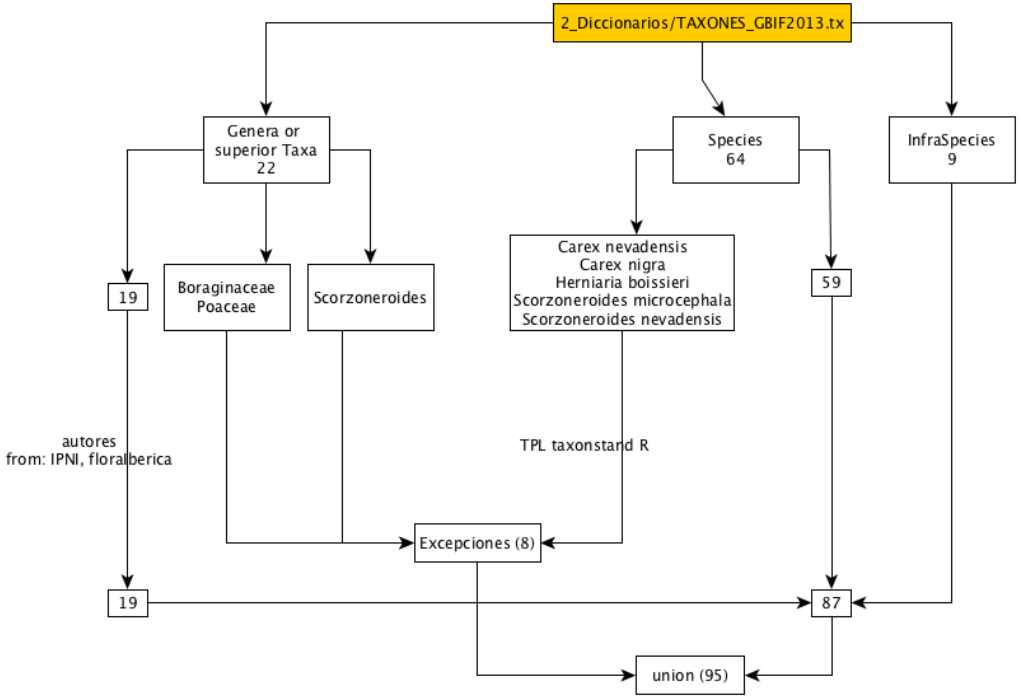
Consultas a la BD_obsnev

Creamos varias consultas sobre la BD_obsnev (la que está en linaria) para poder generar una vista DWC-A. A estas consultas las llamaremos:

- GBIF2014_C1

GBIF2014_C1

- Esta consulta parte de la tabla `TABL_INVENTARIO`
- Sobre ella creamos una consulta, con la restricción de que el campo taxón sea no nulo (`TABL_INVENTARIO.COD_TAXON>0`)
- Creamos un indentificador para cada registro de la consulta, de forma que sea unívoco. Hemos generado un código o clave que es combinación de varios identificadores: ***BORREG-000-XXXXXX-AAAAAAAAA***
  - ***BORREG*** es el nombre de la metodología
  - ***000*** corresponde al *ID_PARCELA*



https://github.com/ajpelu/BorreguilesDP

Calidad de datos de Biodiversidad: Principios

Accesibilidad

Establecer de forma clara los mecanismos de acceso a los datos para que cualquier usuario pueda realizar sus análisis.

Roche et. al. 2015 https://doi.org/10.1371/journal.pbio.1002295

n = 100 conjuntos de datos

56% estaban incompletos.

44% archivados

64% estaban archivados de manera que parcial o totalmente impedían la reutilización.

Datos en **"peligro de extinción"**

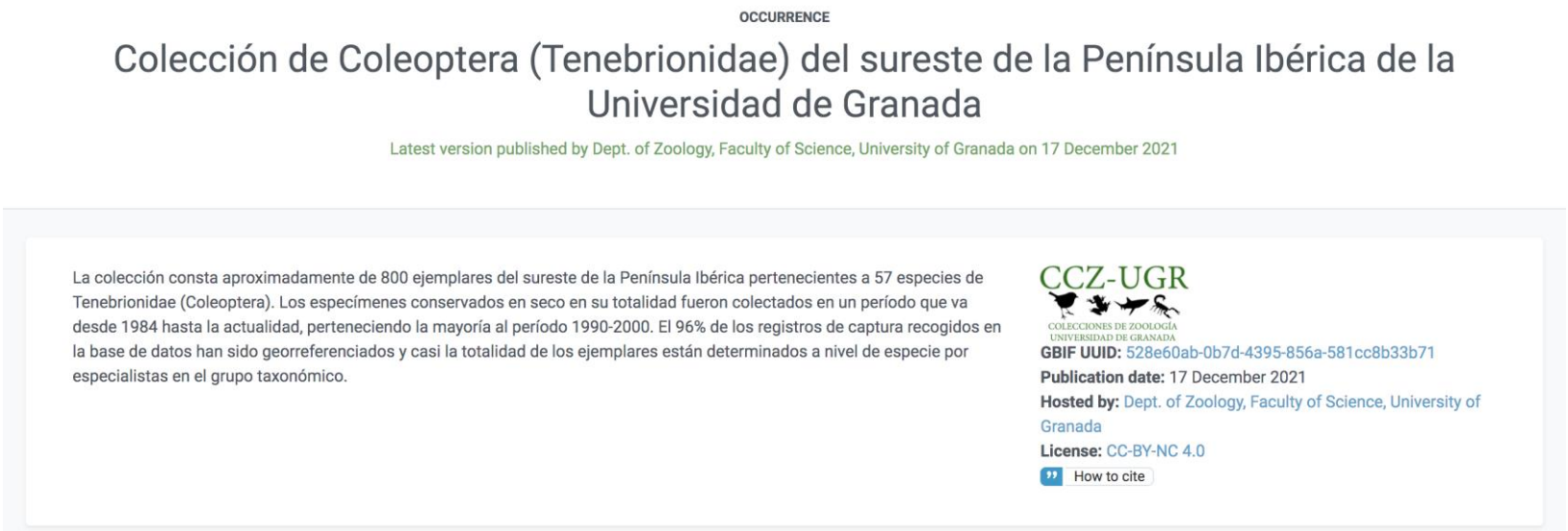
Calidad de datos de Biodiversidad: Principios

Actualización

¿Cuándo fueron los datos actualizados por última vez?

¿Con qué frecuencia se actualizan y son puestos a disposición de los usuarios?

Documentar y concretar la frecuencia de actualización



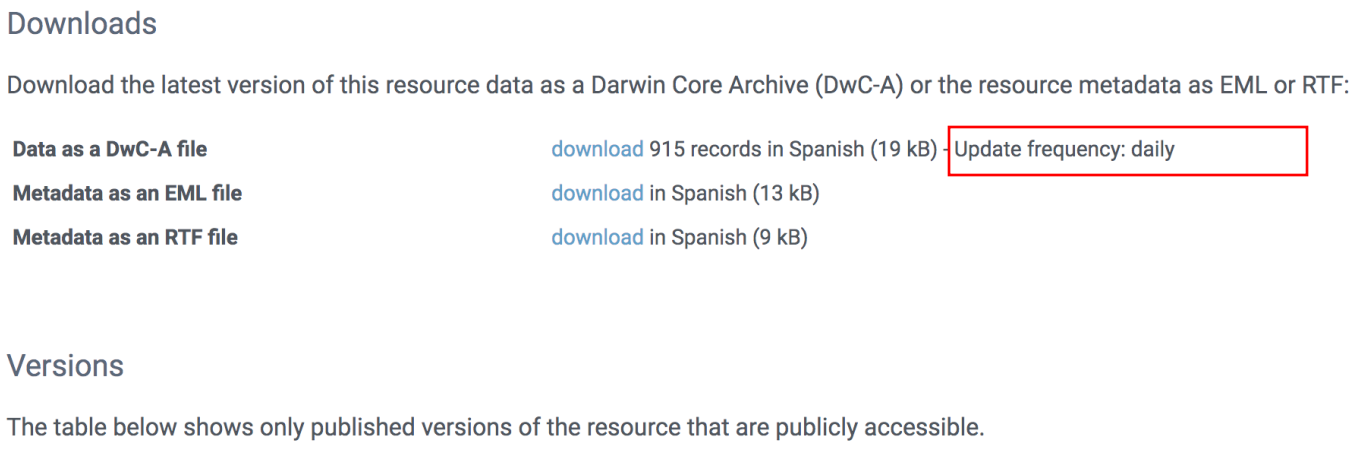
Calidad de datos de Biodiversidad: Principios

Actualización

¿Cuándo fueron los datos actualizados por última vez?

¿Con qué frecuencia se actualizan y son puestos a disposición de los usuarios?

Documentar y concretar la frecuencia de actualización



Downloads

Download the latest version of this resource data as a Darwin Core Archive (DwC-A) or the resource metadata as EML or RTF:

Data as a DwC-A file	download 915 records in Spanish (19 kB) - Update frequency: daily
Metadata as an EML file	download in Spanish (13 kB)
Metadata as an RTF file	download in Spanish (9 kB)

Versions

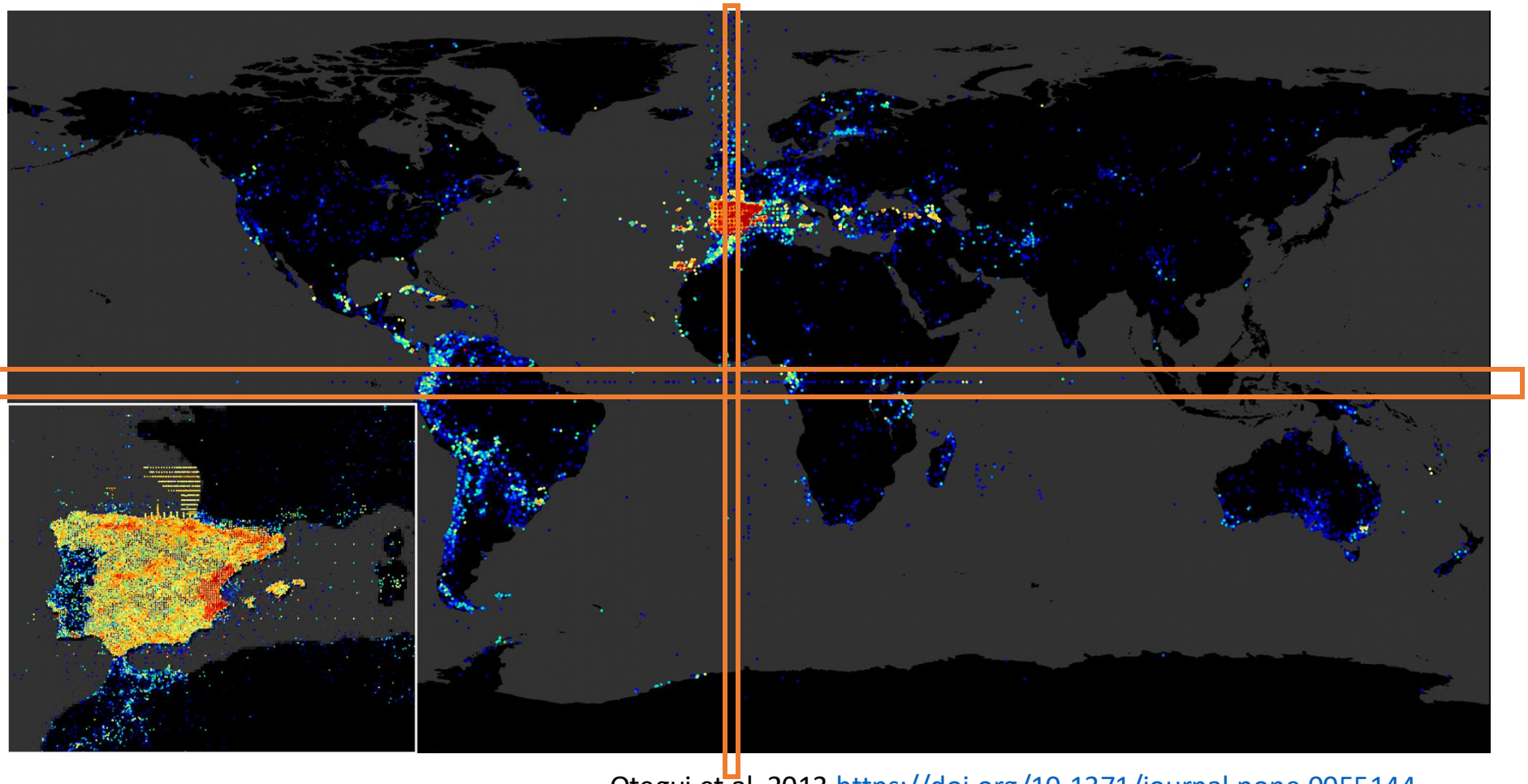
The table below shows only published versions of the resource that are publicly accessible.

Calidad de datos de Biodiversidad: Principios

Depuración y validación de datos

Evaluación de la calidad de los datos: Datos Espaciales

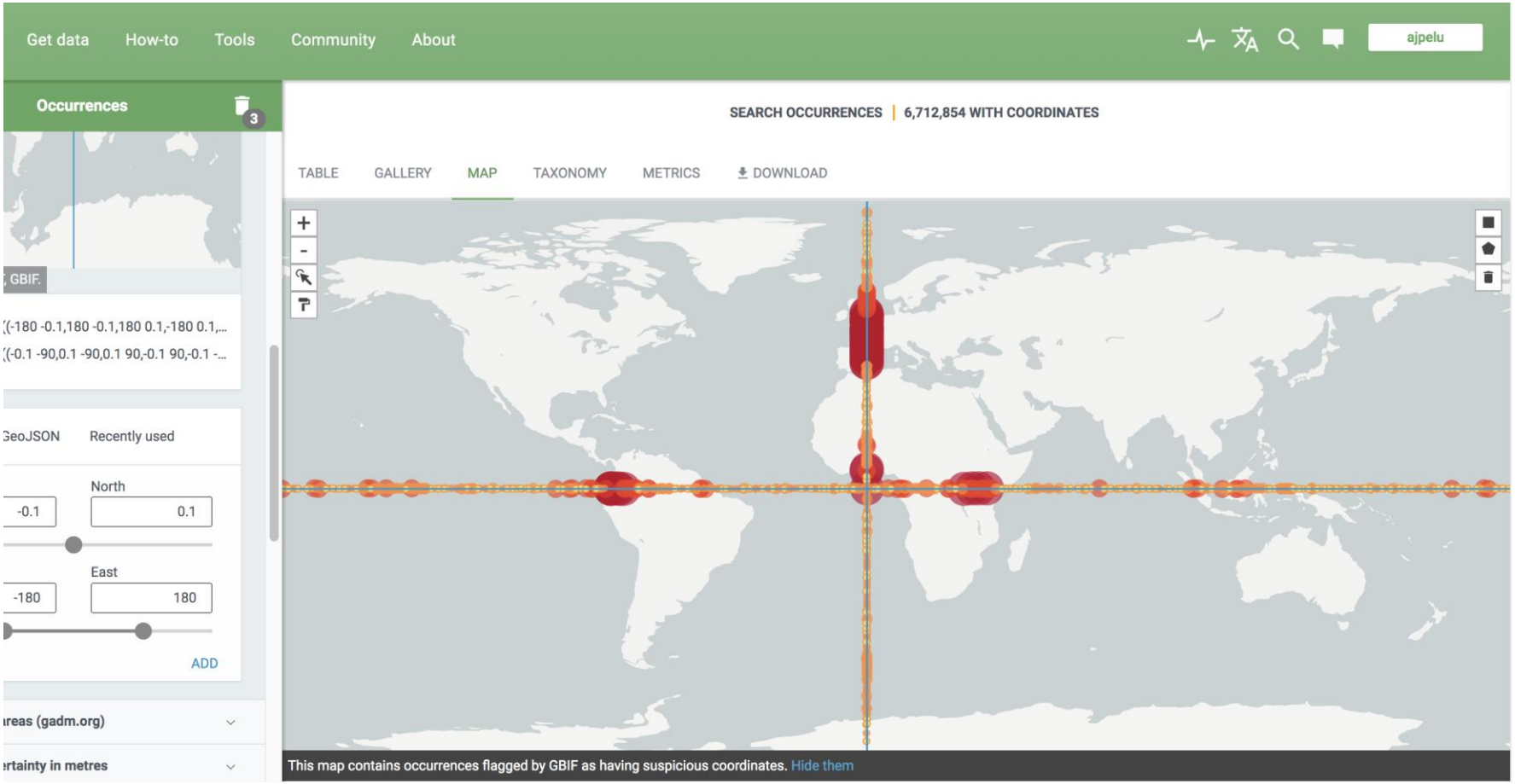
Datos del Nodo Español de GBIF



Otegui et al. 2013 https://doi.org/10.1371/journal.pone.0055144

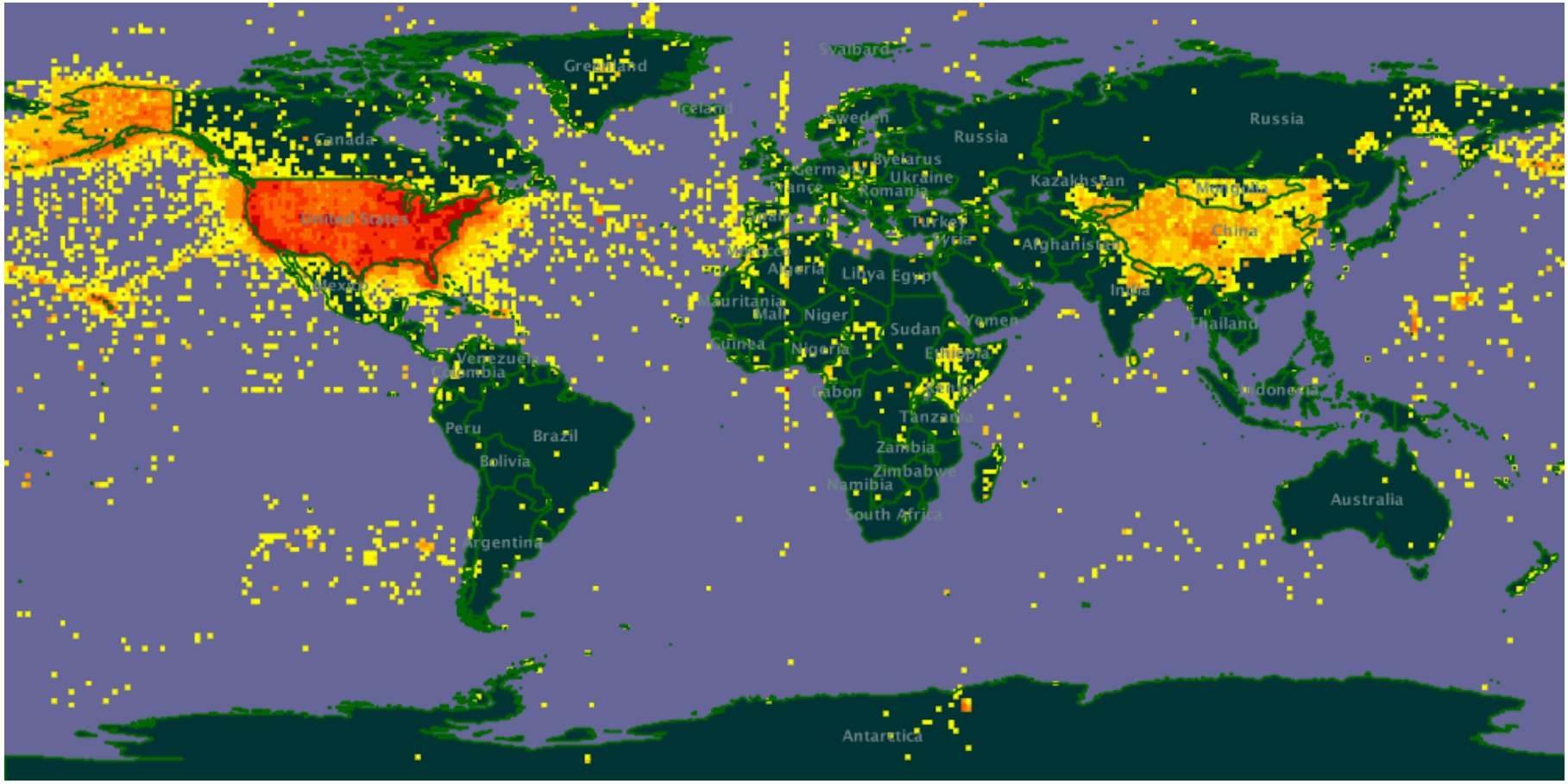
Evaluación de la calidad de los datos: Datos Espaciales

Datos GBIF (2022)



Evaluación de la calidad de los datos: Datos Espaciales

Datos de Ocurrencias EEUU



Evaluación de la calidad de los datos: Datos Temporales

No se ajustan a un estándar

Fechas ausentes o incompletas

No son uniformes, están representadas de múltiples formas en los campos

Fecha de colecta
2001
19/03/2001
Mar 19, 2001

Biodiversity Informatics, 8, 2013, pp. 173-184

ON THE DATES OF THE GBIF MOBILISED PRIMARY BIODIVERSITY DATA RECORDS

JAVIER OTEGUI (1), ARTURO H. ARIÑO (1)*, VISHWAS CHAVAN (2) AND SAMY GAIJI (2)
(1) Department of Zoology and Ecology, University of Navarra, Pamplona, Spain.
(2) Global Biodiversity Information Facility Secretariat, Universitetsparken 15, 2100, Copenhagen, Denmark.
**corresponding author*

Abstract — There are more than 390 million primary biodiversity data records published by hundreds of data publishers through the GBIF network. Thus, the GBIF network is the single most comprehensive index for this kind of data. Ensuring or, at least assessing data quality is of capital importance for the reliability and usability of this data. While conducting a time data gap analysis on this mass of data, we have detected some issues with the way date information is processed and shared. Dates can be obscured or altered under certain circumstances, when a specific combination of publisher's error or date handling features, and faulty or inadequate date parsing and processing routines gets chained together. The extent of the date unreliability (either at the source or through GBIF portal) is relatively low, and problems are concentrated in a few data publishers. The types of errors and misprocessing in dates through the sources and the published records are analysed, impact on the overall data quality of the published index was assessed, and corrective measures are suggested.

May 2001
Feb 2001
19/03/2001
20/03/2001

Evaluación de la calidad de los datos: Vocabularios controlados

sex	
Identifier	http://rs.tdwg.org/dwc/terms/sex
Definition	The sex of the biological individual(s) represented in the Occurrence.
Comments	Recommended best practice is to use a controlled vocabulary.
Examples	`female`, `male`, `hermaphrodite`

24426 Valores únicos*

```
", 1 juvenile",1
", 7 juveniles",1
", Ginandromorfo",2
", Hembra",5016
", Hembra, Hembra",569
", Hembra, Hembra, Hembra",4
", Hembra, Hembra, Hembra, Hembra",1
", Hembra, Macho",1
", Mach",1
", Macho",9549
", Macho, Hembra",2
", Macho, Macho",464
", Macho, Macho, Macho",9
", Marcelo",1
", hembra",1
", macho",2
–,2249
––,42362
0.3,1
0,5728
0 Indeterminado,4
0 female,3
"0 hembras, 1 macho",4
"0 hembras, 14 machos",1
"0 hembras, 2 machos",4
"0 hembras, 4 machos",1
"0 machos, 0 fêmeas",2310
"0 machos, 1 fêmeas",1398
"0 machos, 2 fêmeas",1
```

Property

* Datos de GBIF 2020-04-09 https://github.com/tdwg/dwc-qa/tree/master/data

https://tinyurl.com/2p84z7ue

Recursos

Chapman, A. D. 2005. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

SiB Colombia (2015). Calidad de Datos - Guía de herramientas para mejorar los datos primarios de biodiversidad. Escobar, D., Beltrán, N., Buitrago, L., Plata, C. Delgado, E.; versión 1.0. Bogotá: SiB Colombia

Ortega Maqueda (2007). Principios de Calidad de Datos. Taller de calidad de datos en Bases de datos de Biodiversidad. Unidad de Coordinación GBIF-ES. Real Jardín Botánico de Madrid (España) 13-14 Septiembre 2007

Cezón-García. Taller Online Calidad en Bases de Datos de Biodiversidad. Unidad de Coordinación GBIF-ES. https://www.gbif.es/talleres/calidad-de-datos-biodiversidad-online/

Zermoglio, P (2018). Conceptos básicos de Calidad de Datos. II Taller GBIF.ES: Publicación de datos de biodiversidad en GBIF y en revistas científicas. https://www.gbif.es/talleres/ii-taller-publicacion-datos-biodiversidad/

¿Alguna duda?

ajpelu.bsky.social

antonio.perez@inia.csic.es

Ayuda JDC2022-050056-I financiada por MCIN/AEI /10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR

MINISTERIO DE CIENCIA E INNOVACIÓN
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
AGENCIA ESTATAL DE INVESTIGACIÓN

Si usas esta presentación puedes citarla como:

Pérez-Luque, A.J. (2025). Explorando el ciclo de vida de los datos de biodiversidad y ambientales: desde la recopilación hasta la publicación. Calidad de los Datos. Material Docente de la Asignatura: Ciclo de Gestión de los Datos. Master Universitario en Conservación, Gestión y Restauración de la Biodiversidad. Universidad de Granada. https://ecoinfugr.github.io/ecoinformatica/