

# Herramientas útiles para la calidad y limpieza de datos

Regional capacity enhancement to Latin America by establishing Chile's node

Katia Cezón

[katia@gbif.es](mailto:katia@gbif.es)

GBIF.ES



# CADENA DE INFORMACIÓN DE LOS DATOS DE BIODIVERSIDAD



## 1 PLANIFICACIÓN

No existe

## 2 RECOLECCIÓN Y DOCUMENTACIÓN

- Información taxonómica
- Información espacial
- Datos asociados a la colecta
- Datos descriptivos

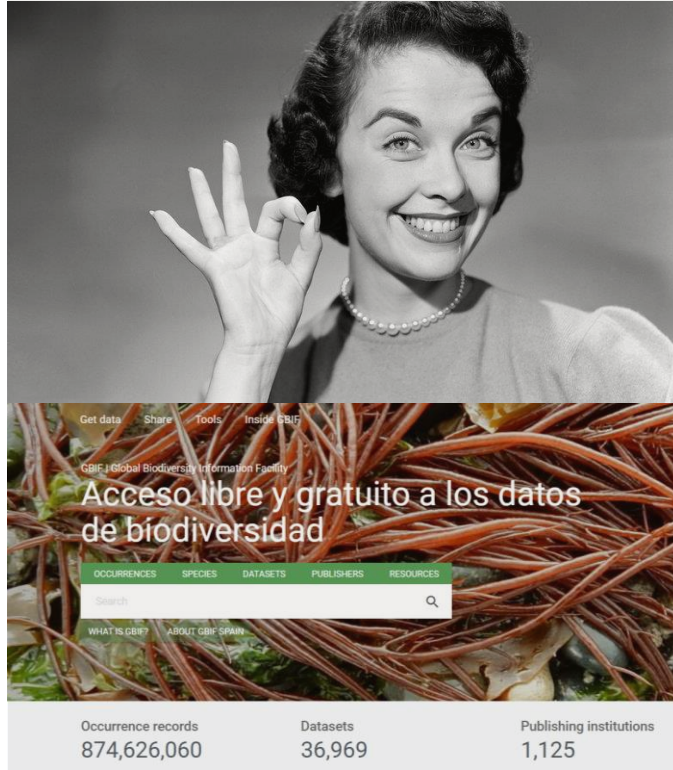
Coste  
corrección  
errores

## 3 DIGITALIZACIÓN

- Experiencia del personal
- Interpretación de los datos
- Diseño base de datos
- Copias de seguridad



# CADENA DE INFORMACIÓN DE LOS DATOS DE BIODIVERSIDAD



## 4 CONTROL DE CALIDAD

Feedback

## 5 PUBLICACIÓN EN GBIF

- Exportación
- Conversiones
- Adaptación de los dato:

Coste  
corrección  
errores



# ASPECTOS A TENER EN CUENTA

- Precio, disponibilidad, licencia
- Requerimientos técnicos
- Facilidad de uso
- Documentación y soporte
- Flexibilidad
- Automatización
- Etc.





**NO EXISTE  
UNA  
RECETA**



# HERRAMIENTAS ÚTILES PARA LA CALIDAD Y LIMPIEZA DE DATOS

## Herramientas de almacenamiento y gestión de datos

- Excel, Access, Open Office, etc.

## Herramientas para el tratamiento de nombres científicos, fechas y coordenadas

- Herramientas para la gestión de nombres científicos (atomización, herramientas para comprobar status, búsqueda de autores, etc.)
- Herramientas geográficas (visualización, comprobación de coordenadas, conversión, etc.)
- Herramientas para el tratamiento de las fechas

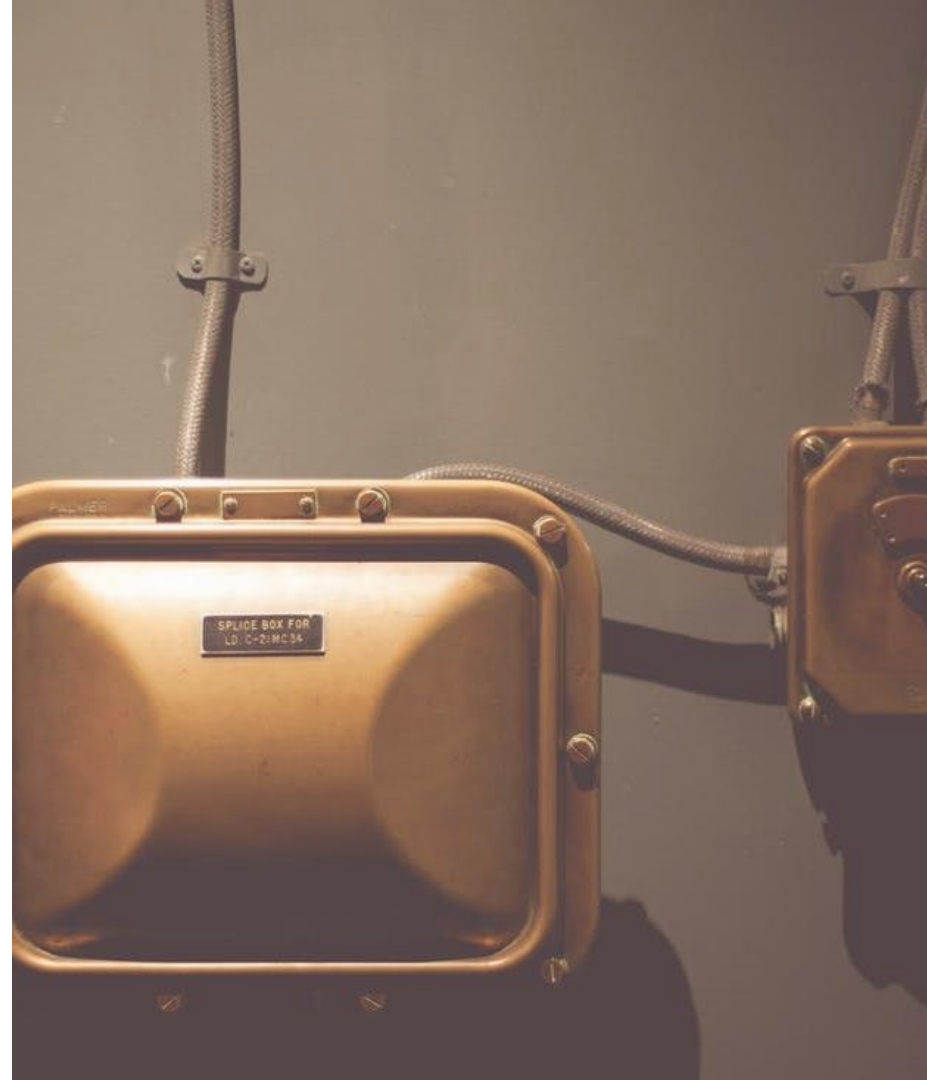
## Herramientas para la limpieza y validación de datos

- Open Refine
- Darwin Test
- Darwin Core Archive Validator



# HERRAMIENTAS DE ALMACENAMIENTO Y GESTIÓN DE DATOS

Permiten manejar los datos mediante tablas (formadas por filas o registros y columnas o variables), crear relaciones entre tablas, elaborar consultas y formularios para introducir datos o informes para extraer la información.





# HERRAMIENTAS PARA EL TRATAMIENTO DE NOMBRES CIENTÍFICOS

Permiten realizar tareas como la **atomización** de sus componentes (género, epíteto específico, etc.), permiten la **resolución de nombres científicos** de cualquier grupo taxonómico, . generan la jerarquía taxonómica, etc.

## Atomización

[GBIF - Name parser](#)

[Name Parser GBIF España](#)

## Resolución

[List Matching Service](#)

[Global Names Resolver](#)

[T-REX](#)

[iPlant](#)

[Species matching GBIF](#)





# HERRAMIENTAS PARA EL TRATAMIENTO DE NOMBRES CIENTÍFICOS



The screenshot shows a Microsoft Access window with a table named 'Separa nombres'. The table contains scientific names and their corresponding taxonomic data. The columns are: name, gen, Name synl, is\_sp, esspaut, infr, infra, infraut, and has\_year. The data is organized into groups of rows for each species, showing different taxonomic treatments.

name	gen	Name synl	is_sp	esspaut	infr	infra	infraut	has_year
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Herichthys pantostictus (Taylor & miller)	Herichthys	-1	0	pantostictu (Taylor & miller, 19				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Astyanax mexicanus De filippi, 1853	Astyanax	-1	0	mexicanus De filippi, 1853				Yes
Xiphophorus helleri Heckel, 1848	Xiphophorus	-1	0	helleri Heckel, 1848				Yes
Girardinichthys viviparus Boustamante, 1837	Girardinichthys	-1	0	viviparus Boustamante, 1837				Yes

# HERRAMIENTAS PARA EL TRATAMIENTO DE NOMBRES CIENTÍFICOS



## Matched Species

[Download to file](#)

Your Data	Scientific Name	Status
Helianthemum squamatum (L.) Pers.	<i>Helianthemum squamatum</i> (L.) Pers.	accepted name
Thymus lacaitae Pau	<i>Thymus lacaitae</i> Pau	accepted name
Thymus lacaitae Pau	<i>Thymus lacaitae</i> Pau	accepted name
Thymus lacaitae Pau	<i>Thymus lacaitae</i> Pau	accepted name
Thymus lacaitae Pau	<i>Thymus lacaitae</i> Pau	accepted name
Thymus vulgaris L.	<i>Thymus vulgaris</i> L.	accepted name
Thymus vulgaris L.	<i>Thymus vulgaris</i> L.	accepted name
Lepidium subulatum L.	<i>Lepidium subulatum</i> L.	accepted name
Lepidium subulatum L.	<i>Lepidium subulatum</i> L.	accepted name
Lepidium subulatum L.	<i>Lepidium subulatum</i> L.	accepted name
Lepidium subulatum L.	<i>Lepidium subulatum</i> L.	accepted name
Lepidium subulatum L.	<i>Lepidium subulatum</i> L.	accepted name
Centaurea hyssopifolia Vahl	<i>Centaurea hyssopifolia</i> Vahl	accepted name
Centaurea hyssopifolia Vahl	<i>Centaurea hyssopifolia</i> Georgi	synonym

# HERRAMIENTAS PARA EL TRATAMIENTO DE NOMBRES CIENTÍFICOS



Resultados de la búsqueda taxonómica

Descargar

XLSX

Fuentes seleccionadas: GBIF Backbone Taxonomy

Total Datos 19

Página Actual 1 de 2

Anterior

Siguiente

Nombre ingresado	Nombre científico resuelto	Fuente de información	Puntaje de coincidencia	Tipo de Coincidencia	Detalles
Herichthys pantostictus (Taylor & miller, 1983)	Nosferatu pantostictus (Taylor & Miller, 1983)	GBIF Backbone Taxonomy	0.999	Coincidencia exacta	<a href="#">Detalles</a>
Agonostomus monticola Bancroft, 1834	Agonostomus monticola (Bancroft, 1834)	GBIF Backbone Taxonomy	0.997	Coincidencia exacta del nombre canónico	<a href="#">Detalles</a>







Species  
matching

# HERRAMIENTAS PARA EL TRATAMIENTO DE NOMBRES CIENTÍFICOS

The screenshot shows the GBIF Species Lookup tool interface. At the top is a green navigation bar with a leaf icon, links for 'Get data', 'Share', 'Tools', and 'Inside GBIF', a search icon, a chat icon, and a search box containing the text 'katia'. Below the navigation bar, the page title 'TOOLS | LOOK UP' is centered. The main content area has a light blue background and contains the following text: 'Normalize species names from a csv file against the GBIF backbone. The file is expected to be have a column called 'scientificName' and an optional column 'kingdom' and 'id'. Below this text are two buttons: 'SIMPLEEXAMPLE.CSV' and 'ADVANCEDEXAMPLE.CSV'. Underneath these buttons is the text 'SELECT FILE' followed by 'or'. At the bottom of the main content area is a large grey circle with the text 'DROP HERE' inside it.

<https://www.gbif.org/tools/species-lookup>



# HERRAMIENTAS PARA EL TRATAMIENTO DE NOMBRES CIENTÍFICOS



## TOOLS | LOOK UP

OriginalName	PreferedKingdom	MatchType	Confidence	ScientificName (Editable)	Status	Rank	Kingdom	Phylum
Atrichum undulatum (Hedw.) P. Beauv.	any	EXACT	100	<a href="#">Atrichum undulatum Pailsot de Beauvois, 1805</a>	ACCEPTED	species	Plantae	Bryophyta
Aulacomnium androgynum (Hedw.) Schwaegr.	any	EXACT	100	<a href="#">Aulacomnium androgynum Schwaegrichen, 1827</a>	ACCEPTED	species	Plantae	Bryophyta
Aulacomnium palustre (Hedw.) Schwaegr.	any	EXACT	100	<a href="#">Aulacomnium palustre Schwaegrichen, 1827</a>	ACCEPTED	species	Plantae	Bryophyta
Barbilophozia kunzeana (Huebener) Müll. Frib.	any	EXACT	97	<a href="#">Barbilophozia kunzeana (Huebener) Müll.Frib.</a>	ACCEPTED	species	Plantae	Marchantioph
Barbula bolleana	any	EXACT	98	<a href="#">Barbula bolleana Brotherus, 1924</a>	ACCEPTED	species	Plantae	Bryophyta
Barbula convoluta Hedw.	any	EXACT	100	<a href="#">Barbula convoluta Hedwig, 1801</a>	ACCEPTED	species	Plantae	Bryophyta
Barbula convoluta Hedw. var. sardoa Bruch & Schimp	any	EXACT	100	<a href="#">Barbula convoluta var. sardoa Schimp.</a>	ACCEPTED	variety	Plantae	Bryophyta
Barbula unguiculata Hedw.	any	EXACT	100	<a href="#">Barbula unguiculata Hedwig, 1801</a>	ACCEPTED	species	Plantae	Bryophyta
Brachythecium albicans (Hedw.) Schimp.	any	EXACT	100	<a href="#">Brachythecium albicans W. P. Schimper in B.S.G., 1853</a>	ACCEPTED	species	Plantae	Bryophyta
Brachythecium dieckii Roll	any	FUZZY	99	<a href="#">Brachythecium dieckel Röhl, 1897</a>	ACCEPTED	species	Plantae	Bryophyta



# HERRAMIENTAS PARA EL TRATAMIENTO DE DATOS GEOGRÁFICOS

Permiten visualizar puntos en mapas, comprobar coordenadas o convertirlas al formato necesario para la publicación, etc.

## Comprobación

[Info XY \(species link tools\)](#)

[Excel to kml \(Earth point\)](#)

[Google Earth, Google Maps,](#)

[Carto](#)

Sistemas de información geográfica

## Conversión

[Canadensys coordinates conversion](#)

Conversor de coordenadas GBIF.ES

[Geotrans](#)

# HERRAMIENTAS PARA EL TRATAMIENTO DE DATOS GEOGRÁFICOS

id , longitude , latitude (decimal degree)

Jilotla	-98.741583	20.551972
Jilotla	-98.741583	20.551972
Jilotla	-98.741583	20.551972
Jilotla	-98.741583	20.551972
Jilotla	-98.741583	20.551972
Jilotla	-98.741583	20.551972
Jihuico	-98.727305	20.541722
Jihuico	-98.727305	20.541722
Jihuico	-98.727305	20.541722
Jihuico	-98.727305	20.541722

output: HTML ▼

see map

Search

## Results

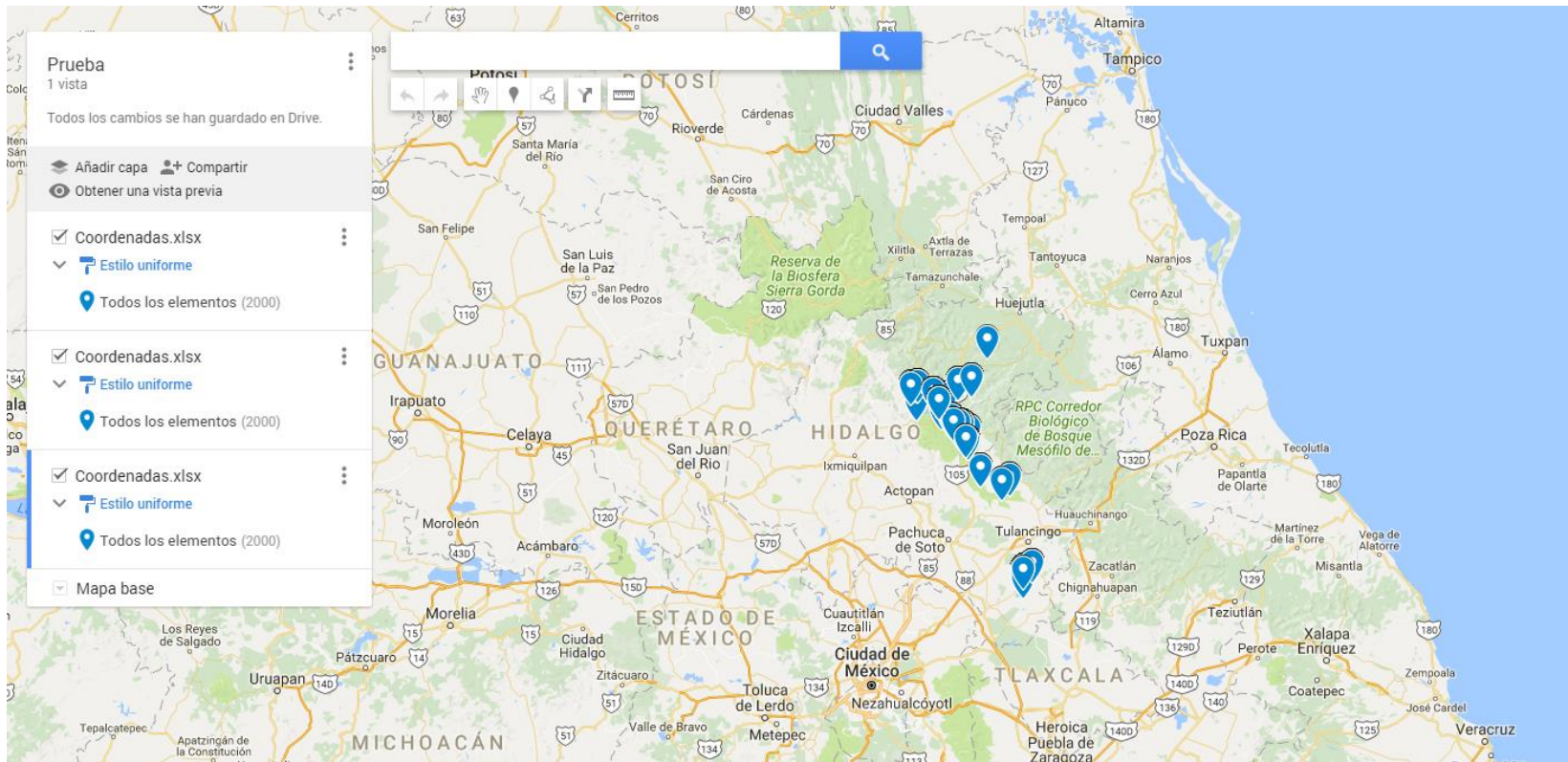
id	longitude	latitude	country	admin1	typeadmin1	admin2	typeadmin2	admin3	typeadmin3	admin4	typeadmin4
Jilotla	-98.741583	20.551972	México	Hidalgo	State	Metztlán	Municipality				
Jilotla	-98.741583	20.551972	México	Hidalgo	State	Metztlán	Municipality				
Jilotla	-98.741583	20.551972	México	Hidalgo	State	Metztlán	Municipality				
Jilotla	-98.741583	20.551972	México	Hidalgo	State	Metztlán	Municipality				
Jilotla	-98.741583	20.551972	México	Hidalgo	State	Metztlán	Municipality				
Jilotla	-98.741583	20.551972	México	Hidalgo	State	Metztlán	Municipality				
Jihuico	-98.727305	20.541722	México	Hidalgo	State	Metztlán	Municipality				
Jihuico	-98.727305	20.541722	México	Hidalgo	State	Metztlán	Municipality				





Google Maps

# HERRAMIENTAS PARA EL TRATAMIENTO DE DATOS GEOGRÁFICOS





# HERRAMIENTAS PARA EL TRATAMIENTO DE DATOS GEOGRÁFICOS

## Coordinate conversion

Use this tool to convert geographic coordinates from DDMSS to decimal degrees. Type coordinate pairs on separate lines or paste latitude and longitude columns from a spreadsheet. Each row may be optionally preceded by an identifier followed by a pipe or tab.

45° 32' 25" N, 129° 40' 31" W

<a href="#">Coordinate conversion</a>
<a href="#">Date parsing</a>
<a href="#">Tools API</a>
<a href="#">About</a>

## Example input

45° 32' 25" N, 129° 40' 31" W

1 | 45.5° N, 129.6° W

2 | 40°26'47"N,74° 0' 21.5022"W

feedback

# HERRAMIENTAS PARA EL TRATAMIENTO DE DATOS GEOGRÁFICOS

Microsoft Access window: Coordenadas a decimal : Base de datos (Access 2007 - 2010) - Microsoft Access

Archivo Inicio Crear Datos externos Herramientas de base de datos Acrobat Herramientas de tabla Campos Tabla

Todos los objetos de Acc... << >> **Coordenadas**

Buscar...

Tablas  
Coordenadas

Consultas  
0\_Convierte coordenadas

Módulos

CatalogNumber	Coordenadas	coordinateUr	decimalLatit	decimalLongitud
30SXH81		7071	38.06	-0.89
30SXJ6243		707	39.226	-1.117
30SXJ69		7071	39.69	-1.08
30SYH01		7071	38.06	-0.66
31TBF61		7071	40.75	0.22
31TBF60		7071	40.66	0.22
30SXH93		7071	38.24	-0.77
30SXJ6243		707	39.226	-1.117
30TYK05		7071	40.22	-0.59
30SXJ83		7071	39.14	-0.86
30SXX62		7071	39.96	-1.07
30SXH7356		707	38.44	-1.012
30SYH07		7071	38.6	-0.65

ČTVRTEK 20 Vendelin ☾

PÁTEK 21 Brigita

SOBOTA 22 Sabina

NEDELE 23 Teodor

# HERRAMIENTAS PARA EL TRATAMIENTO DE LAS FECHAS

Herramientas que permiten la unificación y transformación de distintos formatos de fechas.

[Canadensys date parsing](#)

Función “fechasa” disponible en diversas aplicaciones de GBIF.es

13-VI-1980 → 1980-06-13

13 Junio 1980 → 1980-06-13

13-06-1980 → 1980-06-13





# HERRAMIENTAS PARA EL TRATAMIENTO DE LAS FECHAS

## Date parsing

Use this tool to parse dates into their component parts. Type or paste dates on separate lines, optionally preceded by your own identifier followed by a tab or a pipe.

Jun 13, 2008

[Coordinate conversion](#)

[Date parsing](#)

[Tools API](#)

[About](#)

## Example input

Jun 13, 2008  
15 Jan 2011  
2009 IV 02  
2 VII 1986

1 | 1999/02/24  
2 | 02/17/1921

# HERRAMIENTAS DE VALIDACIÓN Y DEPURACIÓN



*Refine*<sup>OPEN</sup> 

Darwin Test



Darwin Core Archive Validator



## ¿Qué es?

Una herramienta de manipulación de datos

Faceting  
Clustering  
Reconciling



DARWIN\_TEST: Validación de datos exportados a GBIF con Darwincorev2



Data validation and geographic coordinates generalization for Darwin Core datasets <http://www.gbif.es/>

✂ ⚙ ⏻

## 1. Seleccionar

Tipo de datos a validar:

DarwinCore 1.2 (mdb)     DarwinCore 1.4 (mdb)     DarwinCore Archive (zip)

Seleccione el origen de los datos a validar

C:\Darwintest\_3.4\ColeccionesConCaracteresAnomalos\DarwinCore2\_UPNA-H\_20171106\_GBIF.mdb ... Aceptar

## 2. Validación de datos

### 2.1. Metadatos

Detectar posibles errores

↓

Visualización y chequeo de datos de DwC

Detectar caracteres ASCII anómalos

↗

↘

Actualizar tabla DwC vinculada

Crear tabla DwC para corregir base de datos original ?

# HERRAMIENTAS DE VALIDACIÓN Y DEPURACIÓN



Data validator



Get data

Share

Tools

Inside GBIF



katia

TOOLS | DATA VALIDATOR

This an early access version. Please report issues and feedback [here](#).

NEW DATA VALIDATION

ABOUT

SELECT FILE

or

DROP HERE

or Fetch file from location:

SUBMIT

File size limit: 100 mb

<https://www.gbif.org/tools/data-validator>



# HERRAMIENTAS DE VALIDACIÓN Y DEPURACIÓN



Data validator



Get data Share Tools Inside GBIF



katia

SUMMARY

DARWIN CORE EXTENSIONS

NEW VALIDATION

## ● The file can be indexed by GBIF

Some issues were detected by the validator:

GBIF Occurrence  
Interpretation

Recorded date mismatch Taxon match fuzzy Taxon match higherrank  
Coordinate reprojected

File Format: Tabular File (.csv, .tsv)

Media Type: text/plain

Core Row Type: Darwin Core Occurrence

Extensions: 0

This report has been written to <https://www.gbif.org/tools/data-validator/1509439657203> It was generated in 10 minutes and will be deleted after one month. Until then you can revisit the report at your convenience.

### Core

CARIMED\_2016\_OCCURRENCE.TXT

- Term Frequency
- Validation Issues

## CARIMED\_2016\_Occurrence.txt

Row Type: Darwin Core Occurrence

Number Of Lines: 5.979

Number Of Rows With Interpreted Taxon: 5.979

Number Of Interpreted Dates: 5.979

Number Of Interpreted Coordinates: 5.979

### Term Frequency

Term	Count	Percentage	Interpreted
dwc:eventID	5.979	100%	
dcterms:type	5.979	100%	
dcterms:modified	5.979	100%	
dcterms:issued	5.979	100%	

<https://www.gbif.org/tools/data-validator>



Regional capacity enhancement to Latin  
America by establishing Chile's node

**Gracias por la atención**

UNIDAD DE COORDINACIÓN GBIF.ES  
REAL JARDÍN BOTÁNICO-CSIC  
katia@gbif.es