



Herramientas para la limpieza y validación de datos

Leonardo Buitrago



August 26, 2026

Agenda

- Bienvenida y objetivo
- Repaso relámpago: estandarización en Darwin Core
- Conceptos de limpieza de datos
- Herramientas de gestión de datos
- Check de comprensión #2
- Exploración de herramientas
 - Formato de Fechas (*Canadensys – Date Parsing*)
 - Conversión de Coordenadas (*Canadensys – Coordinate conversion*)
 - Validación de nombres científicos (*GBIF Species Matching*)
- *OpenRefine*: el poder para limpiar datos de biodiversidad
- Ejercicio práctico: implementación de herramientas
- Puesta en común - check final y cierre

TALLERES DE CAPACITACIÓN

Taller 1: Estandarización y documentación de datos y metadatos | 10 AGO

Taller 2: Herramientas para la limpieza y validación de datos | 26 AGO

Taller 3: Citación, atribución y reconocimiento | 16 SEP

Objetivo

Conocer los principios de la limpieza de datos y ver en acción las herramientas para validar y estandarizar datos de biodiversidad.



<https://tinyurl.com/ybcadhjc>

Ejercicio práctico - Estandarización



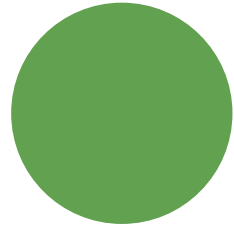
Conceptos de limpieza de datos

Conceptos clave

- ¿Qué es la limpieza de datos?
- Principios de la limpieza de datos



¿Qué es la limpieza de datos?



“Un proceso utilizado para determinar datos inexactos, incompletos o poco razonables y mejorar su calidad corrigiendo los errores y omisiones detectados.”
– Arthur Chapman



El marco de la limpieza de datos



- Definir y determinar los tipos de error;
- Buscar e identificar los casos de error;
- Corregir los errores;
- Documentar los casos y los tipos de error; y,
- Modificar los procedimientos de entrada de datos para reducir errores futuros.

(de Maletic & Marcus, 2000)

¿Por qué limpiar los datos?

Los errores son comunes

Los errores son esperables



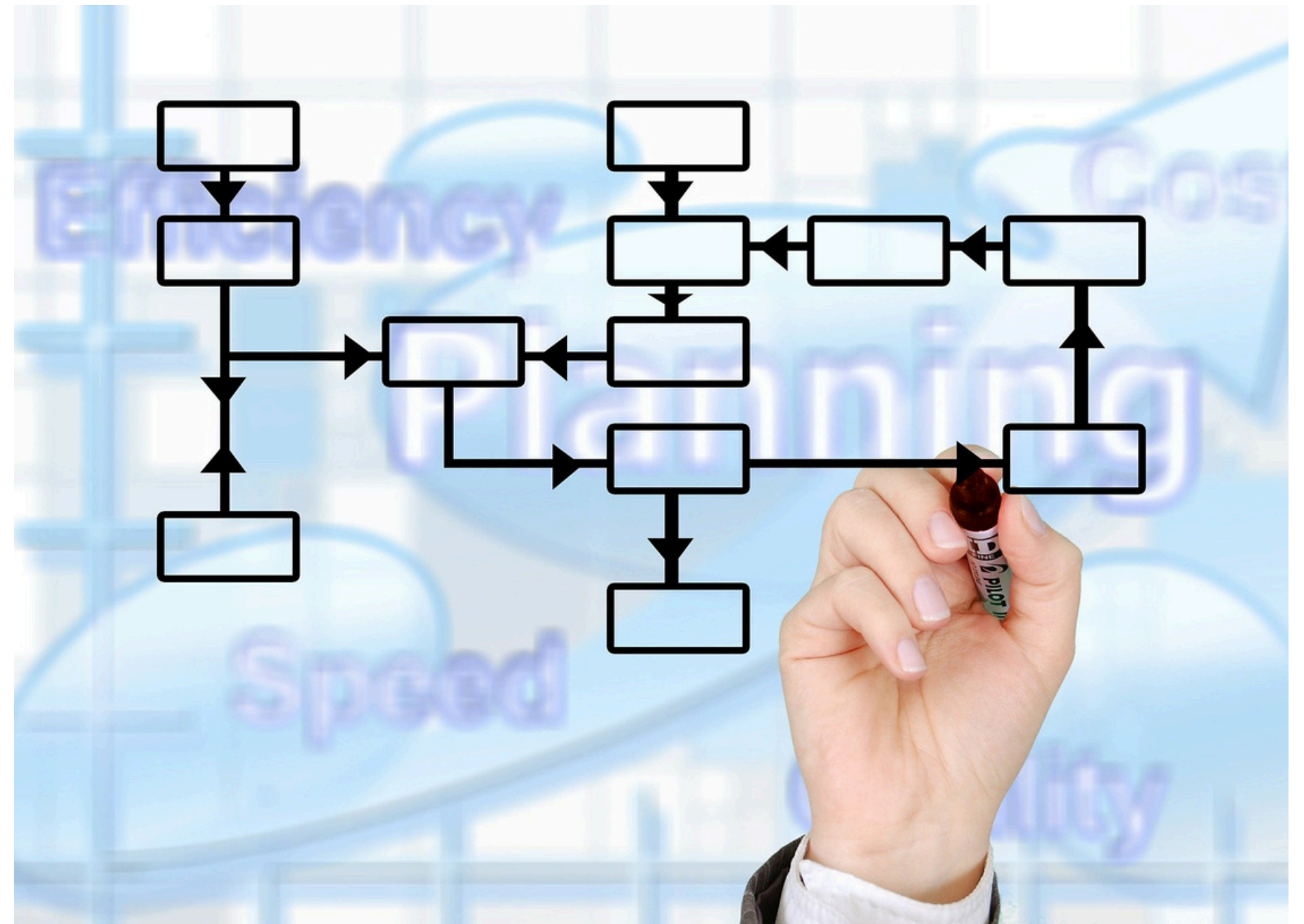
Principios de la limpieza de datos

12 principios clave



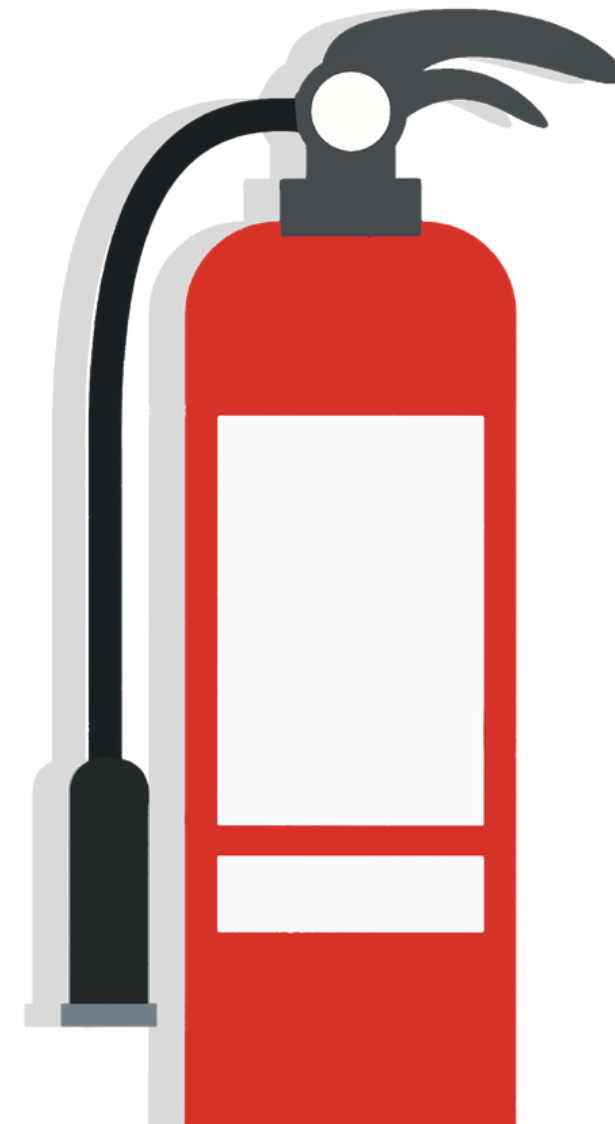
Principios de la limpieza de datos

1. **Planificación**
2. **Organización**
3. Prevención
4. Responsabilidad
5. Alianzas
6. Priorización
7. Medidas de desempeño
8. Optimización
9. Retroalimentación
10. Capacitación
11. Transparencia
12. Documentación



Principios de la limpieza de datos

1. Planificación
2. Organización
3. **Prevención**
4. **Responsabilidad**
5. Alianzas
6. Priorización
7. Medidas de desempeño
8. Optimización
9. Retroalimentación
10. Capacitación
11. Transparencia
12. Documentación



Principios de la limpieza de datos

1. Planificación
2. Organización
3. Prevención
4. Responsabilidad
5. **Alianzas**
6. **priorización**
7. Medidas de desempeño
8. Optimización
9. Retroalimentación
10. Capacitación
11. Transparencia
12. Documentación



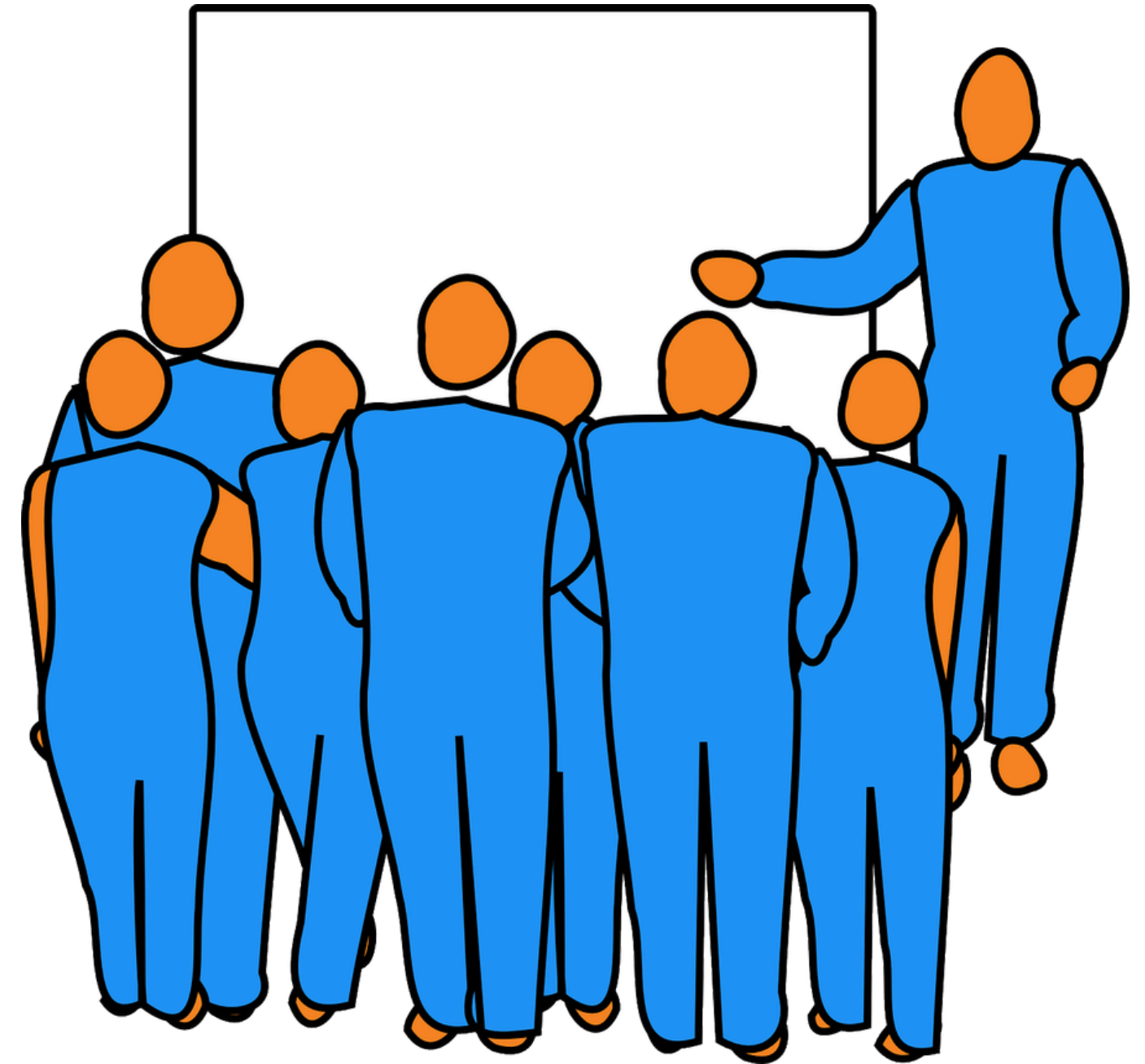
Principios de la limpieza de datos

1. Planificación
2. Organización
3. Prevención
4. Responsabilidad
5. Alianzas
6. Priorización
7. **Medidas de desempeño**
8. **Optimización**
9. Retroalimentación
10. Capacitación
11. Transparencia
12. Documentación



Principios de la limpieza de datos

1. Planificación
2. Organización
3. Prevención
4. Responsabilidad
5. Alianzas
6. Priorización
7. Medidas de desempeño
8. Optimización
9. **Retroalimentación**
10. **capacitación**
11. Transparencia
12. Documentación



Principios de la limpieza de datos

1. Planificación
2. Organización
3. Prevención
4. Responsabilidad
5. Alianzas
6. Priorización
7. Medidas de desempeño
8. Optimización
9. Retroalimentación
10. Capacitación
11. **Transparencia**
12. **documentación**



An aerial photograph of a coastal region. A dark river flows from the top left towards the bottom right. The riverbanks are lined with dense, vibrant green vegetation. On the left side of the river, there are several buildings, some with corrugated metal roofs, and a small cluster of boats. On the right side, a road with a yellow center line runs parallel to the river, with more buildings and trees further inland. The overall scene depicts a mix of natural environment and human settlement.

Herramientas de gestión de datos

La importancia del mantenimiento de datos

Los datos del mundo real son
desordenados

Siempre se necesitan transformaciones

Al transformar se puede perder
información

Conserve el original

¡Automatic!



¿Qué considerar al elegir una herramienta?

Precio, disponibilidad y licencias

Facilidad de uso

Documentación y soporte

Flexibilidad

Rendimiento

Requisitos técnicos

Automatización



¿Qué considerar al elegir una herramienta?



¡USE LO QUE LE FUNCIONE!

Arme su propio conjunto de
herramientas

Consideraciones técnicas básicas

- Use un formato de intercambio
- Valores separados por comas (CSV) o por tabuladores (TSV)
- Codificación: UTF-8
- Documente los formatos y opciones usados al crear un archivo.
- Pruebe diferentes opciones y verifique los resultados.



Herramientas – datos de biodiversidad

 Consultar ▾ Compartir ▾ Recursos ▾ Comunidad ▾ Noticias Acerca de ▾  Buscar



Calidad de datos

Documentos, herramientas y cursos para la validación y limpieza de datos que permiten asegurar su uso por una amplia variedad de audiencias.

La **calidad de datos** se puede definir como una característica esencial y necesaria para que los datos sean "adecuados para su uso" o tengan un "uso potencial valioso" (Chapman, 2005).

Cuando los datos se utilizan sin tener una visión crítica sobre los posibles errores que contienen, los resultados de sus análisis pueden conducir a conclusiones erróneas y decisiones imprudentes basadas en evidencia poco confiable. Los datos tendrán calidad cuando la información derivada de los mismos representa correctamente el mundo real (los hechos).

CONTENIDOS

- Herramientas
 - Limpieza y estructuración de datos
 - Limpieza de fechas
 - Validación de nombres y taxonomía
 - Limpieza geográfica

<https://biodiversidad.co/recursos/calidad-de-datos/>



Herramientas – editores de texto

MAC OS

- BBEdit (TextWrangler)
- Atom

WINDOWS

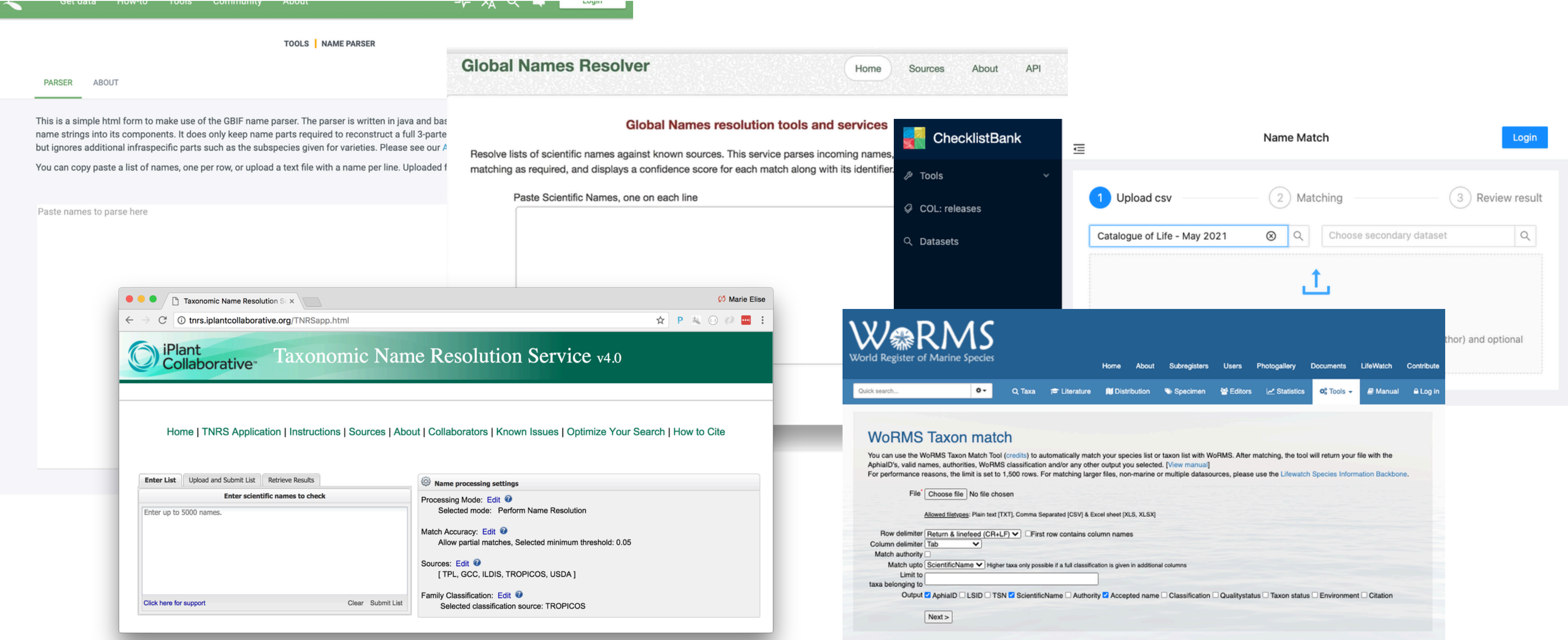
- Notepad++
- Sublime
- EmEditor

UNIX / LINUX

- Emacs
- vi



Herramientas – taxonomía



GBIF Species Matching: <https://www.gbif.org/tools/species-lookup> iPlant TNRS: <http://tnrs.iplantcollaborative.org/>
Global Names Resolver: <http://resolver.globalnames.org> WoRMS: <https://www.marinespecies.org/aphia.php?p=match>
CoL Checklist Bank Name Match: <https://www.checklistbank.org/tools/name-match>

Transformar coordenadas GMS a formato decimal

Grados Minutos Segundos Hemisferio → grados decimales

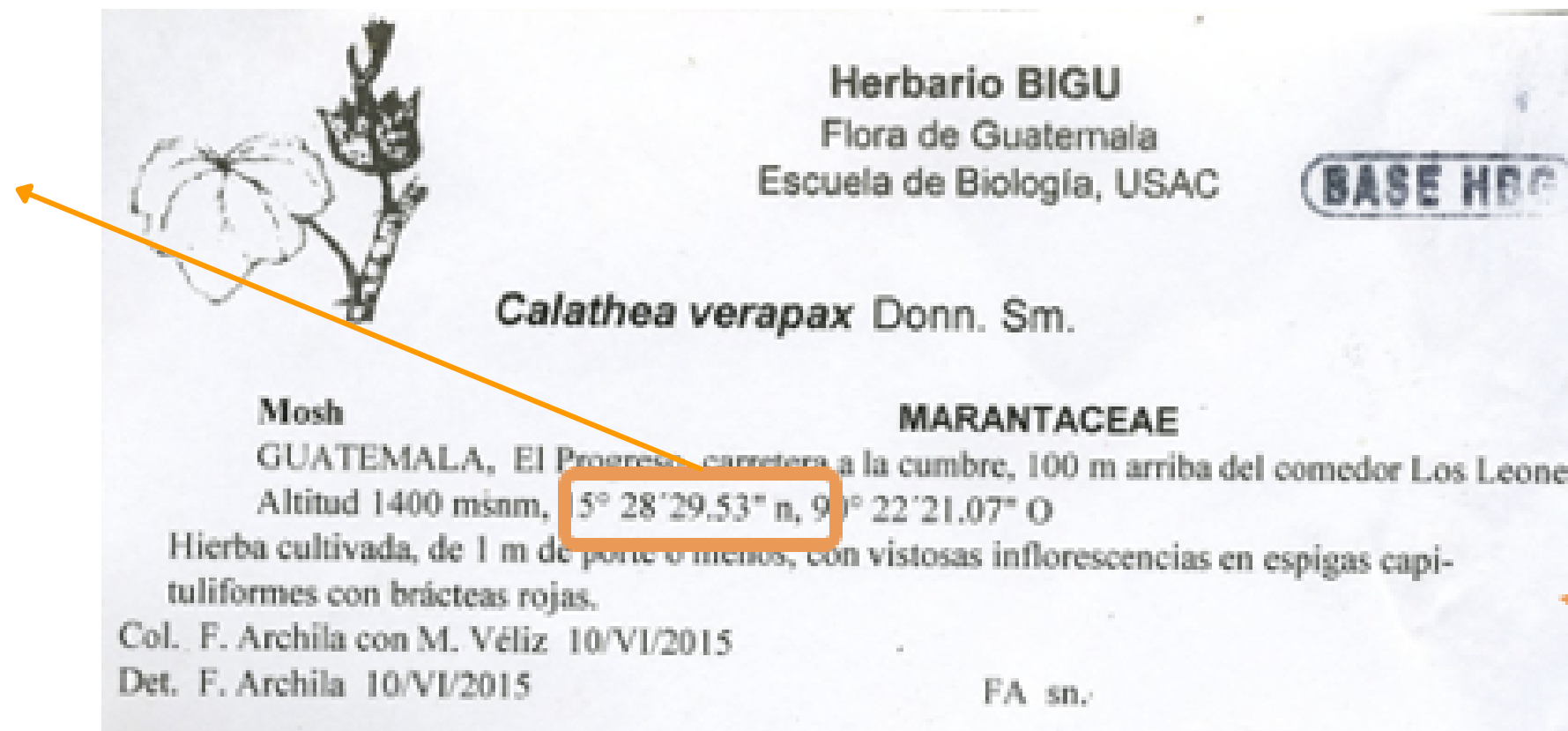
$$\text{GD} = (\text{G} + \text{M}/60 + \text{S}/3600) * [\text{H}]$$

Hemisferio longitudinal: oeste = -1; este = 1

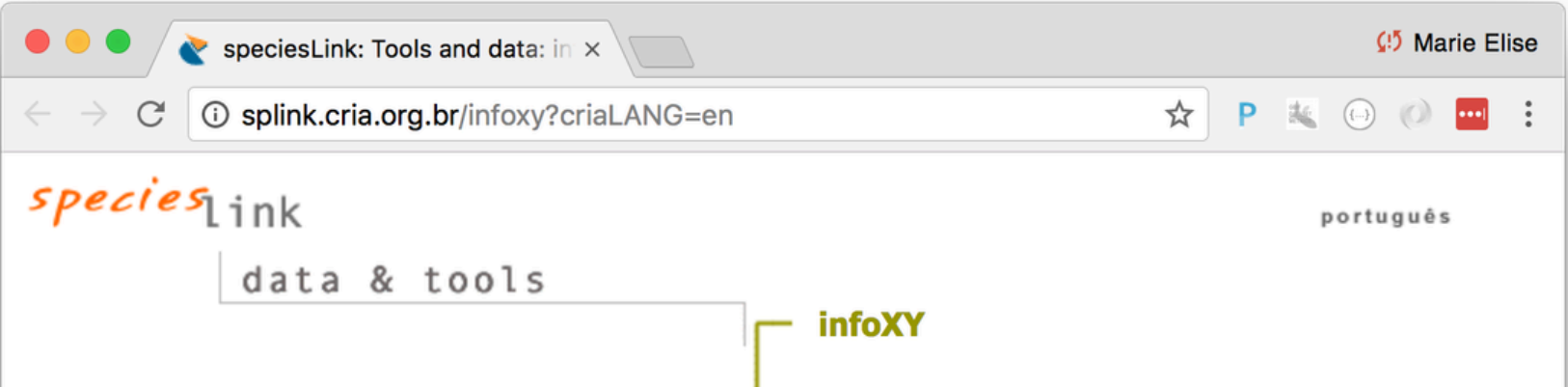
$$\text{GD} = (15 + 28/60 + 29.53/3600) * 1$$

$$\text{GD} = 15.47487$$

15° 28' 29.53" n



Herramientas - georreferenciación

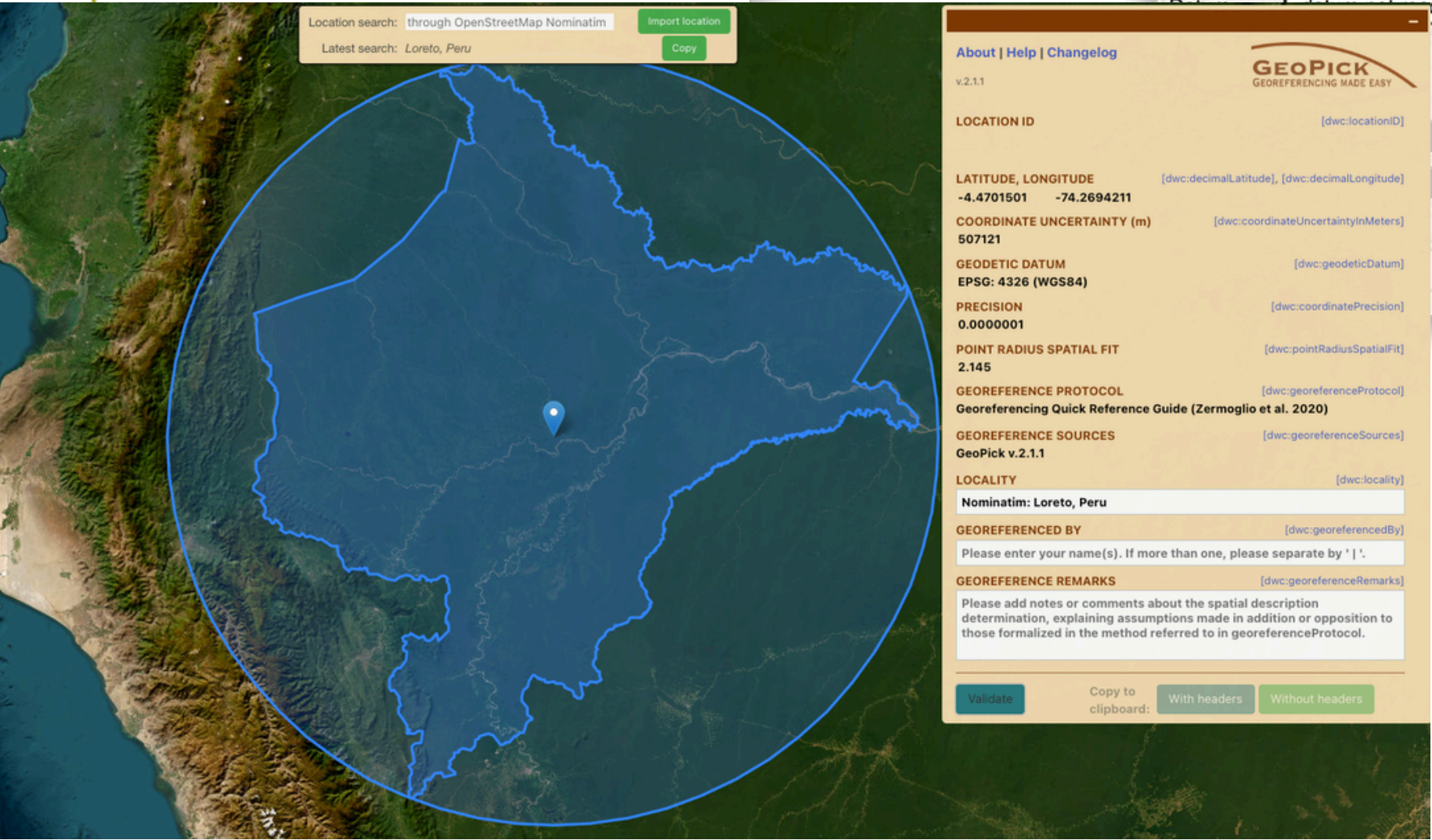


This tool was developed by CRIA with the aim of helping biologists to georeference geographic coordinates the tool returns information about the point and the name of the municipality or district. If the point is in the country, the name of the country. The database used is GADM-Global Administrative Areas. infoXY version 2.0

id , longitude , latitude (decimal degree)

exemploID, -41.5081, -20.7102

search
indicators
dataCleaning
networkManager



English

Georeferencing Calculator

Calculation Type: Coordinates and error - enter the Lat/Long for the named place or starting point

Locality Type: Distance at a heading (e.g., 10 mi E (by air) Bakersfield)

Step 3) Enter all of the parameters for the locality.

Coordinate Source: gazetteer

Coordinate System: decimal degrees

Latitude:

Longitude:

Direction: nearest degree

Offset Distance:

Extent of Named Place:

Measurement Error:

Distance Units: km

Distance Precision: 100 km

Decimal Longitude:

Maximum Error Distance:

Calculate Promote

km mm 1:24000 km

Copyright 2015 Regents of the University of California

InfoXY – <http://splink.cria.org.br/tools?criaLANG=en>
GeoPick: <https://geopick.gbif.org/>
Georeferencing Calculator - <https://georeferencing.org/calculator/>
Google Maps - <https://maps.google.com/>

Herramientas – conversión de coordenadas y fechas

Coordinate conversion results

data.canadensys.net/tools/coordinates

Canadensys data | community

explorer repository tools vascan

Coordinate conversion results

id	original	decimalLatitude	decimalLongitude
1	45.5° N, 129.6° W	45.5	-129.6
2	40°26'47"N, 74° 0' 21.5022"W	40.4463889	-74.0059728

Coordinate conversion

Date parsing

Tools API

About

feedback

Date parsing results

data.canadensys.net/tools/dates

Canadensys data | community

explorer repository tools vascan

Date parsing results

original	year	month	day	ISO 8601
Jun 13, 2008	2008	6	13	2008-06-13
15 Jan 2011	2011	1	15	2011-01-15
2009 IV 02	2009	4	2	2009-04-02
2 VII 1986	1986	7	2	1986-07-02

Coordinate conversion

Date parsing

Tools API

About

feedback

Herramienta de conversion de coordenadas - <http://data.canadensys.net/tools/coordinates>
Herramienta de conversión de fechas - <http://data.canadensys.net/tools/dates>

Herramientas – línea de comandos



Expresiones regulares (potentes, en todas las plataformas)

Scripts (p. ej., Perl, Python)

Bash, DOS (scripts de shell)

R (línea de comandos), RStudio (interfaz gráfica)

Ideas finales



- GBIF Name Parser

<https://www.gbif.org/tools/name-parser>

- Global Names Resolver

<http://resolver.globalnames.org>

- COL ChecklistBank name match

<https://www.checklistbank.org/tools/name-match>

- TNRS

<http://tnrs.iplantcollaborative.org/index.html>

- WoRMS

<https://www.marinespecies.org/aphia.php?p=match>

- InfoXY

<http://splink.cria.org.br/infoxy?criaLANG=en>

- Georeferencing Calculator

<https://georeferencing.org/calculator/>

- Canadensys – herramientas

<http://data.canadensys.net/tools>

¡Gracias!

<https://training.gbif.org/>



europa.eu

<https://www.menti.com/al7h5c3zpg13>



Check de comprensión #2





Exploración de Herramientas

En esta sección aprenderemos sobre:



Formato de Fechas

(Canadensys – Date Parsing)

<https://data.canadensys.net/tools/dates?lang=en>



Conversión de Coordenadas

(Canadensys – Coordinate conversion)

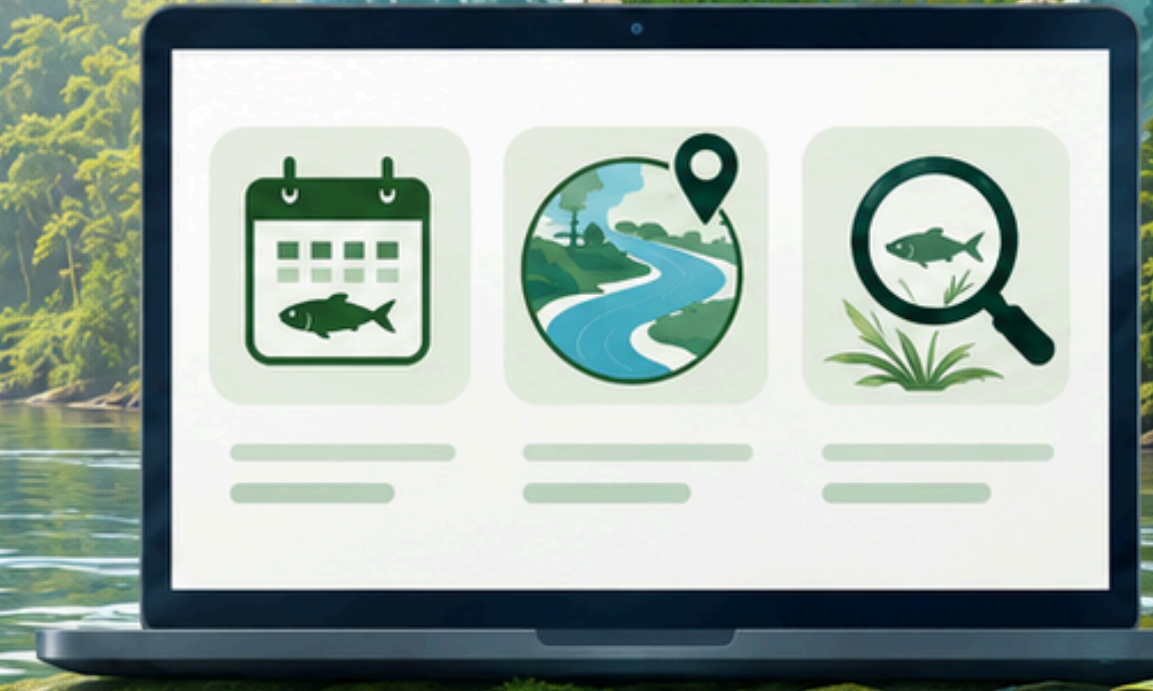
<https://data.canadensys.net/tools/coordinates?lang=en>



Validación de nombres científicos

(GBIF Species Matching)

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





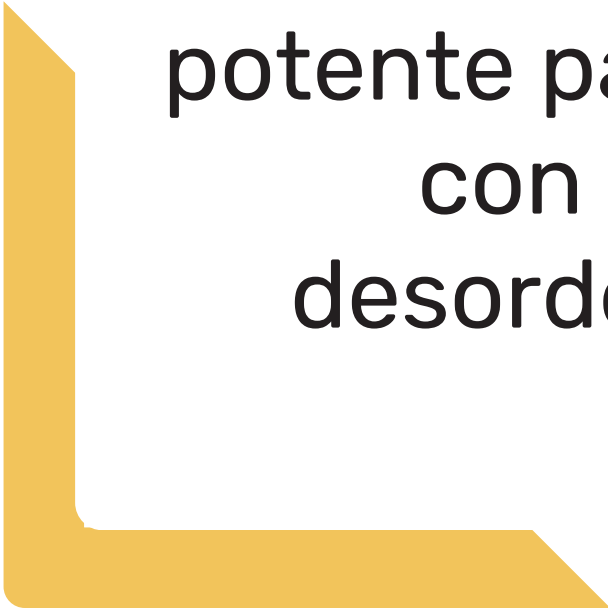
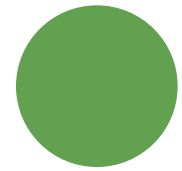
**OpenRefine: el poder para limpiar
datos de biodiversidad**

OpenRefine



- Limpieza de datos
- Estandarización
- <https://openrefine.org/>

¿Qué es OpenRefine?

A yellow L-shaped graphic with a diagonal cut on the top-left and bottom-right corners, positioned to the left of the quote.

“Una herramienta
potente para trabajar
con datos
desordenados.”

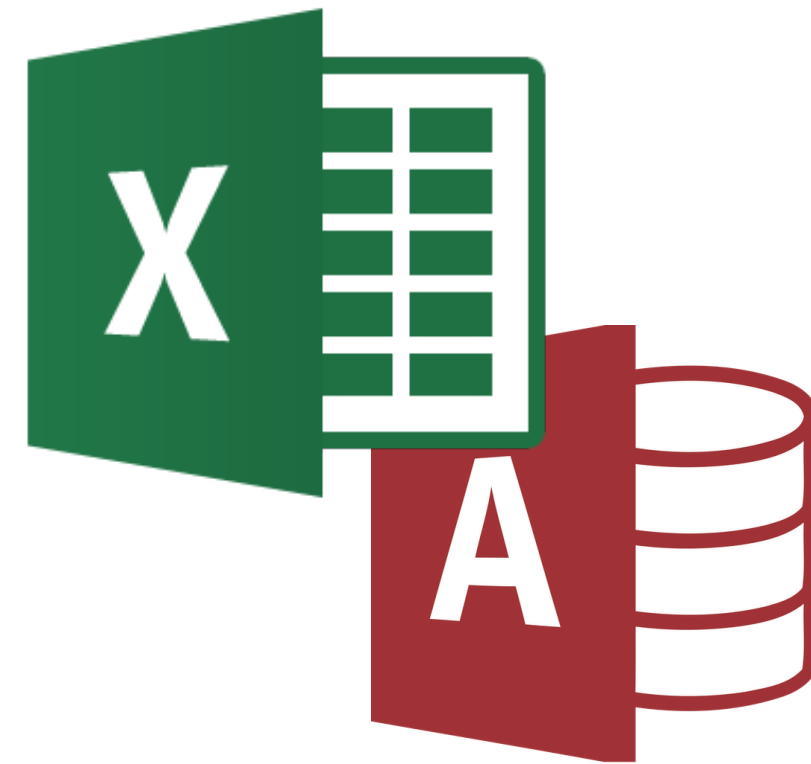
- Facetas (faceting)
- Agrupamiento (clustering)
- Reconciliación (reconciling)

Lo que OpenRefine no es

No es una base de datos



No es una experiencia tipo Excel



Comparación de características



Base de datos



Excel



OpenRefine

Características de OpenRefine

- Visualización interactiva
- Edición de múltiples celdas
- Transformación sencilla de los datos
- Facetado / filtros
- Uso de expresiones regulares y API
- Deshacer ilimitado
- Aplicación de escritorio (sin conexión) con extensiones
- Muchos formatos de importación y exportación
- Gran comunidad de desarrolladores
- Gratuita



Gracias

<https://training.gbif.org/>



europa.eu



<https://tinyurl.com/3z9ank43>

Ejercicio práctico - Limpieza