

Calidad de Datos

Guía de herramientas para mejorar
los datos primarios de biodiversidad



Versión 1.0
2015



Sistema de Información sobre Biodiversidad de Colombia – SiB

Cítese como:

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, XX pp.

ISBN: pendiente

URI persistente: <http://www.sibcolombia.net/repositorio-de-documentos>

Idioma: Español

Copyright © Sistema de Información sobre Biodiversidad de Colombia – SiB. 2015



Este documento se publica bajo una licencia
[Creative Commons Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

Créditos de las imágenes de la portada: EC-SiB Colombia. Rana.

Control del documento:

Versión	Descripción	Fecha de publicación	Autor(es)
1.0		Julio de 2015	Dairo Escobar
			Nestor Beltran
			Leonardo Buitrago
			Camila Plata
			Marcela Delgado

Acerca de SiB

El SiB Colombia es una iniciativa de país que tiene como propósito brindar acceso libre a información sobre la diversidad biológica del país para la construcción de una sociedad sostenible. Esta iniciativa facilita la publicación en línea de datos sobre biodiversidad y su acceso a una amplia variedad de audiencias, apoyando de forma oportuna y eficiente la gestión integral de la biodiversidad.

El SiB Colombia es liderado por un Comité Directivo (CD-SiB), el cual está conformado por el Ministerio de Ambiente y Desarrollo Sostenible, los 5 institutos de investigación del SINA (IAvH, INVEMAR, SINCHI, IIAP E IDEAM) y la Universidad Nacional de Colombia. El CD-SiB se apoya en un Comité Técnico (CT-SiB), grupos de trabajo Ad Hoc para temas específicos y por un Equipo Coordinador (EC-SiB), el cual cumple las funciones de secretaría técnica, que acoge e implementa las recomendaciones del CD-SiB.

En este contexto, el EC-SiB, que por mandato de ley se encuentra alojada en el Instituto Humboldt, es el encargado de asegurar la disponibilidad de información sobre la biodiversidad de Colombia, principalmente a través de procesos de gestión y coordinación inter-institucional en el marco de esta iniciativa nacional.

Introducción

Este documento es una recopilación de las diversas herramientas y prácticas utilizadas en el taller realizado, estas intentan facilitar el proceso de dar calidad a los datos primarios sobre biodiversidad a través de distintas metodologías. Se espera que este texto sirva como base para optimizar los procesos alrededor de la estandarización de datos biodiversidad, empleando como base el estándar *Darwin Core* establecido por el Taxonomic Databases Working Group (TDWG).

La guía se divide en diferentes secciones y va desde el manejo de una herramienta básica para la estructuración de información como Excel, pasando por recursos web para realizar la validación taxonómica y geográfica de los datos de biodiversidad. Finalmente se muestran servicios web que optimizan la visualización de la información (CartoDB, VesPer) y herramientas de validación de datos más robustas y completas como *Open Refine* o *Darwin Test*.

Cada sección presenta una introducción al tema a tratar y una aproximación con el origen y propósito de cada herramienta. En todas ellas se expone un caso práctico, con un paso a paso, que permite al lector desarrollar ejercicios ejemplo para tener un acercamiento más tangible con los recursos y alcances de cada una. La idea final, es que lo aplicado en la práctica de cada herramienta sea escalable a los datos propios de cada investigador y que la limpieza y calidad de datos de biodiversidad sea una tarea mucho más sencilla.

Índice

[Acerca de SiB](#)

[Introducción](#)

[Estándar Darwin Core](#)

[Estandarización y limpieza de datos en Excel](#)

[Ejercicio 1 - Función LIMPIAR](#)

[Ejercicio 2 - Función ESPACIOS](#)

[Ejercicio 3 - Función ESPACIOS\(SUSTITUIR\)](#)

[Ejercicio 4 - Función SUSTITUIR](#)

[Ejercicio 5 - Función CONCATENAR \(1\)](#)

[Ejercicio 6 - Función CONCATENAR \(2\)](#)

[Ejercicio 7 - Función SI](#)

[Recomendaciones adicionales](#)

[Estandarización de Fechas](#)

[Date Parsing](#)

[Caso Práctico](#)

[Estandarización y validación taxonómica](#)

[ECAT Name Parser](#)

[Resolución Taxonómica y de LSIDs](#)

[Taxonomic Name Resolution Service \(TNRS\)](#)

[Estandarización y Validación de Información Geográfica](#)

[Conversión de Coordenadas - Canadensys](#)

[GeoValidador SiB](#)

[CartoDB](#)

[VeSPER](#)

[Darwin Test](#)

[Open Refine](#)

[INSTALACIÓN](#)

[EN WINDOWS](#)

[EN MAC OS](#)

[CONVENCIONES DE LOS EJERCICIOS](#)

[USO BÁSICO](#)

[CARGA DE DATOS](#)

[FACETING](#)

[FILTERING](#)

[CLUSTERING](#)

[EXPORTACIÓN](#)

[USO AVANZADO](#)

[USO DEL API DE EOL](#)

[ENLACES ADICIONALES](#)

Estándar Darwin Core

El SiB Colombia adopta el estándar [Darwin Core](#) (DwC), una extensión del estándar [Dublin Core](#), el cual permite el intercambio efectivo de información sobre biodiversidad a nivel global. Al contar con el DwC como un lenguaje común, se facilita la estandarización semántica y sintáctica de los contenidos. El estándar está compuesto por un grupo de 175 elementos que permiten estructurar y estandarizar datos sobre registros biológicos, listas de especies y eventos de muestreo. Puede obtener más información consultando la [Guía de Referencia Rápida](#) del DwC ajustada por el SiB.

Estandarización y limpieza de datos en Excel

Excel es una herramienta desarrollada por Microsoft en la cual se puede escribir, almacenar, manipular, calcular y organizar todo tipo de información numérica o de texto. Es la hoja de cálculo electrónica más extendida y usada a nivel global. En el campo de la biología es usada ampliamente para estructurar y estandarizar datos de biodiversidad gracias a sus ventajas para interconectar unas celdas con otras mediante el uso de funciones o reglas, de tal forma que si cambia el valor de una celda, automáticamente la hoja de cálculo electrónica recalculará y actualizará los valores de las otras celdas. Este tipo de funcionalidades permite realizar cambios y ajustes en grandes conjuntos de información ahorrando tiempo y optimizando procesos en la normalización de datos.

Caso práctico

Este ejercicio se centra en conocer algunas de las funciones básicas del programa Excel, las cuales facilitan la corrección de errores y en general el manejo y estandarización de los conjuntos de datos. Se espera que algunas de las tareas rutinarias sean realizadas de manera más rápida y eficiente.

Limpieza de datos

Al aplicar las funciones propias de Excel como **Limpiar**, **Espacios** y **Sustituir**, se eliminan todos los caracteres no imprimibles existentes en el texto. Aunque muchas veces no son evidentes, en el momento de cargar el archivo a una instancia (por ejemplo a una herramienta de publicación de datos de biodiversidad como el *Integrated Publishing Toolkit* - [IPT](#), empleado por el SiB Colombia) o realizar filtros y agrupaciones de los registros, pueden presentar inconvenientes en la lectura de la información. De esta forma, se recomienda que estas funciones sean el primer paso en el proceso de limpieza de datos (Tabla 1).

Objetivo: Conocer y aplicar diferentes funciones de la aplicación Excel que permiten limpiar y estandarizar los caracteres y formatos en un conjuntos de datos.

Tabla 1. Resumen de las funciones. Dependiendo de la configuración de Excel, el separador entre textos puede variar de coma (,) a punto y coma (;). Estos casos se indican con (*). La sintaxis de la función se indica en español e inglés.

FUNCIÓN	DESCRIPCIÓN GENERAL
Limpiar	Elimina todos los caracteres no imprimibles del texto. =LIMPIAR(texto) =CLEAN(text)
Espacios	Elimina todos los espacios del texto [espacios ASCII de 7 bits (valor 32)] excepto los espacios individuales entre palabras y los espacios de no separación (ASCII 160). =ESPACIOS(texto) =TRIM(text) Para eliminar este último tipo de espacios utilizar la siguiente fórmula =ESPACIOS(SUSTITUIR(A1, CARACTER(160), CARACTER(32))) =TRIM(SUBSTITUTE(A1, CHAR(160), CHAR(32))) A1 corresponde a la celda objetivo
Sustituir	Elimina el código para la tecla borrar (ASCII 127) =SUSTITUIR(A1, CARACTER(127), "") =SUBSTITUTE(A1, CHAR(127), "")

A1 corresponde a la celda objetivo

Concatenar*	<p>Une varios elementos del texto en uno solo =CONCATENAR(texto1,texto2,...) =CONCATENATE(text1,text2,...)</p> <p>Para concatenar dejando espacios entre palabras utilizar la siguiente fórmula =CONCATENAR(texto1," ",texto2) =CONCATENATE(text1," ",text2)</p> <p>Para concatenar utilizando como separador signos (, ; :) utilizar la siguiente fórmula ubicando el signo entre las comillas =CONCATENAR(texto1,";",texto2) =CONCATENATE(text1,";",text2)</p>
Si*	<p>Comprueba si se cumple una condición y devuelve un valor si se evalúa como verdadero y otro valor si se evalúa como falso. Es útil para comparar la igualdad entre dos columnas =SI(prueba_lógica,valor_si_verdadero,valor_si_falso) =IF(logical_test,value_if_true,value_if_false)</p>

Ejercicio 1 - Función LIMPIAR

Paso 1

Descargue en el escritorio de su equipo el archivo Excel de este vínculo: <http://goo.gl/e3VoNw> . Ahora ábralo y detalle la información que contiene el archivo. Usted encontrará un archivo con siete pestañas, donde cada una corresponde a un ejercicio aplicativo de las funciones de excel

Paso 2

En la hoja "Ejemplo (1)" vaya a la columna "**Nombre Científico**" y cree una columna nueva que puede llamar "**Función Limpiar**"

Paso 3

Frente al nombre científico (celda A2), ingrese la función LIMPIAR tal y como se muestra en la en la Figura 1 =LIMPIAR(A2).

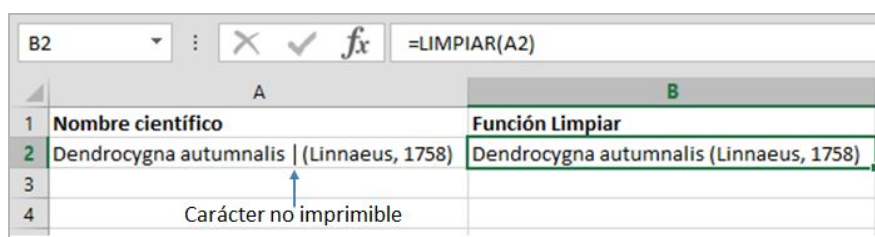


Figura 1. Ejemplo de uso de la función LIMPIAR, la barra de fórmulas muestra el comando que se debe ingresar.

Paso 4

Arrastre el contenido de la celda (esquina inferior derecha) hasta empatar todos los registros de la columna "Nombre Científico".

Paso 5

Reemplace el contenido original de la columna "**Nombre Científico**", copiando los valores de la columna que creó. **Nota:** El contenido debe ser pegado como "valores" (Pegar > Valores). De esta forma, evitamos que la información se altere en un futuro.

Paso 6

Ahora elimine la columna que había creado como "**Función Limpiar**" y guarde los cambios realizados para continuar al siguiente ejercicio.

Ejercicio 2 - Función **ESPACIOS**

Paso 1

Inicie el ejercicio en el mismo archivo que aplicamos la función **LIMPIAR**.

Paso 2

En la hoja "Ejemplo (2)" vaya nuevamente a la columna "**Nombre Científico**" y cree una columna nueva que puede llamar "**Función Espacios**".

Paso 3

Frente al nombre científico (celda A2), la cual presenta a propósito espacios de más, ingrese la función **ESPACIOS** tal y como se muestra en la en la Figura 2 **=ESPACIOS(A2)**.

	A	B
1	Nombre científico	Función Espacios
2	Dendrocygna autumnalis (Linnaeus, 1758)	Dendrocygna autumnalis (Linnaeus, 1758)
3		

Figura 2. Ejemplo de uso de la función **ESPACIOS**, la barra de fórmulas muestra el comando que se debe ingresar.

Paso 4

Arrastre el contenido de la celda (esquina inferior derecha) hasta empatar los registros de la columna "Nombre Científico".

Paso 5

Reemplace el contenido original de la columna "**Nombre Científico**", copiando los valores de la columna que creó. **Nota:** El contenido debe ser pegado como "valores" (Pegar > Valores). De esta forma, evitamos que la información se altere en un futuro.

Paso 6

Ahora elimine la columna que había creado como "**Función Espacios**" y guarde los cambios realizados para continuar con el ejercicio 3.

Ejercicio 3 - Función **ESPACIOS(SUSTITUIR)**

Paso 1

Inicie el ejercicio en el mismo archivo trabajado en los Ejercicios 1 y 2.

Paso 2

En la hoja "Ejemplo (3)" vaya nuevamente a la columna "**Nombre Científico**" y cree una columna nueva que puede llamar "**Función Espacios**".

Paso 3

Frente al nombre científico (celda A2), ingrese la función **ESPACIOS(SUSTITUIR)** tal y como se muestra en la en la Figura 3 **=ESPACIOS(SUSTITUIR(A2,CARACTER(160),CARACTER(32)))**.

B2		fx		=ESPACIOS(SUSTITUIR(A2,CARACTER(160),CARACTER(32)))	
	A		B		
1	Nombre científico		Función Espacios		
2	Dendrocygna autumnalis (Linnaeus, 1758)		Dendrocygna autumnalis (Linnaeus, 1758)		

Figura 3. Ejemplo de uso de la función **ESPACIOS(SUSTITUIR)**, la barra de fórmulas muestra el comando que se debe ingresar.

Paso 4

Arrastre el contenido de la celda (esquina inferior derecha) hasta emparejar los registros de la columna "Nombre Científico".

Paso 5

Reemplace el contenido original de la columna "**Nombre Científico**", copiando los valores de la columna que creó. **Nota:** El contenido debe ser pegado como "valores" (Pegar > Valores). De esta forma, evitamos que la información se altere en un futuro.

Paso 6

Ahora elimine la columna que había creado y guarde los cambios realizados para continuar al siguiente ejercicio.

Ejercicio 4 - Función SUSTITUIR

Paso 1

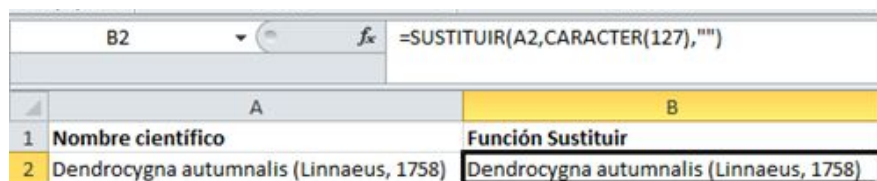
Inicie el ejercicio desde el mismo archivo empleado en los ejercicios anteriores.

Paso 2

En la hoja "Ejemplo (4)" Vaya nuevamente a la columna "**Nombre Científico**" y cree una columna nueva que puede llamar "**Función Sustituir**".

Paso 3

Frente al nombre científico (celda A2), ingrese la función SUSTITUIR tal y como se muestra en la en la Figura 4 =SUSTITUIR(A2,CARACTER(127),"").



	A	B
1	Nombre científico	Función Sustituir
2	Dendrocygna autumnalis (Linnaeus, 1758)	Dendrocygna autumnalis (Linnaeus, 1758)

Figura 4. Ejemplo de uso de la función SUSTITUIR, la barra de fórmulas muestra el comando que se debe ingresar.

Paso 4

Arrastre el contenido de la celda (esquina inferior derecha) hasta empatar los registros de la columna "Nombre Científico".

Paso 5

Reemplace el contenido original de la columna "**Nombre Científico**", copiando los valores de la columna que creó. **Nota:** El contenido debe ser pegado como "valores" (Pegar > Valores). De esta forma, evitamos que la información se altere en un futuro.

Paso 6

Ahora elimine la columna que había creado como "**Función Sustituir**" y guarde los cambios realizados para continuar al siguiente ejercicio.

Ejercicio 5 - Función **CONCATENAR** (1)

Paso 1

Inicie el ejercicio desde el mismo archivo empleado en los ejercicios anteriores.

Paso 2

En la hoja "Ejemplo (5)" vaya a la columna "**Autoría del nombre científico**" e inserte una nueva columna inmediatamente después de esta y nómbrela como "**Nombre científico**". Así podrá ingresar la función a trabajar en la nueva columna.

Paso 3

Frente a la autoría del nombre científico (celda C2), ingrese la función **CONCATENAR** tal y como se muestra en la en la Figura 5 `=CONCATENAR(A2," ",B2," ",C2)`.

	A	B	C	D
1	Género	Epíteto específico	Autoría del nombre científico	Nombre científico
2	Dendrocygna	autumnalis	(Linnaeus, 1758)	Dendrocygna autumnalis (Linnaeus, 1758)
3				

Figura 5. Ejemplo de uso de la función **CONCATENAR**, la barra de fórmulas muestra el comando que se debe ingresar.

Paso 4

Arrastre el contenido de la celda (esquina inferior derecha) hasta empatar los registros de las columnas "Género", "Epíteto específico" y "Autoría del nombre científico".

Paso 5

Copie el contenido de la columna "Nombre científico" y péguelo sobre sí mismo pero como valores. Para esto, debe ir al menú pegar y seleccionar la opción valores. **Nota:** El contenido debe ser pegado como "valores" (Pegar > Valores). De esta forma, evitamos que la información se altere en un futuro.

Paso 6

Guarde los cambios realizados para continuar al siguiente ejercicio.

Ejercicio 6 - Función **CONCATENAR** (2)

Paso 1

Inicie el ejercicio desde el mismo archivo empleado en los ejercicios anteriores.

Paso 2

En la hoja "Ejemplo (6)" vaya a la columna "**Número de catálogo**" e inserte una nueva columna inmediatamente después de esta y nómbrela como "**ID del registro biológico**". Así podrá ingresar la función a trabajar en la nueva columna.

Paso 3

Frente al número de catálogo (celda C2), ingrese la función **CONCATENAR** usando como separador entre palabras el signo (:), tal y como se muestra en la en la Figura 6 **=CONCATENAR(A2,"",B2,"",C2)**.

	A	B	C	D
1	Código de la institución	Código de la colección	Número de catálogo	ID del registro biológico
2	Universidad de Nariño	PSO	12605	Universidad de Nariño:PSO:12605
3				

Figura 6. Ejemplo de uso de la función Concatenar, la barra de fórmulas muestra el comando que se debe ingresar.

Paso 4

Arrastre el contenido de la celda (esquina inferior derecha) hasta empatar los registros de las columnas "Código de la institución", "Código de la colección" y "Número de catálogo".

Paso 5

Inserte una columna inmediatamente después de la columna "**ID del registro biológico**", y copie y pegue el contenido de la columna "**ID del registro biológico**" en ésta. **Nota:** El contenido debe ser pegado como "valores" (Pegar > Valores). De esta forma, evitamos que la información se altere en un futuro.

Paso 6

Ahora reemplace el contenido original de la columna "**ID del registro biológico**" copiando y pegando como valores el contenido de la columna creada en el punto anterior.

Paso 7

Elimine la columna creada en el punto 5 y guarde los cambios realizados para continuar al siguiente ejercicio.

Ejercicio 7 - Función SI

Paso 1

Inicie el ejercicio desde el mismo archivo empleado en los ejercicios anteriores.

Paso 2

En la hoja "Ejemplo (7)" vaya a la columna "**Nombre científico 2**" e inserte una nueva columna inmediatamente después de esta y nómbrela como "**Comparación**". Así podrá ingresar la función a trabajar en la nueva columna.

Paso 3

Frente al nombre científico 2 (celda B2), ingrese la función SI tal y como se muestra en la en la Figura 7 `=SI(A2=B2,"Ok","Revisar")`. **Nota:** En caso que los datos sean iguales, es decir, que el nombre científico 1 y 2 sean el mismo, la función devuelve el valor "Ok" y si encuentra diferencias entre las palabras devuelve el valor "Revisar".

	A	B	C
1	Nombre científico 1	Nombre científico 2	Comparación
2	Dendrocygna autumnalis	Dendrocygna autumnallis	Revisar
3	Cygnus melancoryphus	Cygnus melancorhyphus	Revisar
4	Merganetta armata	Merganetta armata	Ok
5	Piranga flava	Piranga flava	Ok
6	Chlorothraupis olivacea	Chlorothraupis olivacea	Ok
7	Passerina caerulea	Paserina caerulea	Revisar
8	Crypturellus berlepschi	Crypturellus berlepschii	Revisar
9	Nothoprocta taczanowskii	Nothoprocta taczanowski	Revisar

Figura 7. Ejemplo de uso de la función SI, la barra de fórmulas muestra el comando que se debe ingresar.

Paso 4

Arrastre el contenido de la celda (esquina inferior derecha) hasta empatar los registros de las columnas "Nombre científico 1" y "Nombre científico 2".

Paso 5

En la columna "**Comparación**" aparecerán los registros que tienen errores en la nomenclatura con el valor "**Revisar**", se puede utilizar un filtro para escoger únicamente estos registros y analizar con mayor detalle donde se encuentra el error.

Paso 6

Guarde los cambios realizados.

Recomendaciones adicionales

Algunos inconvenientes que se presentan en Excel se deben al cambio de formato en las celdas. Existen dos formas de cambiar el formato (Figura 8):

- Seleccionando la o las celdas de interés y desplegando la pestaña en el panel superior (menú) que da la posibilidad de escoger diferentes categorías como Número, Fecha, Texto etc.
- Haciendo clic derecho sobre la celda que se desee cambiar el formato. Luego se elige "Formato de celdas" desplegándose una ventana emergente que da la posibilidad de escoger diferentes categorías como Número, Fecha, Texto etc.

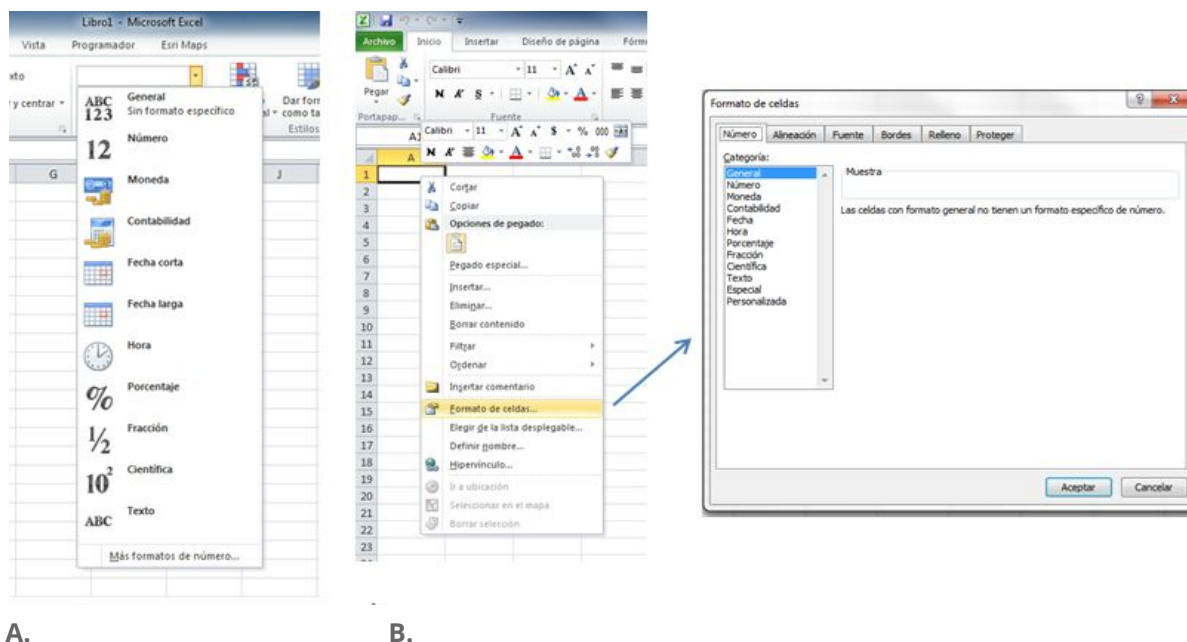


Figura 8. Configuración de formato de celdas en Excel **A.** Botón del formato de celda presente en el menú de Excel. **B** Cambio de formato de celda utilizando el click derecho directamente en la celda objetivo

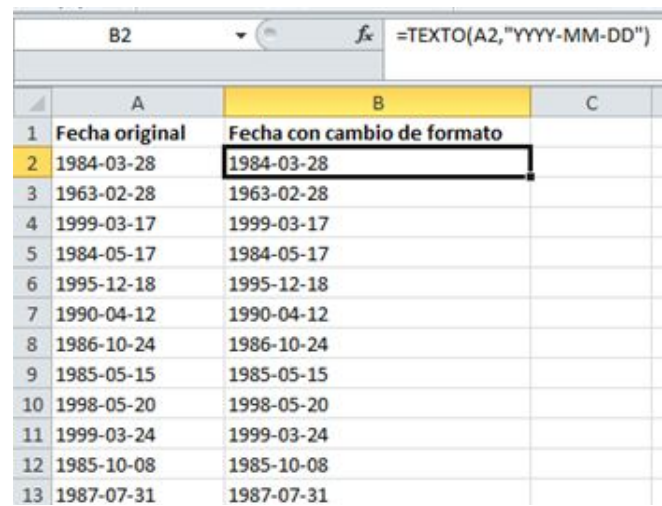
En el caso de los conjuntos de datos estandarizados en DwC, se requiere que las columnas que contengan fechas y coordenadas se encuentren en formato "texto". En ocasiones realizar este cambio de formato puede generar errores en la visibilidad de la fecha, como sucede en el siguiente ejemplo:

	A	B
1	Fecha original	Fecha con cambio de formato
2	1984-03-28	30769
3	1963-02-28	23070
4	1999-03-17	36236
5	1984-05-17	30819
6	1995-12-18	35051
7	1990-04-12	32975
8	1986-10-24	31709

Figura 9. Error común al cambiar una fecha a formato "Texto"

Para evitar este tipo de errores se debe insertar una nueva columna, aplicar la fórmula: `=TEXTO(valor,"YYYY-MM-DD")` y luego pegar estos datos como valores en la columna original que

debe tener el formato de celda "texto". El tema de estandarización de fechas se abordará en detalle en la siguiente sección.



	A	B	C
1	Fecha original	Fecha con cambio de formato	
2	1984-03-28	1984-03-28	
3	1963-02-28	1963-02-28	
4	1999-03-17	1999-03-17	
5	1984-05-17	1984-05-17	
6	1995-12-18	1995-12-18	
7	1990-04-12	1990-04-12	
8	1986-10-24	1986-10-24	
9	1985-05-15	1985-05-15	
10	1998-05-20	1998-05-20	
11	1999-03-24	1999-03-24	
12	1985-10-08	1985-10-08	
13	1987-07-31	1987-07-31	

Figura 10. Ejemplo de uso de la Función Texto, la barra de fórmulas muestra el comando que se debe ingresar.

Estandarización de Fechas

Ingresa y ajustar fechas en un conjunto de datos muchas veces se puede convertir en un trabajo tedioso y frustrante debido a los diferentes formatos en que se puede presentar esta información. Bien sea por la configuración del equipo de cómputo, del programa de sistematización, o de las mismas celdas (en el caso de una hoja de cálculo de Excel). Es importante así contar con herramientas que nos faciliten este proceso de estandarización.

Date Parsing

[Date Parsing](#) es una herramienta en línea, desarrollada por [Canadensys](#) (nodo GBIF Canadá), que permite atomizar y estandarizar fechas de forma masiva, de acuerdo a la ISO 8601 en un formato AAAA-MM-DD; siendo el empleado para el estándar Darwin Core.

Objetivo: Estandarizar y atomizar fechas en el formato ISO 8601 (AAAA-MM-DD) de manera automática y masiva.

Enlace al programa: <http://data.canadensys.net/tools/dates>

Caso Práctico

Usted ha sido contratado por una organización para estandarizar los datos de una colección biológica que posee especímenes con una cobertura temporal de 100 años. Las fechas de colecta se encuentran en diversos formatos y deben estandarizarse de acuerdo a la norma ISO 8601 que es la misma que aplica el Darwin Core en su estándar.

Paso 1

Descargue en el escritorio de su equipo el archivo Excel <http://goo.gl/HLVrwX> . Ahora ábralo y detalle la información que contiene el archivo.

Usted encontrará una lista con 19 fechas de colecta en formatos diferentes y deberá estandarizarlas en el formato AAAA-MM-DD. Usted debe hacerlo bien y en el menor tiempo posible. ¿Lo realizará manualmente? No!

Paso 2

Ingresa a la aplicación [Date Parsing](#) y familiarícese con la herramienta. Identifique los recuadros de la Figura 11

The screenshot shows the 'Date parsing' web application. It features a large text input area on the left, a navigation menu on the top right, and an example input section on the bottom right.

A points to the main text input area where dates are entered. It contains the text 'Jun 13, 2008' and a 'Submit' button at the bottom.

B points to the navigation menu on the top right, which includes links for 'Coordinate conversion', 'Date parsing' (the active tool), 'Tools API', and 'About'.

C points to the 'Example input' section on the bottom right, which displays various date formats: 'Jun 13, 2008', '15 Jan 2011', '2009 IV 02', '2 VII 1986', '1 | 1999/02/24', and '2 | 02/17/1921'.

Figura 11. A. Cuadro de texto para ingresar o pegar las fechas que se desean estandarizar. B. Enlaces a otras herramientas ofrecidas por Canadensys y un breve resumen de esta herramienta. C. Un ejemplo de

las fechas que puede convertir la aplicación. Puede intentar copiar el bloque de fechas y pegarlas en cuadro de texto.

Paso 3

Ahora ingrese las fechas a la aplicación. Para esto, vaya al archivo Excel descargado que contiene las 19 fechas, cópielas (sin incluir el nombre de la columna) y péguelas en el cuadro de texto de la aplicación (Figura 12).

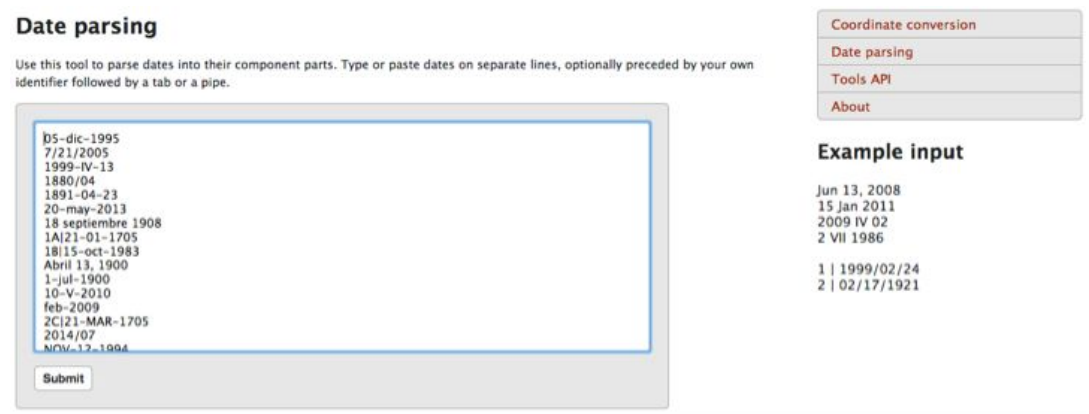


Figura 12. Fechas en diferentes formatos listas para ser convertidas.

Paso 4

Haga clic en el botón **“Submit”** para obtener los resultados. Deberá obtener una tabla similar a la imagen de la Figura 13.

Date parsing results					
id	original	year	month	day	ISO 8601
	05-dic-1995	1995	12	5	1995-12-05
	7/21/2005	2005	7	21	2005-07-21
	1999-IV-13	1999	4	13	1999-04-13
	1880/04	1880	4		1880-04
	1891-04-23	1891	4	23	1891-04-23
	20-may-2013	2013	5	20	2013-05-20
	18 septiembre 1908	1908	9	18	1908-09-18
1A	21-01-1705	1705	1	21	1705-01-21
18	15-oct-1983	1983	10	15	1983-10-15
	Abril 13, 1900	1900	4	13	1900-04-13
	1-jul-1900	1900	7	1	1900-07-01
	10-V-2010	2010	5	10	2010-05-10
	feb-2009	2009	2		2009-02
2C	21-MAR-1705	1705	3	21	1705-03-21
	2014/07	2014	7		2014-07
	NOV-12-1994	1994	11	12	1994-11-12
1C	01-julio-1932	1932	7	1	1932-07-01
	JUNIO 03, 1911	1911	6	3	1911-06-03
	01-dic-2000	2000	12	1	2000-12-01

A B

Figura 13. Visualización de los resultados en *Date parsing*; se presentan en dos secciones: **A**. Columnas con los datos originales. **B**. Columnas con las fechas atomizadas (año, mes, día) y estandarizadas en el formato ISO 8601.

Nota 1: Si desea que los datos cuenten con un identificador, debe incluirlo antes de la fecha y separarlos con un Tab o el símbolo “|”, por ejemplo: 1A|21-03-1987, donde el identificador es 1A

Nota 2: La herramienta no reconoce algunos formatos de fecha como: D-MM-AA, DD-MM-AA, DD-MMM-AA entre otros.

Paso 5

Para finalizar, seleccione la tabla, copíela (incluyendo los encabezados) y péguela como texto en una hoja de Excel vacía. En este punto usted ya tiene 19 fechas estandarizadas en tiempo record!

	A	B	C	D	E	F	G	H
1	id	original	year	month	day	ISO 8601		
2		05-dic-1995	1995	12	5	1995-12-05		
3		7/21/2005	2005	7	21	2005-07-21		
4		1999-IV-13	1999	4	13	1999-04-13		
5		1880/04	1880	4		1880-04		
6		1891-04-23	1891	4	23	1891-04-23		
7		20-may-2013	2013	5	20	2013-05-20		
8		18 septiembre 1908	1908	9	18	1908-09-18		
9	1A	21-01-1705	1705	1	21	1705-01-21		
10	1B	15-oct-1983	1983	10	15	1983-10-15		
11		Abril 13, 1900	1900	4	13	1900-04-13		
12		1-jul-1900	1900	7	1	1900-07-01		
13		10-V-2010	2010	5	10	2010-05-10		
14		feb-2009	2009	2		2009-02		
15	2C	21-MAR-1705	1705	3	21	1705-03-21		
16		2014/07	2014	7		2014-07		
17		NOV-12-1994	1994	11	12	1994-11-12		
18	1C	01-julio-1932	1932	7	1	1932-07-01		
19		JUNIO 03, 1911	1911	6	3	1911-06-03		
20		01-dic-2000	2000	12	1	2000-12-01		

Figura 14. Resultado de la aplicación *Date Parsing* en una hoja de Excel.

Estandarización y validación taxonómica

Homogeneizar y validar la taxonomía de un conjunto de datos representa un importante desafío al estructurar información de biodiversidad. Resulta vital para estructurar datos en el estándar DwC identificar los diferentes niveles taxonómicos que componen el taxón de un registro biológico y definir la categoría a la que pertenece cada uno. Por otro lado, también es común que al ingresar nombres científicos se produzcan errores de tipeo o sistematización al realizar el paso de información física a digital. Este tipo de errores generan inconsistencias que de no corregirse llegan a reflejarse en los portales de datos y páginas web que dan salida a esta información, llevando a resultados no esperados. Actualmente existen herramientas que permiten identificar y corregir estas inconsistencias, o bien completar la información necesaria sobre los niveles taxonómicos superiores de una especie o taxón. A continuación conoceremos algunas de ellas.

ECAT Name Parser

Debido a la complejidad de manejar nombres científicos, los cuales son textos altamente estructurados y cuya sintaxis obedece a un código nomenclatural, GBIF desarrolló la herramienta [ECAT Name Parser](http://tools.gbif.org/nameparser/parser.do). Es una herramienta consolidada en una forma html simple, que hace uso de un analizador (xml), basado en expresiones regulares, que permite separar los nombres científicos en sus distintos componentes y obtener de forma desglosada la información base a partir del original. Permite interpretar la mayoría de nombres científicos y atomizarlos independientemente de su código nomenclatural, siendo capaz de interpretar nombres tan complejos como los nombres híbridos.

Objetivo: aprender a desglosar nombres científicos en sus componentes básicos: género, especie, autor de la especie, rango intraespecífico, epíteto intraespecífico, autor intraespecífico y año. Adicionalmente conocer cómo validar la estructura de los nombres científicos de acuerdo a los códigos nomenclaturales, teniendo en cuenta los alcances y limitaciones de la herramienta.

Enlace al programa: <http://tools.gbif.org/nameparser/parser.do>

Caso Práctico

Usted es un naturalista que desea publicar y compartir a través de GBIF los registros biológicos asociados a sus colectas y observaciones personales, que ha realizado a lo largo de su carrera. Para lograrlo, se ha familiarizado previamente con el estándar que usa GBIF para la publicación de los datos (Darwin Core). En este punto, es consciente que su conjunto de datos requiere un trabajo adicional de depuración para estandarizarlo, donde tendrá que atomizar la información taxonómica, corregir ortografía, entre otros. En su conjunto de datos actual, la única información taxonómica que posee son los nombres científicos en una sola columna.

Usted realizará una serie de ejercicios para la estructuración y estandarización de la información taxonómica, utilizando la herramienta en línea [ECAT Name Parser](http://tools.gbif.org/nameparser/parser.do):

Paso 1

Descargue en el escritorio de su equipo el archivo Excel [Naturalista.xlsx](#). Ahora ábralo y detalle la información que contiene el archivo. Usted encontrará una lista de 30 nombres científicos en una plantilla que contiene algunos elementos del estándar Darwin Core, y deberá documentar los elementos vacíos a partir de la información original. Usted debe hacerlo bien y en el menor tiempo posible. ¿Lo realizará manualmente? No!

Paso 2

Ingresa a la aplicación [ECAT Name Parser](#) y familiarízese con la herramienta. Identifique los recuadros de la siguiente imagen.


The screenshot shows the ECAT Name Parser web application. At the top left is the GBIF logo. To its right is the title 'ECAT Name Parser'. In the top right corner, there are two links: 'parser' and 'api', with the 'parser' link highlighted by a blue box labeled 'A.'. Below the title is a section titled 'Name Parser' with a paragraph of introductory text and a link to 'API documentation'. Below this text is a note: 'You can copy paste a list of names, one per row, or upload a text file with a name per line. Uploaded files have to be encoded as utf8!'. A large text input area is labeled 'B.' and contains three lines of sample text: 'Abies alba Mill.', 'Ge Nicéville 1895', and 'Stagonospora polyspora M.T. Lucas & Sousa da Câmara 1934'. To the left of this area is the text 'Names to parse:' and a note: 'One per line or delimited by the pipe symbol "|"'. Below the text input area is a file upload section labeled 'C.' with the text 'Upload File:' and a button 'Seleccionar archivo' next to the text 'Ningún archivo seleccionado'. To the right of this section is a 'Parse' button labeled 'D.'. Below the file upload section is a 'Load Test Names' button labeled 'E.'.

Figura 15. Vista de la página de inicio de ECAT Name Parser. **A.** Opciones de uso de la herramienta: básico (parser) y avanzado (api). En este ejercicio trabajaremos con la opción básica. Sin embargo, si usted es un usuario avanzado y está familiarizado con el uso de API's, puede consultar la documentación necesaria en la otra opción. **B.** Cuadro de texto para ingresar o pegar los nombres científicos que se desean atomizar. Por defecto encontrará un ejemplo con tres nombres científicos. **C.** Opción para cargar un archivo de texto con los nombres científicos (codificado en UTF-8). **D.** Botón para ejecutar la aplicación. **E.** La aplicación cuenta con una lista de prueba de 274 nombres científicos, con la cual se puede obtener una idea general de los alcances de la herramienta.

Paso 3

Ahora ingrese los nombres científicos en la aplicación. Para esto, vaya al archivo Excel que contiene los 30 nombres científicos, copielos (sin incluir el nombre de la columna) y peguelos en el cuadro de texto de la aplicación. **Nota:** Asegúrese de borrar los 3 nombres científicos que están precargados.

Paso 4

Haga clic en el botón  para obtener los resultados. Deberá obtener una tabla similar a la siguiente imagen:



ECAT Name Parser

[parser](#) [api](#)

Parsed Names

30 name parsed. 22 wellformed, 0 hybrid formulas and 1 doubtful names. See legend for [parsing types](#). **A.**

[Show extended parsing](#) **B.**

Original	Genus	Infrageneric	Specific	Rank	Notho	InfraSpecific	Authorship	Year	(Authorship)	(Year)
Aeschynomene americana glandulosa (Poir. ex Lam.) Rudd	Aeschynomene		americana	INFRASPECIFIC_NAME		glandulosa	Rudd		Poir. ex Lam.	
Anyphaenidae	Anyphaenidae			FAMILY						
Caecilia subnigrans Dunn, 1942	Caecilia		subnigrans	SPECIES			Dunn	1942		
Caiman										

Figura 16. Vista de resultados en ECAT Name Parser **A.** Resumen de los resultados **B.** Opción para expandir la tabla de resultados. **C.** Vista previa de los resultados en una tabla simplificada de 11 columnas.

Paso 4

Ahora visualice los resultados completos haciendo clic en “[Show extended parsing](#)”.

La vista extendida de los resultados (Figura 17) es una tabla de 18 columnas (ver Tabla 2). Dependiendo del tamaño de su pantalla, es posible que necesite utilizar la barra de desplazamiento para visualizarlas todas .



ECAT Name Parser

[parser](#) [api](#)

Parsed Names

30 name parsed. 15 wellformed, 0 hybrid formulas and 9 doubtful names. See legend for [parsing types](#)

[Hide extended parsing](#)

Original	Genus	Infrageneric	Specific	Rank	Notho	InfraSpecific	Authorship	Year	(Authorship)	(Year)	Type	Standardised	Canonical	Sensu	NomStatus	Remarks	GBIF Nub
Aeschynomene americana glandulosa (Poir. ex Lam.) Rudd	Aeschynomene		americana	INFRASPECIFIC_NAME		glandulosa	Rudd		Poir. ex Lam.		WELLFORMED	Aeschynomene americana glandulosa (Poir. ex Lam.) Rudd	Aeschynomene americana glandulosa				search
Anyphaenidae	Anyphaenidae			FAMILY							WELLFORMED	Anyphaenidae	Anyphaenidae				search
Caecilia subnigrans Dunn, 1942	Caecilia		subnigrans	SPECIES							DOUBTFUL	Caecilia subnigrans	Caecilia subnigrans				search
Caiman crocodilus Linnaeus, Caiman 1758)			crocodilus	SPECIES							DOUBTFUL	Caiman crocodilus	Caiman crocodilus				search

Figura 17. Vista extendida de resultados en ECAT Name Parser

A partir de los nombres científicos ingresados, la herramienta atomiza la información nomenclatural en sus partes básicas. La información obtenida es analizada siguiendo diferentes códigos nomenclaturales, principalmente el zoológico (ICZN) y botánico (ICBN).

Tabla 2. Resumen de los elementos que se observan en la vista extendida de ECAT Name Parser.

<i>Original</i>	El nombre científico que fue ingresado
<i>Genus</i>	El género interpretado a partir del nombre científico
<i>Infrageneric</i>	El subgénero o sección interpretado a partir del nombre científico
<i>Specific</i>	El epíteto específico interpretado a partir del nombre científico

<i>Rank</i>	La categoría del taxón o clasificación taxonómica del nombre más específico en el "Nombre científico". Nota: La herramienta utiliza un vocabulario controlado distinto al DwC para este elemento
<i>Notho</i>	La categoría del taxón o clasificación taxonómica de la parte del nombre que corresponde a un híbrido. Nota: Este elemento es útil para los nombre botánicos donde es común tener especímenes híbridos
<i>InfraSpecific</i>	El epíteto infraespecífico interpretado a partir del nombre científico. Por ejemplo: variedad, subespecie.
<i>Authorship</i>	La autoría interpretada a partir del nombre científico
<i>Year</i>	El año de descripción del taxón interpretado a partir del nombre científico
<i>(Authorship)</i>	La autoría infraespecífica interpretada a partir del nombre científico. Nota: Este elemento es útil para los nombre botánicos, los cuales pueden incluir la autoría de la especie y del rango infraespecífico
<i>(Year)</i>	El año de la descripción del taxón interpretado a partir del nombre científico. Nota: Este elemento es útil para los nombre botánicos los cuales pueden incluir el año de descripción del rango infraespecífico
<i>Type</i>	El tipo del nombre de acuerdo a su validez nomenclatural
<i>Standardised</i>	El nombre ingresado después de ser normalizado (espacios y comas ajustados)
<i>Canonical</i>	El nombre canónico interpretado, este solo contiene la información nomenclatural y excluye información taxonómica adicional
<i>Sensu</i>	Indica cuando el nombre científico es utilizado según uno o varios autores, diferentes al autor original. Nota: <i>Sensu</i> , término en latín que significa "en el sentido de". Por ejemplo: sec. Berendsohn 2001, sensu latu.
<i>NomStatus</i>	Un comentario al estatus taxonómico. Por ejemplo; <i>nom.illeg.</i> Generalmente inicia con <i>nom.</i> or <i>nomen</i>
<i>Remarks</i>	Comentarios adicionales a la información nomenclatura que contenga el nombre ingresado
<i>GBIF num</i>	URL a la búsqueda del nombre en " GBIF Backbone taxonomy "

Paso 5

Identifique posibles errores en los nombres científicos. Para esto, seleccione y copie la tabla de resultados (incluyendo los encabezados); y péguela en una hoja de Excel vacía dentro del archivo *Naturalista.xlsx*. Para pegarla utilice la ruta Pegar> Pegado especial >Texto

Paso 6

Agregue un filtro en la primera fila como se muestra en la Figura 18.

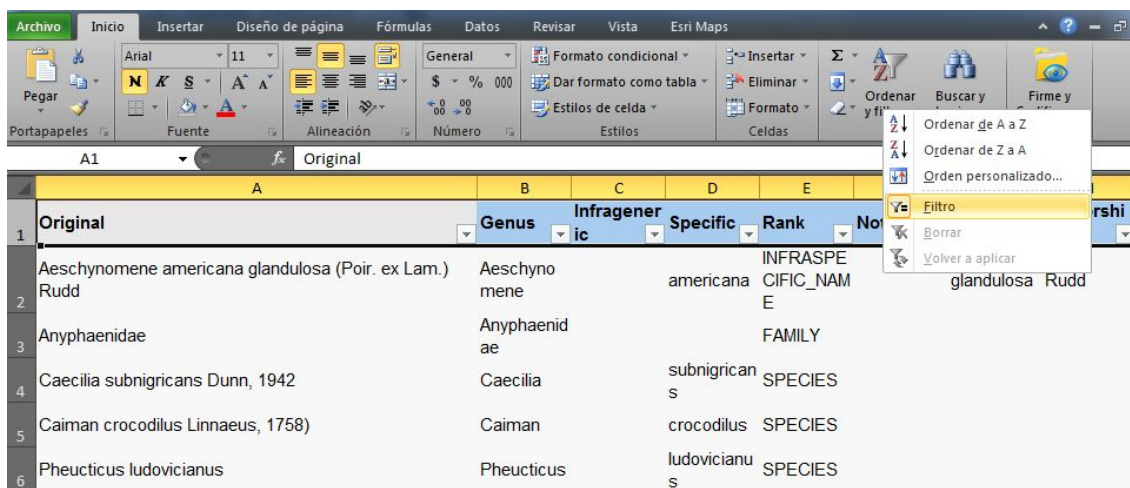


Figura 18. Filtros en Excel de los resultados obtenidos en ECAT Name Parser

Paso 7

Diríjase a la columna del elemento “Type” y con el filtro explore las categorías (Figura 19).

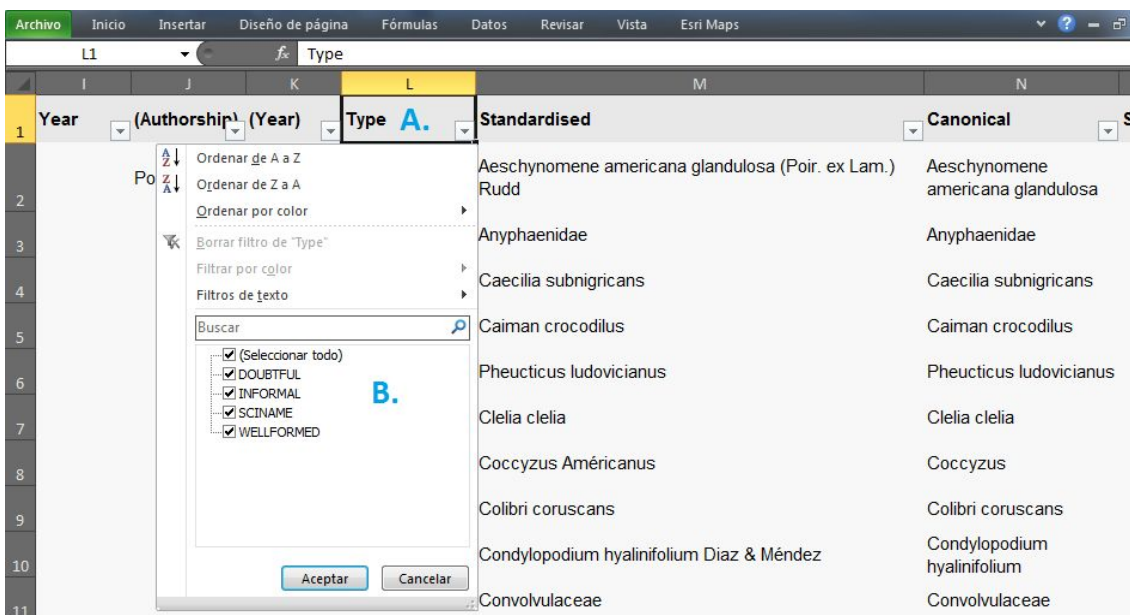


Figura 19. Filtros en Excel por tipo de categoría obtenida en ECAT Name Parser.

¿Recuerda el resumen de resultados de la herramienta en línea?. Aquel resumen (Figura 20) hace referencia al elemento “Type”, el cual le permite identificar algunos errores en los nombres originales.

Parsed Names

30 name parsed. 22 wellformed, 0 hybrid formulas and 1 doubtful names. See legend for [parsing types](#).

Figura 20. Resumen obtenido en la página ECAT Name Parser

Al revisar el resumen general y explorar las categorías, usted encontrará las categorías que se explican en la Tabla 3.

Tabla 3. Resumen de las categorías identificadas por ECAT Name Parser en los taxones ingresados

Categoría	#	Interpretación
WELLFORMED	22	El formato de estos nombres se ajusta a las reglas nomenclaturales
SCINAME	5	Nombres científicos que no están bien formados
INFORMAL	2	Nombres científicos que contienen un calificador de la identificación como <i>cf.</i> y <i>aff.</i>
DOUBTFUL	1	Se duda si los nombres son realmente nombres científicos

Paso 8

Teniendo en cuenta el resumen que arrojó la aplicación y el número de inconsistencias, corrija los errores de formato. La mejor forma, como se puede observar en las Figuras 21 y 22, es filtrando los nombres por categorías (una por una). Filtre primero por “SCINAME”.

Original	InfraSpecific	Type	Standardised	Canonical
Caiman crocodilus Linnaeus, 1758	A.	SCINAME	Caiman crocodilus Linnaeus, 1758	Caiman crocodilus
Clelia clelia Daudin 1803	B.	SCINAME	Clelia clelia Daudin, 1803	C. Clelia clelia
Cupania sylvatica var. sylvatica Seem.	sylvatica	SCINAME	Cupania sylvatica var. sylvatica	Cupania sylvatica sylvatica
Liophis dilepis Cope 1862	B.	SCINAME	Liophis dilepis Cope, 1862	C. Liophis dilepis
Patagioenas fasciata albiLínea	albi	D.	Patagioenas fasciata albi	E. Patagioenas fasciata albi

Figura 21. Errores e inconsistencias detectadas con ECAT Name Parser filtrando por la categoría “SCINAME”. **A.** Falta un paréntesis antes de la autoría. **B.** El nombre no cumple con las reglas nomenclaturales zoológicas, ya que la autoría y el año deben estar separados por una coma. **C.** La herramienta es capaz de detectar, en la mayoría de los casos, este tipo de errores de formato y los corrige automáticamente en el elemento **Standardised**. En este caso no es necesario realizar ninguna corrección. **D.** Observe que el epíteto infraespecífico contiene una mayúscula y una tilde, por esta razón la herramienta hace una interpretación errónea. **E.** Ajuste el epíteto infraespecífico en todos los elementos. Cambie *albiLínea* por *albilinea*.

Ahora filtre por la categoría “INFORMAL” y corrija.

Original	InfraSpecific	Type	Standardised	Canonical
Rupicola peruvianus cf. aequatorialis	A.	aequatorialis	Rupicola peruvianus aequatorialis	Rupicola peruvianus aequatorialis
Stenochrus aff. portoricensis Chamberlin, 1922	B.	INFORMAL	Stenochrus portoricensis Chamberlin, 1922	C. Stenochrus portoricensis

Figura 22. Errores e inconsistencias detectadas con ECAT Name Parser filtrando por la categoría “INFORMAL”. **A.** Quien identificó el espécimen tenía dudas e indicó que es necesario confirmar (*cf.*) la identificación. **B.** Quien identificó el espécimen tenía dudas e indicó que la especie es similar o afín (*aff.*) a *Stenochrus portoricensis*. **C.** La herramienta detecta los términos *aff.* y *cf.* y los elimina para cumplir con el código nomenclatural.

Nota: Para no perder los calificadores de la identificación, vaya al archivo original y documéntelos en la columna del elemento “**Calificador de la identificación**”. Para los dos casos anteriores, los calificadores de la identificación serían: “*cf. aequatorialis*” y “*aff. portoricensis*”.

Ahora filtre por la categoría "DOUBTFUL":

	A	G	L	M	N
1	Original	InfraSpecific	Type	Standardised	Canonical
11	Cordia bullata subsp. bullata (L.) Roem. Y Schult. A.	bullata	DOUBTFUL	Cordia bullata subsp. bullata B.	Cordia bullata bullata
32					
33					
34					

Figura 23. Errores e inconsistencias detectadas con ECAT Name Parser filtrando por la categoría "DOUBTFUL" **A.** La "Y" en la autoría puede ser un error pero por el momento no podemos tener certeza de ello. **B.** La herramienta estandariza los nombres omitiendo las autorías ya que no hay certeza sobre estas. **Nota:** si los nombres ingresados contienen caracteres no imprimibles es posible que la herramienta categorice el nombre como "DOUBTFUL" a pesar de que este sea correcto.

En este punto ya ha identificado y corregido los nombres científicos de acuerdo al resultado que arrojó la herramienta. Ahora, en el archivo original, debe completar las columnas (elementos) vacías con el resto de la información.

Paso 8

Copie los datos de la columna **Standardised** en la columna del elemento DwC **Nombre científico**. **Nota:** El nombre científico original se reemplaza por el nombre científico de la columna **Standardised**, ya que la herramienta ajusta las autorías, el uso de comas y elimina espacios sobrantes entre otros. Sin embargo, recuerde siempre guarda la información original en una columna diferente. Por ejemplo: "Nombre científico original".

Paso 9

Copie los datos de la columna **Genus** en la columna del elemento DwC **Género**. **Nota:** es posible que la herramienta ponga los nombres de las familias en la columna **Genus** utilice un filtro en la columna **Rank** para verificar y ajustar la información.

Paso 10

Copie los datos de la columna **Specific** en la columna del elemento DwC **Epíteto específico**.

Paso 11

Copie los datos de la columna **InfraSpecific** en la columna del elemento DwC **Epíteto infraespecífico**. **Nota:** asegúrese de haber ajustado la información de este elemento de acuerdo a las indicaciones del paso 8.

Paso 12

Copie los datos de la columna **Rank** en la columna del elemento DwC **Categoría del taxón**. **Nota:** Las categorías de **Rank** deben ser ajustadas al vocabulario controlado del elemento categoría del taxón. Este punto se tocará nuevamente en la siguiente guía.

Paso 13

Es posible reconstruir la autoría del nombre científico a partir de la información en las columnas **Autorship**, **Year**, **(Autorship)** y **(Year)**, en el elemento DwC **Autoría del nombre científico**. Para ello, puede ayudarse de la función de Excel "concatenar". Haga el ejercicio solo para dos autorías, en los siguientes ejercicios prácticos, se verán maneras más eficientes de obtener la autoría.

Después de completar la información el archivo de Excel, debe contar con una hoja de cálculo similar a la Figura 24.

DT	DU	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF
Calificador de la identificación	Nombre científico	Reino	Filo	Clase	Orden	Familia	Género	Epíteto específico	Epíteto infraespecífico	Categoría del taxón	Autoría del nombre científico
	Aeschynomene americana glandulosa (Poir. ex Lam.) Rudd					Aeschynomene	americana	glandulosa		INFRASPECIFIC_NAME	(Poir. ex Lam.) Rudd
	Anyphaenidae					Anyphaenidae				FAMILY	
	Caecilia subnigricans Dunn, 1942					Caecilia	subnigricans			SPECIES	
	Caiman crocodilus Linnaeus, 1758					Caiman	crocodilus			SPECIES	
	Clelia clelia Daudin 1803					Clelia	clelia			SPECIES	
	Coccyzus Americanus					Coccyzus	americanus				
	Colibri coruscans (Gould, 1846)					Colibri	coruscans			SPECIES	
	Condylopodium hyalinifolium Diaz & Méndez					Condylopodium	hyalinifolium			SPECIES	Diaz & Méndez
	Convolvulaceae					Convolvulaceae				FAMILY	
	Cordia bullata subsp. bullata (L.) Roem. Y Schult.					Cordia	bullata	bullata		SUBSPECIES	
	Cupania sylvatica var. sylvatica Seem.					Cupania	sylvatica	sylvatica		VARIETY	
	Cydista aequinoctialis (L.) Miers					Cydista	aequinoctialis			SPECIES	(L.) Miers
	Huperzia reflexa (Lam.) Trevis.					Huperzia	reflexa			SPECIES	(Lam.) Trevis.
	Ischnothele caudata Ausserer, 1875					Ischnothele	caudata			SPECIES	Ausserer, 1875
	Liophis dilepis Cope 1862					Liophis	dilepis			SPECIES	Cope, 1862
	Lygromma chamberlini Gertsch, 1941					Lygromma	chamberlini			SPECIES	Gertsch, 1941
	Lygromma quindio Platnick & Shadab, 1976					Lygromma	quindio			SPECIES	Platnick & Shadab, 1976
	Pardosa fastosa (Keyserling, 1877)					Pardosa	fastosa			SPECIES	(Keyserling, 1877)
	Patagioenas fasciata albilinea					Patagioenas	fasciata	albilinea		INFRASPECIFIC_NAME	
	Pheucticus ludovicianus					Pheucticus	ludovicianus			SPECIES	
	Playa cayana nigricrissa					Playa	cayana	nigricrissa		INFRASPECIFIC_NAME	
	Pseudoboa neuwiedii (Duméril, Bibron & Duméril, 1854)					Pseudoboa	neuwiedii			SPECIES	
	Psilorhinus mexicanus					Psilorhinus	mexicanus			SPECIES	
cf. aequatorialis	Rupicola peruvianus cf. aequatorialis					Rupicola	peruvianus	aequatorialis		INFRASPECIFIC_NAME	
	Scinax x-signatus (Spix, 1824)					Scinax	x-signatus			SPECIES	(Spix, 1824)
	Steatornis caripensis					Steatornis	caripensis			SPECIES	
aff. portoricensis	Stenochrus aff. portoricensis Chamberlin, 1922					Stenochrus	portoricensis			SPECIES	Chamberlin, 1922
	Thaumasia					Thaumasia					
	Thraupis cyanocephala annectens					Thraupis	cyanocephala	annectens		INFRASPECIFIC_NAME	
	Typhlonectes natans (Fischer in Peters, 1880)					Typhlonectes	natans			SPECIES	

Figura 24. Resultado final en Excel luego de la estandarización y estructuración de la información taxonómica con ECAT Name Parser.

Paso 14

En este punto ya tiene estructurada y estandarizada la información referente al nombre científico de sus registros biológicos. Guarde el archivo *Naturalista.xlsx* con los ajustes realizados en esta práctica, ya que los necesitará en próximos ejercicios.

Para tener en cuenta: la versión en línea no permite exportar los resultados, lo cual puede ser una desventaja. Sin embargo, si usted es un usuario avanzado puede utilizar la API para hacerlo. ECAT Name Parser le permitirá iniciar la estructuración de los siguientes elementos del estándar Dwc: Calificador de la identificación, Nombre científico, Género, Epíteto específico, Epíteto infraespecífico, Categoría del taxón y Autoría del nombre científico; para los cuales validará el formato nomenclatural. Recuerde que para corroborar la validez taxonómica es necesario utilizar otras herramientas como las se verán las siguientes prácticas.

Enlaces de interés: Puede obtener más información sobre el desarrollo y funcionamiento de esta herramienta en el blog [Working with Scientific Names, GBIF developer Blog](#)

Resolución Taxonómica y de LSIDs

El SiB Colombia desarrolló una aplicación web asistida para la obtención de las categorías taxonómicas con sus respectivos nombres, autorías y LSIDs (*Life Science Identifier*) —es un identificador alfanumérico global, único y persistente que es usado por la comunidad científica para identificar los objetos taxonómicos—. Esta herramienta, basada en la lista anual de [Catalogue of Life](#), del año 2012, facilita toda la recuperación y resolución taxonómica a partir de nombres científicos, a todo nivel (reino, filo, familia, especie, etc.).

La herramienta automatiza las siguientes tareas:

- ❑ Obtención de las categorías taxonómicas con su respectivo nombre y la autoría.
- ❑ Obtención de los LSIDs. Estos son usados bajo el esquema de publicación del SiB Colombia como un identificador único, global y persistente del nombre a publicar.

El límite de nombres científicos que pueden someterse por búsqueda es de 5000. Si para un nombre científico determinado no se encuentra una coincidencia en *Catalogue of Life* (2012), el LSID y la taxonomía superior no serán generados.

El código de la aplicación se encuentra disponible en repositorio de GitHub del [SiB Colombia](#).

Objetivo: Aprender a obtener la taxonomía superior a partir de los nombres científicos, así como sus LSIDs de manera masiva, teniendo en cuenta los alcances y limitaciones de la herramienta.

Enlace al programa: <http://tools.sibcolombia.net/taxon/index.php/taxon/busqueda>

Caso Práctico

Usted es un naturalista que desea publicar y compartir a través de GBIF los registros biológicos asociados a sus colectas y observaciones personales, que ha realizado a lo largo de su carrera. Para lograrlo, se ha familiarizado previamente con el estándar que usa GBIF para la publicación de los datos (Darwin Core). En este punto, es consciente que su conjunto de datos requiere un trabajo adicional de depuración para estandarizarlo. Después de usar la herramienta [ECAT Name Parser](#), usted ya cuenta con nombres científicos estructurados y ha documentado algunos elementos DwC. Sin embargo, aún necesita completar la información de los niveles taxonómicos superiores a género, para lo cual decide utilizar la herramienta [Resolución Taxonómica y de LSIDs](#).

Paso 1

Abra el archivo *Naturalista.xlsx* con los resultados de la **Práctica 2**.

Paso 2

Ingresa a la aplicación web [Resolución Taxonómica y de LSIDs](#) y familiarícese con la herramienta. Identifique los recuadros de la Figura 25.

A.

B. Ingresar Lista **C.** Subir Lista **D.** Instrucciones **E.** Acerca de

Ingrese el archivo con los nombres.

Ingrese hasta 5000 nombres

F. **BUSCAR** **G.** **LIMPIAR**

H. **RESULTADOS** **I.** **+ Exportar Datos**

Desplegando 1-1 de 1 resultado.

Nombre Id Ingresado	Reino	Filo	Clase	Orden	Familia	Género	Epíteto Específico	Epíteto Infraespecífico	Categoría	Autor	Nombre Científico	LSID
-	-	-	-	-	-	-	-	-	-	-	-	-

Figura 25. Vista de la página de inicio del Servicio de Resolución Taxonómica y de LSIDs del SiB Colombia. **A.** La herramienta está disponible en español e inglés, ajústela a su preferencia de idioma. En la parte superior hay cuatro pestañas: **B.** Ingresar lista: sección para ingresar los nombres haciendo copy & paste; **C.** Subir lista: sección para subir archivos Excel o TXT con la lista de nombres científicos; **D.** Instrucciones: sección con la información básica para utilizar la herramienta; **E.** Acerca de: sección con la documentación de la herramienta. **F.** Botón para ejecutar la aplicación. **G.** Botón para borrar los nombres cargados en el área de texto y/o los archivos subidos. **H.** Tabla que muestra la vista previa de la información taxonómica encontrada. **I.** Botón para exportar los resultados en un archivo de Excel, el cual contiene la versión completa de los resultados.

Paso 3

Ingrese los nombres en la herramienta. Para esto, vaya a la hoja del archivo Naturalista.xlsx que contiene los resultados de la herramienta ECAT name parser, copie los nombres científicos de la columna **Canonical** (sin incluir el nombre de la columna) y peguelos en el cuadro de texto de la aplicación.

Nota: Para optimizar la búsqueda en la herramienta, los nombres deben estar estandarizados y no deben incluir espacios adicionales, calificadores de identificación, ni autorías. Abreviaturas taxonómicas como **var.** y **subsp.** para nombres científicos por debajo de especie, son necesarias para obtener una resolución taxonómica exitosa. En el ejercicio con la herramienta [ECAT Name Parser](#), se obtuvo una columna con los nombres canónicos (**Canonical**), la cual cumple con los requerimientos anteriores.

Paso 4

Con los nombre ya ingresados en la herramienta, haga clic en el botón **BUSCAR** para obtener los resultados. La aplicación deberá mostrarle una tabla similar a la de la Figura 26.

SiB | PORTAL DE DATOS

PORTAL GENERAL

Ingrese hasta 5000 nombres

Aeschynomene americana glandulosa
Anyphaenidae
Caecilia subnigricans
Caiman crocodilus
Clelia clelia
Coccyzus americanus
Colibri coruscans

BUSCAR LIMPIAR

RESULTADOS

Exportar Datos

Desplegando 1-10 de 30 resultados

Id	Nombre Ingresado	Reino	Filo	Clase	Orden	Familia	Género	Epíteto Específico	Epíteto Infraespecífico	Categoría	Autor	Nombre Científico	LSID
1	Aeschynomene americana glandulosa	-	-	-	-	-	-	-	-	-	-	-	-
2	Anyphaenidae	Animalia	Arthropoda	Arachnida	Araneae	Anyphaenidae	-	-	-	family	-	Anyphaenidae	urn:lsid:catalogueoflife.org:taxon:29c1-102b-9a4a-00304854f820:col20120124
3	Caecilia subnigricans	Animalia	Chordata	Amphibia	Gymnophiona	Caeciliidae	Caecilia	subnigricans	-	species	Dunn, 1942	Caecilia subnigricans Dunn, 1942	urn:lsid:catalogueoflife.org:taxon:29c1-102b-9a4a-00304854f820:col20120124

Figura 26. Vista de resultados en el Servicio de Resolución Taxonómica y de LSIDs del SiB Colombia.

Nota: En la tabla de resultados sólo aparecerá una fila por cada nombre único. Es decir, si usted ingresó una lista de nombres científicos en la cual aparece más de una vez una misma especie, en los resultados dicha especie aparecerá una sola vez.

Paso 5

Ahora haga clic en el botón **Exportar Datos** para descargar los resultados completos en un archivo Excel. Guárdelo en su equipo.

En el archivo de Excel encontrará: (i) El nombre original que ingresó (columna **Taxón**); (ii) ocho categorías taxonómicas, cada una con su respectivo LSID; (iii) la categoría del taxón (columna **Rango**); (iv) la autoría del nombre científico; (v) el nombre científico extraído de la base de datos y (vi) su LSID. Su archivo Excel debe ser similar al de la Figura 27.

Archivo Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista Esri Maps								
S35 fx								
A	B	C	D	R	S	T		
#	Taxón	Reino	Reino ID	Rango	Autor	Nombre Científico	LSID	
1	Aeschynomene americana glandulosa	-	-	-	-	-	-	
2	Anyphaenidae	Animalia	urn:lsid:cc:family	-	-	Anyphaenidae	urn:lsid:cata	
3	Caecilia subnigricans	Animalia	urn:lsid:cc:species	Dunn, 1942	Caecilia subnigricans Dunn, 1942	Caecilia subnigricans Dunn, 1942	urn:lsid:cata	
4	Caiman crocodilus	Animalia	urn:lsid:cc:species	Linnaeus 1758	Caiman crocodilus Linnaeus 1758	Caiman crocodilus Linnaeus 1758	urn:lsid:cata	
5	Clelia clelia	Animalia	urn:lsid:cc:species	DAUDIN 1803	Clelia clelia DAUDIN 1803	Clelia clelia DAUDIN 1803	urn:lsid:cata	
6	Coccyzus americanus	Animalia	urn:lsid:cc:species	(Linnaeus, 1758)	Coccyzus americanus (Linnaeus, 1758)	Coccyzus americanus (Linnaeus, 1758)	urn:lsid:cata	
7	Colibri coruscans	Animalia	urn:lsid:cc:species	(Gould, 1846)	Colibri coruscans (Gould, 1846)	Colibri coruscans (Gould, 1846)	urn:lsid:cata	
8	Condylopodium hyalinifolium	Plantae	urn:lsid:cc:species	Diaz & Méndez	Condylopodium hyalinifolium Diaz & Méndez	Condylopodium hyalinifolium Diaz & Méndez	urn:lsid:cata	
9	Convolvulaceae	Plantae	urn:lsid:cc:family	-	-	Convolvulaceae	urn:lsid:cata	
10	Cordia bullata bullata	-	-	-	-	-	-	
11	Cupania sylvatica sylvatica	-	-	-	-	-	-	
12	Cydista aequinoctialis	Plantae	urn:lsid:cc:species	(L.) Miers	Cydista aequinoctialis (L.) Miers	Cydista aequinoctialis (L.) Miers	urn:lsid:cata	
13	Huperzia reflexa	Plantae	urn:lsid:cc:species	(Lam.) Trevisan	Huperzia reflexa (Lam.) Trevisan	Huperzia reflexa (Lam.) Trevisan	urn:lsid:cata	

Figura 27. Archivo Excel con los datos obtenidos desde el Servicio de Resolución Taxonómica y de LSIDs del SiB Colombia.

Paso 6

Ahora complete la información taxonómica del archivo *Naturalista.xlsx* a partir de los resultados obtenidos. Los elementos a completar haciendo copy & paste son: **Género**, **Reino**, **Filo**, **Clase**, **Orden** y **Familia**.

También puede documentar el elemento **Categoría del taxón**, en este caso particular no es necesario, debido a que los resultados no ofrecen información adicional a la que ya está documentada.

Paso 7

Al comparar la información de la autoría del nombre científico con la obtenida en la práctica de ECAT Name Parser, usted puede identificar que ahora tiene las autorías para más nombres. Re-ajuste el elemento **Autoría del nombre científico** con la nueva información.

Nota: Puede filtrar la columna de autorías y hacer una revisión rápida en busca de inconsistencias. Si detecta autorías en mayúscula o sin la coma entre el autor y el año, debe ajustarlas.

En este punto usted ya ha estructurado y validado la mayoría de la información taxonómica bajo el estándar Darwin Core. Su conjunto de datos deberá verse como la Figura 28.

Archivo													
Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista Esri Maps													
EC45													
DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC		
Calificador de la identificación	Nombre científico	Reino	Filo	Clase	Orden	Familia	Género	Epíteto específico	Epíteto infraespecífico	Categoría del taxón	Autoría del nombre científico		
2	Aeschynomene americana	glandulosa (Poir. ex Lam.) Rudd						Aeschynomene americana	glandulosa	INFRASPECIFIC_NAME	(Poir. ex Lam.) Rudd		
3	Anypheidae	Animalia	Arthropoda	Arachnida	Araneae	Anypheidae	Anypheidae			FAMILY			
4	Caecilia subnigrans	Dunn	Animalia	Chordata	Amphibia	Gymnophiona	Caeciliidae	Caecilia	subnigrans	SPECIES	Dunn, 1942		
5	Caiman crocodilus	Linnaeus	Animalia	Chordata	Reptilia	Crocodylia	Crocodylidae	Caiman	crocodilus	SPECIES	Linnaeus 1758		
6	Clelia clelia	Daudin 1803	Animalia	Chordata	Reptilia	Squamata	Colubridae	Clelia	clelia	SPECIES	Daudin, 1803		
7	Coccyzus americanus		Animalia	Chordata	Aves	Cuculiformes	Cuculidae	Coccyzus	americanus		(Linnaeus, 1758)		
8	Colibri coruscans	(Gould, 1846)	Animalia	Chordata	Aves	Apodiformes	Trochilidae	Colibri	coruscans	SPECIES	(Gould, 1846)		
9	Condylopodium hyalinifolium	Plantae	Magnoliops	Magnoliops	Asteraceae	Condylopodium	hyalinifolium			SPECIES	Diaz & Méndez		
10	Convolvulaceae	Plantae	Magnoliops	Magnoliops	Solanales	Convolvulaceae				FAMILY			
11	Cordia bullata subsp. bullata (L.) R.								bullata	SUBSPECIES			
12	Cupania sylvatica var. sylvatica See								sylvatica	VARIETY			
13	Cydista aequinoctialis (L.) Miers	Plantae	Magnoliops	Magnoliops	Bignoniaceae	Cydista	aequinoctialis			SPECIES	(L.) Miers		
14	Huperzia reflexa (Lam.) Trevisan	Plantae	Lycopodiophyta	Lycopodiophyta	Lycopodiaceae	Huperzia	reflexa			SPECIES	(Lam.) Trevisan		
15	Ischnothele caudata Ausserer, 1875	Animalia	Arthropoda	Arachnida	Araneae	Dipluridae	Ischnothele	caudata		SPECIES	Ausserer, 1875		
16	Liophis dilepis Cope 1862	Animalia	Chordata	Reptilia	Squamata	Colubridae	Liophis	dilepis		SPECIES	Cope, 1862		
17	Lygromma chamberlini Gertsch, 1941	Animalia	Arthropoda	Arachnida	Araneae	Prodromidae	Lygromma	chamberlini		SPECIES	Gertsch, 1941		
18	Lygromma quindio Platnick & Shadab, 1976	Animalia	Arthropoda	Arachnida	Araneae	Prodromidae	Lygromma	quindio		SPECIES	Platnick & Shadab, 1976		
19	Pardosa fastosa (Keyserling, 1877)	Animalia	Arthropoda	Arachnida	Araneae	Lycosidae	Pardosa	fastosa		SPECIES	(Keyserling, 1877)		
20	Patagioena fasciata albilinea (Bonaparte, 1854)	Animalia	Chordata	Aves	Columbiformes	Columbidae	Patagioena	fasciata	albilinea	INFRASPECIFIC_NAME	(Bonaparte, 1854)		
21	Pheucticus ludovicianus (Linnaeus, 1766)	Animalia	Chordata	Aves	Passeriformes	Cardinalidae	Pheucticus	ludovicianus		SPECIES	(Linnaeus, 1766)		
22	Playa cayana nigricrisa (Cabanis, 1862)	Animalia	Chordata	Aves	Cuculiformes	Cuculidae	Playa	cayana	nigricrisa	INFRASPECIFIC_NAME	(Cabanis, 1862)		
23	Pseudoboa newwedii (Duméril, Bibron & Duméril, 18	Animalia	Chordata	Reptilia	Squamata	Colubridae	Pseudoboa	newwedii		SPECIES	Duméril, Bibron & Duméril, 18		
24	Psilorhinus mexicanus (Rueppell, 1837)	Animalia	Arthropoda	Insecta	Coleoptera	Curculionidae	Psilorhinus	mexicanus		SPECIES	Rueppell, 1837		
25	Rupicola peruviana cf. ae	Animalia	Chordata	Aves	Passeriformes	Cotingidae	Rupicola	peruviana	ae	INFRASPECIFIC_NAME	Taczanowski, 1889		
26	Scinax x-signatus (Spix, 1824)	Animalia	Chordata	Amphibia	Anura	Hylidae	Scinax	x-signatus		SPECIES	(Spix, 1824)		
27	Stenomantis caripensis (Humboldt, 1817)	Animalia	Chordata	Amphibia	Strigiformes	Stenomantis	caripensis			SPECIES	Humboldt, 1817		
28	Stenomantis caripensis (Chamberlin, 1922)	Animalia	Arthropoda	Arachnida	Schizomida	Hubbardiidae	Stenomantis	caripensis		SPECIES	Chamberlin, 1922		
29	Thaumasia (Zimmer, 1944)	Animalia	Arthropoda	Arachnida	Araneae	Pisauridae	Thaumasia				(Zimmer, 1944)		
30	Thraupis cyanocephala (Fischer in Peters, 1880)	Animalia	Chordata	Aves	Passeriformes	Thraupidae	Thraupis	cyanocephala	annectens	INFRASPECIFIC_NAME	(Fischer in Peters, 1880)		
31	Typhlonectes natans (Fischer in Peters, 1880)	Animalia	Chordata	Amphibia	Gymnophiona	Caeciliidae	Typhlonectes	natans		SPECIES	(Fischer in Peters, 1880)		
32													

Figura 28. Archivo estructurado en DwC con los datos obtenidos a partir Servicio de Resolución Taxonómica y de LSIDs del SiB Colombia.

Paso 8

Identifique los vacíos de información en la tabla. Hay casos donde la herramienta no logra obtener la información esperada. Por ejemplo:

- ☐ Falta la taxonomía superior de tres géneros.
- ☐ La información de la columna **Categoría del taxón** está incompleta y no se ajusta completamente al vocabulario controlado sugerido por el estándar Darwin Core. Usted puede utilizar la opción de buscar y reemplazar para actualizar para actualizar elemento con el [vocabulario controlado](#) del estándar DwC.
- ☐ En la columna **Autoría del nombre científico**, falta documentar el autor para varios taxa.

Tenga en cuenta los puntos anteriores, pues estos vacíos serán resueltos en el siguiente ejercicio.

Paso 9

Los LSIDs son indispensables si usted desea publicar a través de GBIF una lista de especies a partir de la información que está estructurando. Por tal razón, conserve los LSIDs que obtuvo con la herramienta. Para hacerlo, copie los identificadores de la última columna (**LSID**) del archivo de resultados, que corresponde a los LSIDs de los taxa ingresados; y péguela en el elemento **ID del taxón**.

Herramientas alternativas para obtener LSID's

Catalogue of Life es probablemente la base de datos de nombres científicos más grande que existe, debido a que integra información de diferentes fuentes y grupos biológicos. Sin embargo, tiene algunos vacíos de información para ciertos grupos biológicos y es necesario complementarlos con otras bases de datos como: [Universal Biological Indexer and Organizer uBio](#) y [Index to Organism Names: ION](#).

Haciendo uso de otras bases de datos como [uBio](#) y [ION](#), complete los identificadores faltantes de los nombres científicos que identificó en el paso 8 del **Ejercicio del SIB**.

Nota: De estas bases de datos usted también puede extraer la información taxonómica superior.

Paso 1

Busque en una de las bases de datos sugeridas (Figura 29) los nombres de las especies que faltan por LSID.



Figura 29. Búsqueda del nombre “*Cupania sylvatica* Seem. var. *sylvatica*” en la base de datos de uBio

Paso 2

Obtenga el LSID a través de la base de datos seleccionada (Figura 30) y complete la información en el conjunto de datos en los nombres científicos que carecían de identificador.

Find name...

- [uBio Project](#)
 - [Introduction](#)
 - [News](#)
- [Servers](#)
 - [NameBank](#)
 - [ClassificationBank](#)
 - [List Classifications](#)
 - [Statistics](#)

Cupania sylvatica Seem. var. sylvatica
NamebankID: 9084180
LSID: <urn:lsid:ubio.org:namebank:9084180> ([view metadata](#))

Synonyms:

Cupania sylvatica sylvatica

Figura 30. Resultado de la búsqueda en la base de datos de uBio.

Para tener en cuenta: la herramienta [Resolución Taxonómica y de LSIDs](#), tiene la ventaja de permitir exportar sus resultados a un archivo Excel facilitando el trabajo de estandarización y estructuración de los siguientes elementos del estándar Dwc: [Nombre científico](#), [Reino](#), [Filo](#), [Clase](#), [Orden](#), [Familia](#), [Género](#), [Epíteto específico](#), [Epíteto infraespecífico](#), [Categoría del taxón](#) y [Autoría del nombre científico](#).

Por el momento la mayor desventaja de esta herramienta es que está basada en la lista anual del año 2012 de [Catalogue of Life](#), por lo cual hay varios nombres científicos que no podrá resolver a pesar de que estos ya se encuentran en la versión actual de [Catalogue of Life](#); en los próximos meses la herramienta contará con la última versión.

Enlaces de interés:

Catalogue of Life

<http://www.catalogueoflife.org/content/frequently-asked-questions#1>

LSID Tester, a tool for testing Life Science Identifier resolution services

<http://www.scfbm.org/content/3/1/2>

<http://darwin.zoology.gla.ac.uk/~rpage/lsid/tester/>

Catalogue of Life and LSIDs

<http://iphylo.blogspot.com/search/label/LSID>

Taxonomic Name Resolution Service (TNRS)

Cuando se manejan conjuntos de datos que incluyen información de múltiples taxa, herramientas más generales como [ECAT Name Parser](#) y [Resolución Taxonómica y de LSIDs](#) son útiles para estandarizar la información de manera masiva. Sin embargo, estas herramientas no siempre son suficientes para estandarizar todos nuestros datos. Para ello, se pueden aprovechar bases de datos especializadas como *Taxonomic Name Resolution Service* (TNRS). Esta es una aplicación desarrollada por iPlant, que permite estandarizar los nombres científicos botánicos a partir de fuentes como [Tropicos](#), [USDA Names](#), GCC y [NCBI](#). Para ello, corrige errores ortográficos y palabras alternativas para una lista estándar de nombres científicos, y convierte los sinónimos detectados a los nombres actualmente aceptados. TNRS puede procesar muchos nombres al mismo tiempo, ahorrando horas de corrección manual y disminuyendo la probabilidad de error. Para los nombres que no se pueden resolver de forma automática, TNRS presenta una lista de posibilidades y proporciona herramientas para investigar y seleccionar el nombre adecuado.

Objetivo: validar nombres científicos de plantas de manera automática y actualizar los registros biológicos con APGIII, teniendo en cuenta los alcances y limitaciones de la herramienta.

Enlace al programa: <http://tnrs.iplantcollaborative.org/TNRSapp.html>

Caso Práctico

Usted es un naturalista que desea publicar y compartir a través de GBIF los registros biológicos asociados a sus colectas y observaciones personales, que ha realizado a lo largo de su carrera. Para lograrlo, se ha familiarizado previamente con el estándar que usa GBIF para la publicación de los datos (Darwin Core). En este punto, es consciente que su conjunto de datos requiere un trabajo adicional de depuración para estandarizarlo. En los dos ejercicios anteriores usted estructuró los nombres científicos y completó la información taxonómica de los niveles superiores, ahora desea validar si los nombres de su conjunto son aceptados o si deben ser actualizados.

EJERCICIO 1

Paso 1

Abra el archivo Excel *Naturalista.xlsx* que trabajó en la sección Resolución Taxonómica y de LSIDs.

Paso 2

Ingresa a la aplicación [Taxonomic Name Resolution Service \(TNRS\)](#) y familiarícese con la herramienta. Identifique los recuadros de la Figura 31.

A. Enter List

Enter scientific names to check

Enter up to 5000 names.

Click here for support

Clear Submit List

B. Upload and Submit List

C. Retrieve Results

Name processing settings

Processing Mode: Edit
Selected mode: Perform Name Resolution

Match Accuracy: Edit
Allow partial matches, Selected minimum threshold: 0.05

Sources: Edit
[TROPICOS, GCC, USDA]

Family Classification: Edit
Selected classification source: TROPICOS

Figura 31. Vista de la página de inicio de TNRS. La herramienta cuenta con tres pestañas o secciones: **A.** Enter List: Sección para ingresar la lista de nombres haciendo copy & paste; **B.** Upload and Submit List: Sección para cargar archivos TXT con la lista de nombres. Para utilizar esta opción debe proveer un correo electrónico donde será notificado cuando los resultados estén listos. **C.** Retrieve Results: si usted sometió los nombres en un archivo de texto, aquí puede ingresar la clave que llegó a su correo para visualizar los resultados. **D.** Aquí puede seleccionar el modo en desea que la herramienta ejecute la tarea: (i) *Perform Name Resolution* separa, estandariza y valida los nombre científicos con una o varias fuentes, y (ii) *Parse Names Only* realiza una tarea similar a [ECAT Name Parse](#). **E.** Aquí puede ajustar la sensibilidad de la herramienta durante el proceso de búsqueda. Si desea obtener coincidencias exactas, mueva la barra hacia "Exact match only"; si por el contrario desea obtener coincidencias parciales, mueva la barra hacia el otro extremo. Tener coincidencias parciales permitirá que la herramienta resuelva los nombres ingresados así existan algunos errores de escritura o formato. **F.** Aquí puede seleccionar una o más fuentes contras las cuales desea que los nombres científicos sean comparados. **G.** Aquí puede seleccionar el sistema de clasificación de su preferencia: APGIII o NCBI.

Paso 3

Ingrese los nombres en la herramienta. Para esto, vaya al archivo Excel que contiene los nombres científicos. De los 30 nombres, solo algunos están asociados a plantas, así que debe filtrarlos haciendo uso de la columna **Reino** (Figura 32). Tenga en cuenta que las tres especies que no pudo resolver con [Resolución Taxonómica y de LSIDs](#), también son plantas.

Archivo	Inicio	Insertar	Diseño de página	Fórmulas	Datos	Revisar	Vista	Esri Maps
D49								
	C	D	E	F				
2	Nombre científico	Reino	Filo	Clase				
3	Aeschynomene americana glandulosa (Poir. ex Lam.) Rudd	-	-	-				
4	Cordia bullata subsp. bullata (L.) Roem. Y Schult.	-	-	-				
5	Cupania sylvatica var. sylvatica Seem.	-	-	-				
29	Condylopodium hyalinifolium Diaz & Méndez	Plantae	Magnoliophyta	Magnoliopsida				
30	Convolvulaceae	Plantae	Magnoliophyta	Magnoliopsida				
31	Cydista aequinoctialis (L.) Miers	Plantae	Magnoliophyta	Magnoliopsida				
32	Huperzia reflexa (Lam.) Trevis.	Plantae	Lycopodiophyta	Lycopodiopsida				

Figura 32. Filtro en Excel de la información a validar en TNRS.

Paso 4

Copie los nombres científicos de la columna **Nombre científico** (sin incluir el nombre de la columna) y péguelos en el cuadro de texto de la aplicación (Figura 33).

Figura 33. Ingreso de información en TNRS a ser validada.

Nota: Para este ejercicio mantenga la herramienta con la configuración por defecto.

Paso 5

Haga clic en el botón “**Submit List**” para obtener los resultados. Deberá obtener una tabla como la Figura 34.

Name Submitted	Name Matched	Name Source	Overall Score	Taxonomic Status	Accepted Name	Details
Aeschynomene americana glandulosa (Poir. ex Lam.) Rudd (+1 more)	Aeschynomene americana var. glandulosa (Poir. ex Lam.) Rudd (+1 more)	TROPICOS	.97	Synonym	Aeschynomene americana L.	Details
Cordia bullata subsp. bullata (L.) ...	Cordia bullata subsp. bullata	TROPICOS	1.00	Accepted	Cordia bullata subsp. bullata	Details
Cupania sylvatica var. sylvatica S...	Cupania sylvatica var. sylvatica	TROPICOS	1.00	Accepted	Cupania sylvatica var. sylvatica	Details
Condylopodium hyalinifolium Dia...	Condylopodium hyalinifolium S.Díaz & G.P.Méndez (+1 more)	GCC	.92	Accepted	Condylopodium hyalinifolium S.Díaz	Details
Convolvulaceae	Convolvulaceae (+1 more)	USDA	1.00	Accepted	Convolvulaceae	Details
Cydista aequinoctialis (L.) Miers	Cydista aequinoctialis (L.) Miers (+1 more)	USDA	1.00	Accepted	Cydista aequinoctialis (L.) Miers	Details
Huperzia reflexa (Lam.) Trevis.	Huperzia reflexa (Lam.) Trevis. (+1 more)	USDA TROPICOS	1.00	Accepted	Huperzia reflexa (Lam.) Trevis.	Details

Figura 34. Resultados obtenidos en TNRS. **A.** Botón para descargar los resultados completos.

Paso 6

Descargue los resultados completos haciendo clic en el botón “**Download results**”. Ajuste las preferencias de descarga en la ventana emergente (Figura 35).

Figura 35. (Izquierda) Ventana emergente para configurar preferencias de descarga en TNRS. **A.** Para un mismo nombre científico puede haber más de una coincidencia. En este caso, seleccione “Best matches only” para obtener la coincidencia de mejor puntaje para cada nombre. **B.** Usted puede descargar los resultados en un formato resumido o detallado. Seleccione el formato “Simple”. **C.** Tenga en cuenta el formato de codificación del archivo para que pueda visualizar correctamente los resultados. Hágalo con UTF-8.

Paso 7

Finalmente haga clic en el botón “**OK**” para que inicie la descarga del archivo TXT con los resultados.

Paso 8

Visualice los resultados abriendo el archivo TXT desde la aplicación Excel (Archivo > Abrir...). De esta forma se mantiene el formato y no se altera la información. Hágalo con el **"Asistente para importar texto"** de Excel con los siguientes pasos:

- a. Ajuste el formato de codificación (UTF-8), luego haga clic en **Siguiente**.

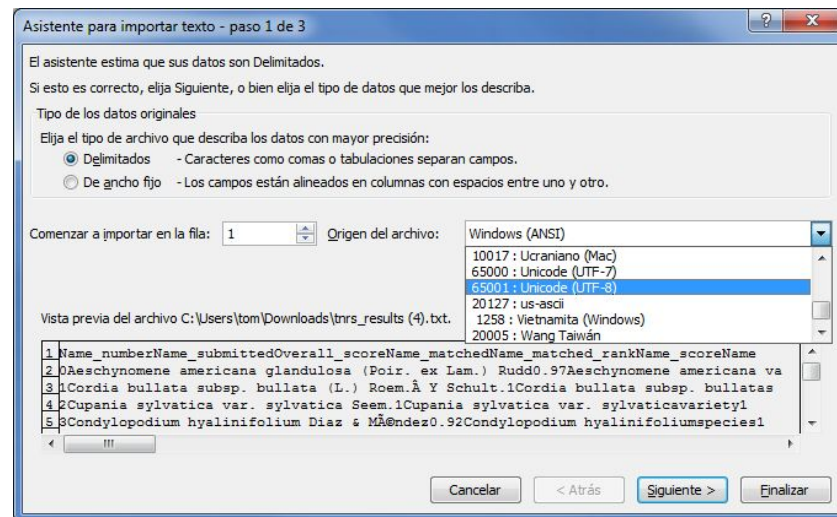


Figura 36. Ajuste de codificación de datos en Excel en UTF-8.

Nota: Si encuentra caracteres extraños como **"S.DÃ-az & G.P.MÃ!cndez"**, abra el archivo nuevamente y asegúrese de seleccionar el formato de codificación en el cual se realizó la descarga de los resultados - UTF-8.

- b. El archivo que descargó está delimitado por tabulaciones, en la lista de separadores escoja **Tabulación**, luego haga clic en **Siguiente**.

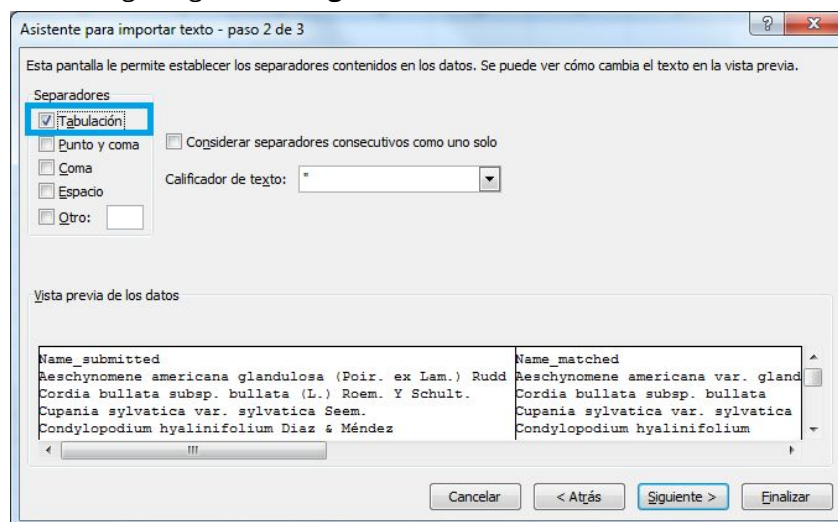


Figura 37. Ajuste de separadores de datos en Excel.

- c. Seleccione para todas las columnas el formato **Texto**, esto evitará que Excel haga interpretaciones erróneas del formato, luego haga clic en **Siguiente**.

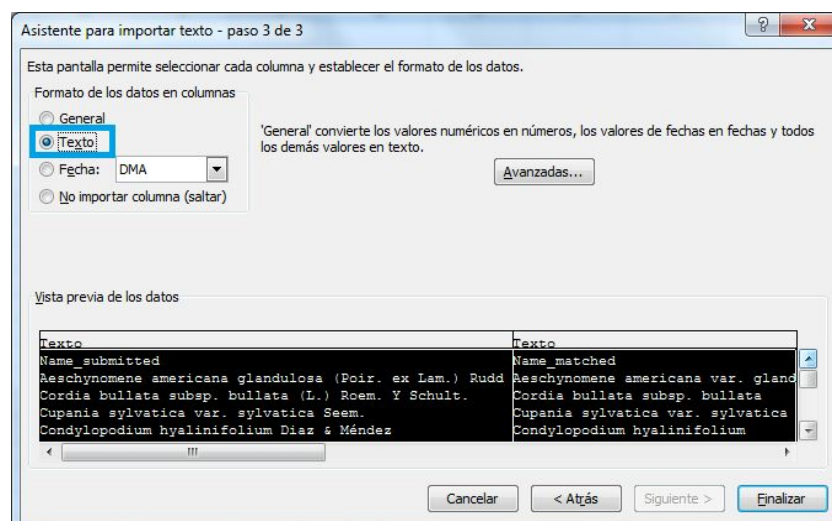



Figura 38. Ajuste de formato de los datos en Excel.

En la Tabla 4 se describen los elementos de la tabla de resultados.

Tabla 4. Elementos de la tabla de resultados obtenidos desde TNRS

<i>Name_submitted</i>	El nombre científico que fue ingresado.
<i>Name_matched</i>	Nombre científico con el mejor puntaje.
<i>Author_matched</i>	Autoría del nombre científico con el mejor puntaje.
<i>Overall_score</i>	Este puntaje indica la similitud entre el nombre original, y el encontrado incluida la autoría, donde 1 corresponde a la mejor coincidencia.
<i>Taxonomic_status</i>	El estado taxonómico del nombre con el mejor puntaje. Por ejemplo: aceptado, inválido, sinónimo.
<i>Accepted_name</i>	El nombre canónico aceptado para el taxón.
<i>Accepted_author</i>	Autor del nombre canónico aceptado para el taxón.
<i>Accepted_family</i>	Familia aceptada para el nombre con el mejor puntaje.
<i>Source</i>	La fuente donde se encontró el nombre con el mejor puntaje. Por ejemplo: TROPICOS , USDA , GCC
<i>Warnings</i>	Si no hay coincidencias aparecerá una alerta. En la vista en línea de los resultados estas se indican con el icono  .
<i>Accepted_name_Isid</i>	El identificador del nombre aceptado para el taxón. Sólo cuando la fuente es GCC , se obtiene este identificador.

Paso 9

Identifique la información útil para mejorar la calidad de sus datos. En la Figura 39 se destacan algunas.

ArchivoInicioInsertarDiseño de páginaFórmulasDatosRevisarVistaEsri Maps					
E21					
	B	E	F	G	H
1	Name_matched	Taxonomic_status	Accepted_name	Accepted_author	Accepted_family
2	Aeschynomene americana var. glandulosa	Synonym	Aeschynomene americana	L.	Fabaceae
3	Cordia bullata subsp. bullata	Accepted	Cordia bullata subsp. bullata		Boraginaceae
4	Cupania sylvatica var. sylvatica	Accepted	Cupania sylvatica var. sylvatica		Sapindaceae
5	Condylopodium hyalinifolium	Accepted	Condylopodium hyalinifolium	S.Díaz & G.P.Méndez	Asteraceae
6	Convolvulaceae	Accepted	Convolvulaceae		
7	Cydistia aequinoctialis	Accepted	Cydistia aequinoctialis	(L.) Miers	Bignoniaceae
8	Huperzia reflexa	Accepted	Huperzia reflexa	(Lam.) Trevis.	Lycopodiaceae
9					
10					

Figura 39. Interpretación de los resultados en Excel obtenidos de TNRS. **A.** Revisando los valores de la columna **Taxonomic_status** puede determinar que la subespecie *Aeschynomene americana* var. *glandulosa* es un sinónimo de la especie *Aeschynomene americana*, actualice su archivo de trabajo con el nombre y autor aceptado. **B.** Use la columna **Accepted_author** para actualizar su archivo con las autorías aceptadas. **C.** Use la columna **Accepted_family** para actualizar su archivo con las familias aceptadas.

Paso 10

A partir de las familias validadas usted puede resolver la taxonomía superior de *Aeschynomene americana*, *Cordia bullata* subsp. *bullata* y *Cupania sylvatica* var. *sylvatica*, utilizando la herramienta [Resolución Taxonómica y de LSIDs](#). Hágalo a partir de las familias como se muestra en la Figura 40.

SiB PORTAL DE DATOS													
RESULTADOS													
Desplegando 1-3 de 3 resultados.													
Id	Nombre Ingresado	Reino	Filo	Clase	Orden	Familia	Género	Epíteto Específico	Epíteto Infraespecífico	Categoría	Autor	Nombre Científico	LSID
1	Fabaceae	Plantae	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	-	-	-	family		Fabaceae	urn:lsid:catalogueoflife.org:29c1-102b-9a4a-00304854f820:col2012012
2	Boraginaceae	Plantae	Magnoliophyta	Magnoliopsida	Lamiales	Boraginaceae	-	-	-	family		Boraginaceae	urn:lsid:catalogueoflife.org:29c1-102b-9a4a-00304854f820:col2012012
3	Sapindaceae	Plantae	Magnoliophyta	Magnoliopsida	Sapindales	Sapindaceae	-	-	-	family		Sapindaceae	urn:lsid:catalogueoflife.org:29c1-102b-9a4a-00304854f820:col2012012

Figura 40. Resolución de taxonomía superior a partir del Servicio de Resolución Taxonómica y LSID's del SiB.

Para tener en cuenta: [TNRS](#) es una de las herramientas más robustas para resolver nombres científicos adicionalmente la posibilidad de ajustar la sensibilidad de la búsqueda y proveer un puntaje de similitud permite que el usuario pueda tomar una mejor decisión sobre la validez de los nombres ingresados. Por otro lado la herramienta no provee la taxonomía superior más allá del nivel de familia, lo cual es una desventaja ya que hace necesario utilizar otras herramientas para completar la información taxonómica. Con esta herramienta usted podrá estandarizar y estructurar los siguientes **elementos del estándar Dwc**: **Nombre científico**, **Familia**, **Género**, **Epíteto específico**, **Epíteto infraespecífico**, **Categoría del taxón** y **Autoría del nombre científico**

Recursos adicionales

Instrucciones completas para el uso de TNRS:

<http://tnrs.iplantcollaborative.org/instructions.html#Status>

List Matching Service

Después de validar los nombres de plantas con TNRS usted puede estar seguro, que a la fecha, corresponden a nombres aceptados. Muy seguramente le gustaría asegurarse que los nombres zoológicos también lo fueran. Para lograrlo usted va a utilizar la herramienta [List Matching Service](#).

[List Matching Service](#) es una herramienta de [Catalogue of Life](#) que permite comparar de manera masiva nombre científicos debajo del nivel de género con la lista dinámica de [Catalogue of Life](#) incluyendo sinónimos y nombres aceptados.

Paso 1

Ingrese a la aplicación [List Matching Service](#) y familiarícese con la herramienta. Identifique los recuadros de la siguiente imagen.

The screenshot shows the 'List Matching Service' interface. At the top, the title 'List Matching Service' is on the left and the '4D4Life' logo is on the right. Below the title, there are several input fields and a 'send' button. Label A points to the 'Column delimiter:' dropdown menu, which is currently set to 'Space'. Label B points to the 'First row contains headers:' checkbox, which is unchecked. Label C points to the 'Only accepted names:' checkbox, which is also unchecked. Label D points to the 'Input type:' section, which has two radio buttons: 'File upload' (selected) and 'Text area' (unselected). Below these radio buttons, there are two buttons: 'Seleccionar archivo' and 'Ningún archivo seleccionado'. Label E points to the 'send' button located at the bottom right of the form.

Figura 41. Vista de la página de inicio de List Matching Service de [Catalogue of Life](#). **A.** Aquí puede indicar cómo están delimitadas las columnas de su archivo (Ejemplo: espacio, tabulación). **B.** Al seleccionar esta opción le está indicando a la herramienta si las columnas del archivo tienen encabezados. **C.** Al seleccionar esta opción le está indicando a la herramienta que los resultados no deben incluir sinónimos, solo nombres aceptados. **D.** Aquí puede subir un archivo de texto con los datos (file upload) o pegarlos en un área de texto (text area). **E.** Botón para ejecutar la aplicación.

Paso 2

Ingrese los nombres en la herramienta. Para esto, vaya al archivo Excel que contiene los 30 nombres científicos estandarizados en la práctica con TNRS, copielos de la columna **Nombre científico** (sin incluir el nombre de la columna) y peguelos en el cuadro de texto de la aplicación.

Nota: No seleccione ninguna opción, ya que en este ejercicio no pegará los encabezados y le interesa identificar los sinónimos presentes en sus datos.

Paso 3

Con los nombre ya ingresados en la herramienta, haga clic en el botón “**send**” para obtener los resultados. La aplicación deberá mostrarle una tabla similar a la Figura 42.

Matched Species			Download to file
Your Data	Scientific Name	Status	
Aeschynomene americana	<i>Aeschynomene americana</i> L.	accepted name	
Caecilia subnigricans Dunn, 1942	<i>Caecilia subnigricans</i> Dunn, 1942	accepted name	
Caiman crocodilus (Linnaeus, 1758)	<i>Caiman crocodilus</i> (Linnaeus, 1758)	accepted name	
Caiman crocodilus (Linnaeus, 1758)	<i>Caiman crocodilus</i> Fuchs, 1974	synonym	
Caiman crocodilus (Linnaeus, 1758)	<i>Caiman crocodilus</i> Fuchs, 1971	synonym	
Caiman crocodilus (Linnaeus, 1758)	<i>Caiman crocodilus</i> Müller & Hellmich, 1936	synonym	
Clelia clelia Daudin, 1803	<i>Clelia clelia</i> (Daudin, 1803)	accepted name	
Clelia clelia Daudin, 1803	<i>Clelia clelia</i> Dunn, 1944	synonym	
Clelia clelia Daudin, 1803	<i>Clelia clelia</i> Greer, 1965	synonym	
Clelia clelia Daudin, 1803	<i>Clelia clelia</i> Smith, 1942	synonym	

Figura 42. Vista de resultados en la página de de List Matching Service. **Nota:** la herramienta cuenta con un botón de descarga que exporta la lista de nombres ingresados, no exporta los resultados.

Al final de la tabla aparecen los nombres que no tuvieron coincidencia (**Unmatched Species**). En la Figura 43 se muestra una imagen de los nombres científicos que no se resolvieron para este ejercicio y una breve explicación del por qué.

Unmatched Species	
Anyphaenidae	A.
Convolvulaceae	
Cordia bullata subsp. bullata (L.) Roem. Y Schult.	B.
Cupania sylvatica var. sylvatica Seem.	
Thaumasia	A.

Figura 43. Resultados de los nombres no coincidentes en List Matching Service. **A.** La herramienta sólo resuelve nombres científicos a nivel de especie, por lo cual las familias no fueron validadas. **B.** Estas especies aún no se encuentran registradas en *Catalog of Life*. Sin embargo, ya fueron validadas con TNRS.

Paso 4

A partir de los resultados identifique si debe actualizar algún nombre. En la Figura 44 se muestra un ejemplo de cómo interpretar la tabla de resultados.

Your Data	Scientific Name	Status	
Clelia clelia Daudin, 1803	<i>Clelia clelia</i> (Daudin, 1803)	accepted name	A.
Clelia clelia Daudin, 1803	<i>Clelia clelia</i> Dunn, 1944	synonym	
Clelia clelia Daudin, 1803	<i>Clelia clelia</i> Greer, 1965	synonym	
Liophis dilepis Cope, 1862	<i>Liophis dilepis</i> Hoge Et Al., 1981	synonym	B.
Huperzia reflexa (Lam.) Trevis.	<i>Huperzia reflexa</i> (Lam.) Rothm.	synonym	
Huperzia reflexa (Lam.) Trevis.	<i>Huperzia reflexa</i> (Lam.) Trevis.	synonym	

Figura 44. Resultados de los nombres buscados y coincidentes en el List Matching Service. **A.** Si el nombre ingresado es aceptado se listaran tanto el nombre aceptado encontrado en la base de datos, como los sinónimos. **B.** Si el nombre ingresado es un sinónimo solo se mostrara las diferentes formas de dicho sinónimo. El nombre aceptado no aparece como con TNRS, por lo cual se debe hacer una búsqueda manual para obtenerlo.

Paso 5


Para los nombres científicos que fueron catalogados como sinónimos, es necesario hacer una búsqueda simple en *Catalogue of Life* para determinar el nombre aceptado. Para ello, ingrese a <http://www.catalogueoflife.org/col/> y realice la búsqueda. La Figura 45 le da una idea de cómo hacer esta búsqueda y cómo se visualiza el resultado.

Buscar la edición Catalogue of Life

Buscar por:

☐ Búsqueda exacta

A.

Nombre ▲	Rango	Estado del nombre	Grupo	Base de datos de origen
<i>Liophis dilepis</i> Hoge Et Al., 1981	Especie	sinónimo para <i>Lygophis dilepis</i> Cope, 1862	Animalia	

B.

Figura 45. A. Búsqueda simple y **B.** Resultado en Catalogue of Life.

Conclusiones finales

Las herramientas utilizadas en las prácticas de Estandarización y validación taxonómica le permitieron pasar de manera masiva de una lista simple de nombres científicos a un conjunto de datos de 15 elementos completamente estructurados y validados bajo el estándar Darwin core. Cuando realice la validación de sus propios datos tenga en cuenta los alcances y limitaciones de cada herramienta, esto le permitirá escoger la ruta de trabajo más eficiente para tener datos de mejor calidad.

Estandarización y Validación de Información Geográfica

Sin duda los procesos de estructuración y validación de información geográfica representan una inversión de tiempo considerable a la hora de dar calidad a un conjunto de datos de biodiversidad. Los investigadores y científicos reconocen la dificultad que existe en ocasiones para manejar esta información debido a los diferentes formatos en que se recopila, por ejemplo para el caso de las coordenadas (geográficas, planas, decimales, entre otros); o de otro lado las inconsistencias que se evidencian al documentar las entidades geográficas asociada a esas coordenadas y registros. Esta información es fundamental para que los registros biológicos presenten espacialidad y de esta manera sea posible realizar análisis posteriores, ajustados a la realidad, que tengan en cuenta la posición o distribución de las especies en diferentes entidades y escalas geográficas. Actualmente existen desarrollos y herramientas que simplifican la estandarización y validación de esta información lo cual se evidencia en ganancia de tiempo para realizar procesos masivos de estos datos y en calidad en estos datos para su reuso.

Conversión de Coordenadas - *Canadensys*

Esta herramienta fue creada por el equipo de [Canadensys](http://data.canadensys.net) para realizar la conversión de coordenadas geográficas (grados, minutos y segundos) a grados decimales. De esta forma, el nodo GBIF para Canadá espera darle un valor agregado a los registros publicados. En la web existen varias herramientas que permiten realizar esta conversión, pero no de manera masiva. Esta aplicación recibe dos tipos de valores de entrada y realiza el proceso de manera masiva.

Objetivo: Transformar y estandarizar coordenadas geográficas GMS a grados decimales, de manera automática y masiva.

Enlace al programa: <http://data.canadensys.net/tools/coordinates>

Caso Práctico

Usted ha sido contratado por una organización para lograr la publicación de varios conjuntos de datos a través de GBIF en un tiempo récord. Desafortunadamente los conjuntos de datos no se encuentran estandarizados y mucho menos en Darwin Core. Su primer reto es transformar las N coordenadas geográficas en grados decimales en el menor tiempo posible.

Paso 1

Ingresa a la aplicación [Conversión de coordenadas](http://data.canadensys.net/tools/coordinates) y familiarícese con la herramienta. Identifique los recuadros de la Figura 46.

The screenshot shows the 'Coordinate conversion' web application. The main area is a large text input field containing the coordinates '45° 32' 25" N, 129° 40' 31" W'. Below the input field is a 'Submit' button. To the right of the input field is a sidebar with a table of links: 'Coordinate conversion', 'Date parsing', 'Tools API', and 'About'. Below the links is a section titled 'Example input' showing two rows of coordinates: '1 | 45.5° N, 129.6° W' and '2 | 40°26'47" N, 74° 0' 21.5022" W'.

Coordinate conversion	
1	45.5° N, 129.6° W
2	40°26'47" N, 74° 0' 21.5022" W

Figura 46. Vista página de inicio de la aplicación de Conversión de coordenadas. **A.** Cuadro de texto para ingresar o pegar las coordenadas que se desean convertir. **B.** Enlaces a otras herramientas ofrecidas por Canadensys y un breve resumen de esta herramienta. **C.** Un ejemplo de los valores que

puede convertir la aplicación. Puede intentar copiar el bloque de coordenadas y pegarlas en cuadro de texto.

Nota 1: Los valores de entrada para realizar la conversión a grados decimales puede ser de dos formas: (i) la pareja de coordenadas concatenadas en una sola celda o (ii) las coordenadas en celdas independientes.

Nota 2: Es importante que las coordenadas a convertir estén realmente en grados, minutos y segundos. La herramienta no recibe otros valores. Por ejemplo: coordenadas planas.

Paso 2

Antes de descargar los datos verifique que su equipo lea las cifras decimales con punto "." y no coma "," (Figura 47). Esto se realiza dirigiéndose a Inicio > Panel de Control > Configuración regional y de Idioma > Configuración adicional > Símbolo decimal: verifique o coloque "." como separador decimal; igualmente escoja "," como Símbolo de separación de miles.

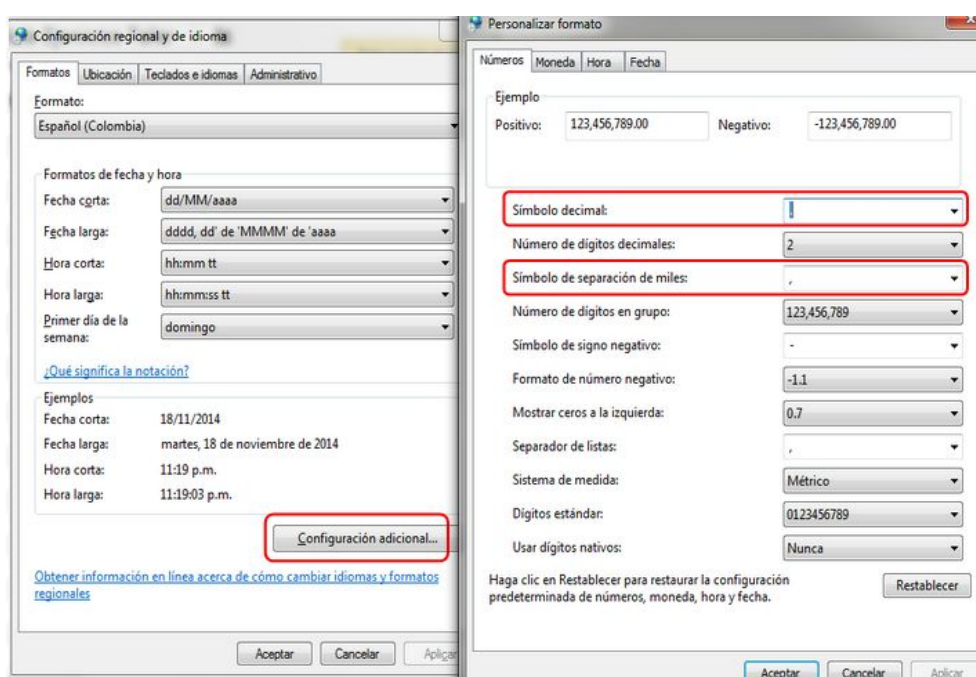


Figura 47. Configuración en Windows del símbolo empleado para separar cifras decimales.

Paso 3

Descargue en su equipo el archivo Excel "[Datos ejemplo geo.xlsx](#)" (clic en el hipervínculo). Ahora ábralo y detalle la información que contiene el archivo.

Usted encontrará una tabla con 18 columnas que documentan información de 30 registros biológicos, cada uno se encuentra georeferenciado con información básica de su ubicación (Departamento y Municipio) y su taxonomía. Revise los elementos: **Coordenadas originales**, **Latitud original**, **Longitud original** podrá encontrar inconsistencias claras, como por ejemplo, espacios en blanco dentro de algunas coordenadas. Sin embargo, la conversión de coordenadas se puede realizar a partir de diferentes valores de entrada (incluyendo este tipo de inconsistencias).

Paso 4

Ahora ingrese las coordenadas geográficas en la aplicación. Para esto, vaya al archivo Excel que contiene las 30 coordenadas, copie la información de la columna **Coordenadas originales** (sin

incluir el nombre de la columna) y péguela en el cuadro de texto de la aplicación. Deberá verse como la Figura 48.

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.

7° 09' 04.70" N, 73° 32' 24.40" W
7° 11' 39.56" N, 73° 54' 23.69" W
7° 13' 40.00" N, 73° 50' 22.26" W
7° 30' 52.29" N, 74° 53' 33.90" W
7° 57' 30.20" N, 75° 11' 52.99" W
6° 08' 27.34" N, 75° 24' 40.19" W
3° 41' S, 70° 15' W
3° 44' 53.8" S, 70° 27' 35.24" W
3° 46' 47.5" S, 70° 20' 24.8" W
3° 46' 52" S, 70° 21' 10.9" W
1° 56' 21.3" N, 76° 35' 52.7" W
1° 04' N, 76° 44" W
1° 56' 21.3" N, 76° 35' 52.7" W
0° 7' S, 74° 56' W
6° 49' 52.9" N, 71° 5' 55.5" W

Submit

Coordinate conversion

Date parsing

Tools API

About

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

Figura 48. Coordenadas en formato GMS copiadas de la columna **Coordenadas originales**, listas para ser convertidas.

Nota: También puede hacerlo copiando la información de las columnas **Latitud original**, y **Longitud original**; obtendrá el mismo resultado.

Paso 5

Haga clic en el botón **“Submit”** para obtener los resultados. Aparecerán los resultados de la conversión de las coordenadas en la herramienta, mostrando, la información original y transformada. Deberá obtener una tabla similar a la Figura 49.

Coordinate conversion results

original	decimalLatitude	decimalLongitude
5° 48' 16.9" N, 73° 6' 17.6" W	5.8046944	-73.1048889
5° 49' 20.2" N, 73° 9' 20.7" W	5.8222778	-73.15575
7° 30' 52.29" N, 74° 53' 33.90" W	7.514525	-74.89275
7° 57' 30.20" N, 75° 11' 52.99" W	7.9583889	-75.1960528
9° 13' 14.13" N, 75° 47' 56.94" W	9.2205917	-75.79915
9° 24' 05" N, 75° 40' 48" W	9.4013889	-75.68
6° 26' 15" N, 75° 19' 50" W	6.4375	-75.3305556
6° 06' 01.13" N, 75° 29' 39.00" W	6.1003139	-75.4941667
6° 08' 27.34" N, 75° 24' 40.19" W	6.1409278	-75.4111639
6° 12' 44.80" N, 75° 29' 52.38" W	6.2124444	-75.4978833
6° 13' 40.88" N, 75° 19' 58.49" W	6.2280222	-75.3329139
7° 04' 16.68" N, 73° 51' 35.00" W	7.0713	-73.8597222
7° 06' 06.03" N, 73° 24' 32.37" W	7.101675	-73.4089917
7° 06' 21.41" N, 73° 24' 57.69" W	7.1059472	-73.416025
7° 08' 51.39" N, 73° 28' 27.46" W	7.1476083	-73.4742944
7° 09' 04.70" N, 73° 32' 24.40" W	7.1513056	-73.5401111
7° 11' 39.56" N, 73° 54' 23.69" W	7.1943222	-73.9065806
7° 13' 40.00" N, 73° 50' 22.26" W	7.2277778	-73.8395167

Coordinate conversion

Date parsing

Tools API

About

Figura 49. Visualización de los resultados en la herramienta *Conversión de coordenadas*. Los resultados de la transformación de coordenadas se presentan en dos secciones. **A.** Columna con las coordenadas originales. **B.** Columnas con las coordenadas transformadas a grados decimales y separadas en latitud y longitud.

Paso 6

Abra una hoja nueva de Excel, seleccione las tres primeras columnas y elija la opción de valores **“Texto”** para evitar que los datos a pegar se interpreten de modo diferente (Figura 50). Seleccione la tabla de Resultados de la conversión de coordenadas en Canadensys (Figura 49), copíela (incluyendo los encabezados) y péguela como texto (clic derecho **“Pegado especial”** > texto) en la hoja de Excel nueva.

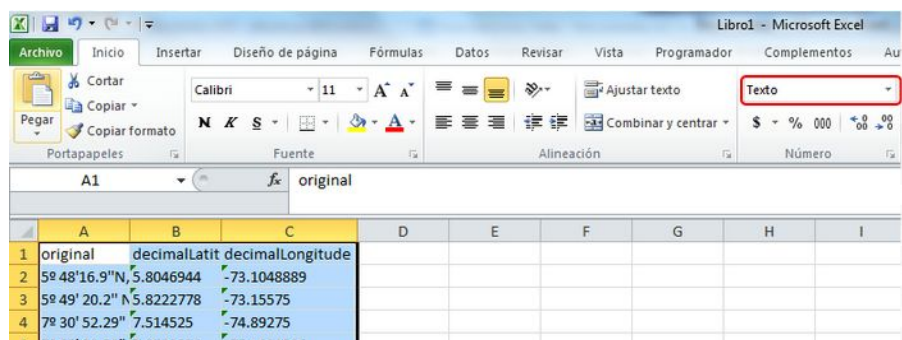


Figura 50. Inserción de los resultados de la aplicación *Conversión de coordenadas* en Excel.

Paso 7

En este punto usted ya tiene 30 coordenadas geográficas transformadas a grados decimales en tiempo record! Ya puede completar las columnas **Latitud decimal** y **Longitud decimal** en el archivo Excel "*Datos_ejemplo_geo.xlsx*". Guarde la información ya que empleará este archivo en los ejercicios posteriores.

	A	B	C	D	E	F
1	ID del registro biológico	Coordenadas originales	Latitud original	Longitud original	Latitud decimal	Longitud decimal
2	UDEA:CLUA:ARTR:COLE:0001	5° 48' 16.9" N, 73° 6' 17.6" W	5° 48' 16.9" N	73° 6' 17.6" W	5.8046944	-73.1048889
3	UDEA:CLUA:ARTR:COLE:0002	5° 49' 20.2" N, 73° 9' 20.7" W	5° 49' 20.2" N	73° 9' 20.7" W	5.8222778	-73.15575
4	UDEA:CLUA:ARTR:COLE:0003	7° 30' 52.29" N, 74° 53' 33.90" W	7° 30' 52.29" N	74° 53' 33.90" W	7.514525	-74.89275
5	UDEA:CLUA:ARTR:COLE:0004	7° 57' 30.20" N, 75° 11' 52.99" W	7° 57' 30.20" N	75° 11' 52.99" W	7.9583889	-75.1980528
6	UDEA:CLUA:ARTR:COLE:0005	9° 13' 14.13" N, 75° 47' 56.94" W	9° 13' 14.13" N	75° 47' 56.94" W	9.2205917	-75.79915
7	UDEA:CLUA:ARTR:COLE:0006	9° 24' 05" N, 75° 40' 48" W	9° 24' 05" N	75° 40' 48" W	9.4013889	-75.68
8	UDEA:CLUA:ARTR:COLE:0007	6° 26' 15" N, 75° 19' 50" W	6° 26' 15" N	75° 19' 50" W	6.4375	-75.3305556
9	UDEA:CLUA:ARTR:COLE:0008	6° 06' 01.13" N, 75° 29' 39.00" W	6° 06' 01.13" N	75° 29' 39.00" W	6.1003139	-75.4941667
10	UDEA:CLUA:ARTR:COLE:0009	6° 08' 27.34" N, 75° 24' 40.19" W	6° 08' 27.34" N	75° 24' 40.19" W	6.1400278	-75.4111639

Figura 51. Coordenadas transformadas en la aplicación *Conversión de coordenadas* agregadas al archivo "*Datos_ejemplo_geo.xlsx*".

Para tener en cuenta: Aunque para este ejercicio no se realizó, tenga presente que la herramienta también permite que cada fila o registro cuente con un ID. Esta opción es útil en caso que usted requiera vincular las coordenadas con un ID en su conjunto de datos (Figura 52).

A		B		Coordinate conversion results			
1	ID del registro biológico	Coordenadas originales		id	original	decimalLatitude	decimal
2	UDEA:CLUA:ARTR:COLE:0001	5° 48' 16.9" N, 73° 6' 17.6" W		UDEA:CLUA:ARTR:COLE:0001	5° 48' 16.9" N, 73° 6' 17.6" W	5.8046944	-73.1048889
3	UDEA:CLUA:ARTR:COLE:0002	5° 49' 20.2" N, 73° 9' 20.7" W		UDEA:CLUA:ARTR:COLE:0002	5° 49' 20.2" N, 73° 9' 20.7" W	5.8222778	-73.15575
4	UDEA:CLUA:ARTR:COLE:0003	7° 30' 52.29" N, 74° 53' 33.90" W		UDEA:CLUA:ARTR:COLE:0003	7° 30' 52.29" N, 74° 53' 33.90" W	7.514525	-74.89275
5	UDEA:CLUA:ARTR:COLE:0004	7° 57' 30.20" N, 75° 11' 52.99" W		UDEA:CLUA:ARTR:COLE:0004	7° 57' 30.20" N, 75° 11' 52.99" W	7.9583889	-75.1980528

UDEA:CLUA:ARTR:COLE:0001
UDEA:CLUA:ARTR:COLE:0002
UDEA:CLUA:ARTR:COLE:0003
UDEA:CLUA:ARTR:COLE:0004
UDEA:CLUA:ARTR:COLE:0005
UDEA:CLUA:ARTR:COLE:0006
UDEA:CLUA:ARTR:COLE:0007
UDEA:CLUA:ARTR:COLE:0008
UDEA:CLUA:ARTR:COLE:0009
UDEA:CLUA:ARTR:COLE:0010
UDEA:CLUA:ARTR:COLE:0011
UDEA:CLUA:ARTR:COLE:0012
UDEA:CLUA:ARTR:COLE:0013
UDEA:CLUA:ARTR:COLE:0014
UDEA:CLUA:ARTR:COLE:0015
UDEA:CLUA:ARTR:COLE:0016

5° 48' 16.9" N, 73° 6' 17.6" W
5° 49' 20.2" N, 73° 9' 20.7" W
7° 30' 52.29" N, 74° 53' 33.90" W
7° 57' 30.20" N, 75° 11' 52.99" W
9° 13' 14.13" N, 75° 47' 56.94" W
9° 24' 05" N, 75° 40' 48" W
6° 26' 15" N, 75° 19' 50" W
6° 06' 01.13" N, 75° 29' 39.00" W
6° 08' 27.34" N, 75° 24' 40.19" W
6° 12' 44.80" N, 75° 29' 52.38" W
6° 13' 40.88" N, 75° 19' 58.49" W
7° 04' 16.68" N, 73° 51' 35.00" W
7° 06' 06.03" N, 73° 24' 32.37" W
7° 06' 21.41" N, 73° 24' 57.69" W
7° 08' 51.39" N, 73° 28' 27.46" W
7° 09' 04.70" N, 73° 32' 24.40" W

Enviar

Figura 52. Registros con coordenadas acompañados de Identificadores y su conversión realizada con la herramienta de Canadensis.

GeoValidador SiB

Esta herramienta nace con el fin de optimizar y facilitar el proceso de validación de datos biológicos georeferenciados para Colombia, que rodea la publicación de datos de biodiversidad a través del SiB- Colombia. Actualmente se encuentra en una *versión BETA* y corre como un ejecutable de escritorio que debe ser almacenado directamente en el PC.

Consiste en una aplicación creada en JavaScript por el equipo de Infraestructura Informática del SiB- Colombia, alimentada con archivos cartográficos en formato shape proporcionados por el Instituto Geográfico Agustín Codazzi. Se está trabajando actualmente para que a futuro se valide completamente la toponimia de los departamentos y municipios de Colombia de acuerdo a la División Político Administrativa de Colombia, proporcionada por el Departamento Administrativo Nacional de Estadística - DANE. Una vez se tengan estos ajustes, la herramienta quedará disponible a través de una plataforma web para que cualquier persona la pueda utilizar.

Objetivo: Validar coordenadas geográficas respecto al Departamento y Municipio documentado para registros biológicos ubicados en Colombia.

Caso Práctico

Su institución tiene un conjunto de datos que desea publicar lo antes posible a través del SiB-Colombia. Para ello usted debe enviar estos registros que se encuentran georeferenciados y garantizar la validez de las coordenadas respecto al Departamento y Municipios documentados, pues este es un requisito para que sus datos sean aceptados; ¿lo podrá realizar en tiempo record?

Paso 1

Ubique y tenga presente que la carpeta “**Validaciones**” compartida para el taller esté almacenada en el disco C:// de su equipo. Esta carpeta siempre debe albergarse allí y resulta fundamental para que la herramienta se ejecute correctamente.

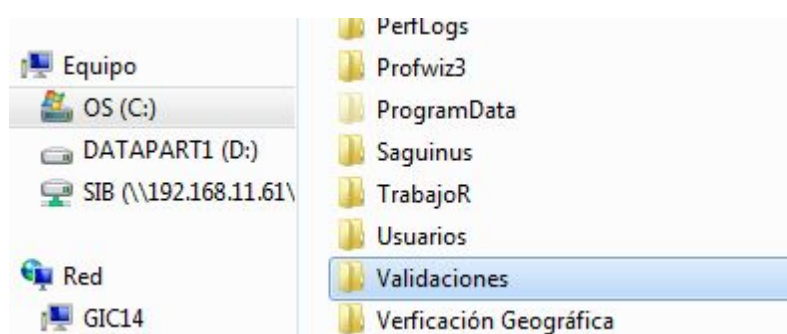


Figura 53. Ubicación correcta de la carpeta *Validaciones* en el PC

Paso 2

Del archivo resultante de Conversión de coordenadas - Canadensys (“*Datos_ejemplo_geo.xlsx*”) copie los datos de las columnas **Latitud decimal**, **Longitud decimal**, **Municipio** y **Departamento** (en este estricto orden y sin el encabezado de las columnas) en una nueva hoja de cálculo de Excel (Figura 54). Las celdas que se encuentren vacías deben documentarse con “NA”, de lo contrario el ejecutable arroja un error.

	A	B	C	D
1	5.804694	-73.104889	PAIPA	BOYACÁ
2	5.822278	-73.15575	SOTAQUIRA	CASANARE
3	7.514525	-74.89275	ZARAGOZA	ANTIOQUIA
4	7.958389	-75.198053	CAUCASIA	ANTIOQUIA
5	9.220592	-75.79915	LORICA	CÓRDOBA
6	9.401389	-75.68	COVEÑAS	SUCRE
7	6.4375	-75.330556	BARBOSA	ANTIOQUIA
8	6.100314	-75.494167	NA	CUNDINAMARCA

Figura 54. Elementos a validarse seleccionados en Excel.

Paso 3

Guarde este archivo como un archivo de texto (.txt) en la ruta **C://Validaciones/Municipios/input** con el nombre de **"municipios"** (por favor guarde el archivo exactamente con ese nombre).

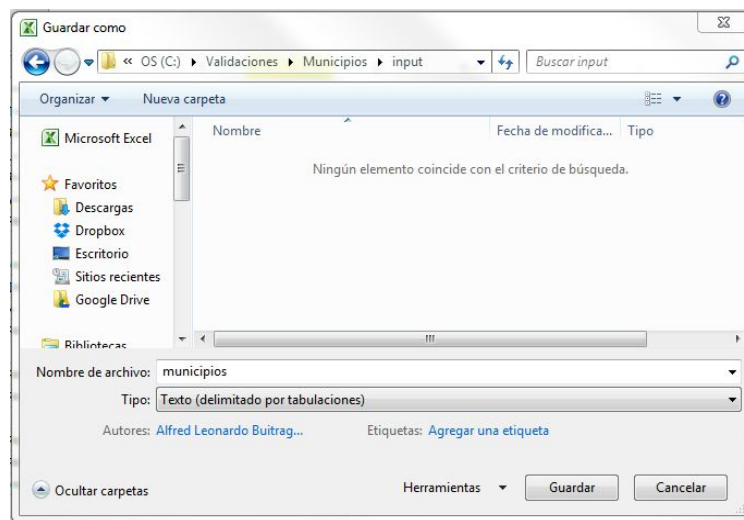


Figura 55. Guardado del archivo de texto en la carpeta *Validaciones*.

Haga clic en **"Aceptar"** y **"Si"** a las dos ventanas siguientes de que le indican que guardará un libro de Excel en formato .txt.

Paso 4

En la carpeta **C://Validaciones** encontrará un archivo ejecutable llamado **"Validar"**. Haga doble clic sobre él.

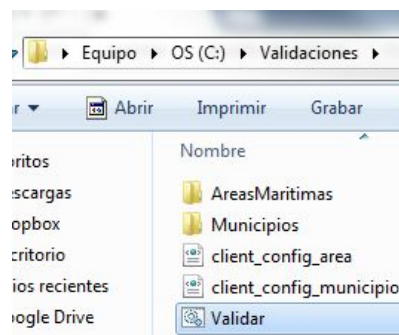
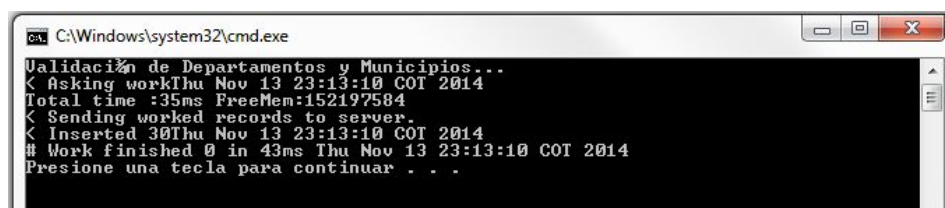


Figura 56. Archivo ejecutable "Validar"

Paso 5

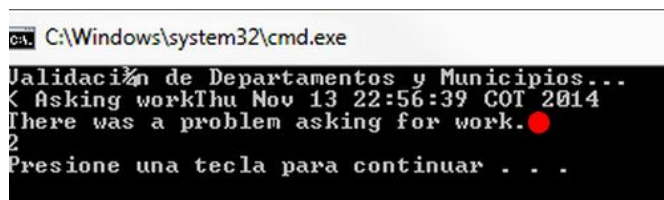
Se abrirá una ventana de comandos. En ella se debe presionar una tecla para continuar, tal como le indica el ejecutable (Figura 57). Hasta ese momento la herramienta valida los departamentos y municipios para Colombia.



```
C:\Windows\system32\cmd.exe
Validación de Departamentos y Municipios...
< Asking workThu Nov 13 23:13:10 COT 2014
Total time :35ms FreeMem:152197584
< Sending worked records to server.
< Inserted 30Thu Nov 13 23:13:10 COT 2014
# Work finished 0 in 43ms Thu Nov 13 23:13:10 COT 2014
Presione una tecla para continuar . . .
```

Figura 57. Archivo ejecutable de JavaScript indicando la validación inicial..

Nota: Tenga en cuenta que si los datos presentan algún problema como espacios vacíos o el orden en que está dispuesta la información, el sistema arrojará un error (Figura 58).

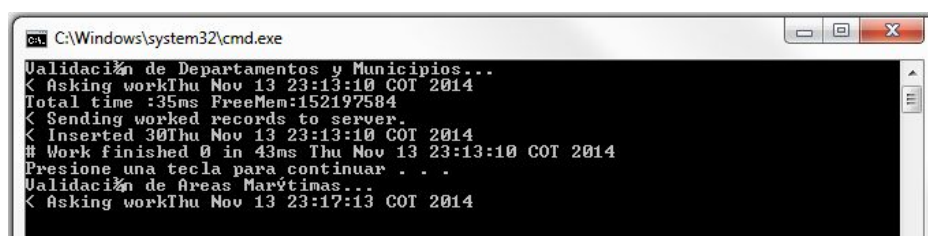


```
C:\Windows\system32\cmd.exe
Validación de Departamentos y Municipios...
< Asking workThu Nov 13 22:56:39 COT 2014
There was a problem asking for work.
Presione una tecla para continuar . . .
```

Figura 58. Detección de un error al ejecutar la aplicación, tal como indica la línea de texto del punto rojo.

Paso 6

En este momento la herramienta realiza la validación de áreas marítimas. **Nota:** Estas áreas aún se encuentran en proceso de ajuste para que la herramienta realice la validación a esta capa adecuadamente.



```
C:\Windows\system32\cmd.exe
Validación de Departamentos y Municipios...
< Asking workThu Nov 13 23:13:10 COT 2014
Total time :35ms FreeMem:152197584
< Sending worked records to server.
< Inserted 30Thu Nov 13 23:13:10 COT 2014
# Work finished 0 in 43ms Thu Nov 13 23:13:10 COT 2014
Presione una tecla para continuar . . .
Validación de Areas Marítimas...
< Asking workThu Nov 13 23:17:13 COT 2014
```

Figura 59. Validación de áreas marítimas (es una validación parcial que aún está en ajustes de desarrollo)

Paso 7

Cierre la ventana de comandos y diríjase a la carpeta **Municipios > output**. Allí encontrará un archivo de texto con el nombre **"municipiosValidacion_[fecha]"**.

Paso 8

Importe ese archivo de texto a una hoja de cálculo Excel. Puede simplemente arrastrarlo a la hoja, o ingrese en Excel y desde el menú **Datos** seleccione el archivo de texto (.txt).

Paso 9

En el archivo importado inserte una fila de encabezados al inicio y documéntela así:

Latitud | Longitud | Depto | Depto_Sugerido | Valid_Depto | Mpio | Mpio_Sugerido | Valid_Mpio

Estos nombres no son obligatorios textualmente, sólo se recomiendan para que usted sepa a qué corresponde cada columna y facilite el filtro de información.

	A	B	C	D	E	F	G	H
1	Latitud	Longitud	Depto	Depto_Sugerido	Valid_Depto	Mpio	Mpio_Sugerido	Valid_Mpio
2	5.8046944	-73.1048889	BOYACÁ	BOYACÁ	true	PAIPA	PAIPA	true
3	5.8222778	-73.15575	CASANARE	BOYACÁ	false	SOTAQUIRA	SOTAQUIRÁ	false
4	7.514525	-74.89275	ANTIOQUIA	ANTIOQUIA	true	ZARAGOZA	ZARAGOZA	true
5	7.9583889	-75.1980528	ANTIOQUIA	ANTIOQUIA	true	CAUCASIA	CAUCASIA	true
6	9.2205917	-75.79915	CÓRDOBA	CÓRDOBA	true	LORICA	LORICA	true

Figura 60. Inserción de fila de encabezados en Excel de los datos validados.

Paso 10

En las columnas de **Valid_Depto** y **Valid_Mpio** usted encontrará los dominios “true” y “false”. El primero indica que hubo correspondencia luego del proceso de validación y el segundo indica lo contrario.

Latitud	Longitud	Depto	Depto_Sugerido	Valid_Depto	Mpio	Mpio_Sugerido	Valid_Mpio
5.8222778	-73.15575	CASANARE	BOYACÁ	false	SOTAQUIRA	SOTAQUIRÁ	false
6.1003139	-75.4941667	CUNDINAMARCA	ANTIOQUIA	false	NA	RETIRO	false
7.0713	-73.8597222	ARAUCA	SANTANDER	false	BARRANCABERMEJA	BARRANCABERMEJA	false
7.101675	-73.4089917	NA	SANTANDER	false	NA	GIRÓN	false
7.1513056	-73.5401111	BOLIVAR	SANTANDER	false	NA	SAN VICENTE DE CHUCURÍ	false
7.514525	-74.89275	SANTANDER	ANTIOQUIA	false	RIONEGRO	ZARAGOZA	false
-3.7482778	-70.4597889	PUTUMAYO	AMAZONAS	false	PUERTO NARIÑO	PUERTO NARIÑO	false
-3.7798611	-70.3402222	NA	AMAZONAS	false	NA	PUERTO NARIÑO	false

Figura 61. Filtro de datos “false” (con errores).

Paso 11

Finalmente tome los datos validados de **Depto_sugerido**, **Mpio_Sugerido** y péguelos en el archivo **Datos_ejemplo_geo.xlsx** reemplazando los que había originalmente en las columnas **Departamento** y **Municipio**. De este modo finalizará el proceso de validación; guarde los cambios realizados en este archivo Excel pues lo empleará en el caso práctico de Carto DB.

Para tener en cuenta: En este ejercicio práctico se tomarán los datos sugeridos como totalmente válidos (“true”), pero dentro de los procesos de estandarización y publicación de datos usted debe tomarse el trabajo de corroborar y corregir los datos “false” a través de una herramienta web de visualización de coordenadas (por ejemplo: [GPS Visualizer](#), [CartoDB](#), [Esri Maps for Office](#), entre otros) o un Software de Información Geográfica (por ejemplo: ArcGIS, Quantum GIS, etc.). Si se valida el primer dato del archivo “*Datos_ejemplo_geo.xlsx*” a través de una de estas herramientas, se puede observar que la coordenada sí se encuentra mal documentada y realmente corresponde al departamento de Boyacá, (no a Casanare). Igualmente el municipio presenta un error ortográfico es su documentación que también fue validado gracias a esta herramienta.

	A	B	C	D	E	F	G	H
1	Latitud	Longitud	Depto	Depto_Sugerido	Valid_Depto	Mpio	Mpio_Sugerido	Valid_Mpio
3	5.8222778	-73.15575	CASANARE	BOYACÁ	false	SOTAQUIRA	SOTAQUIRÁ	false
9	6.1003139	-75.4941667	CUNDINAMARCA	ANTIOQUIA	false	NA	RETIRO	false




Figura 62. Verificación de errores “false” a través de una herramienta de visualización de datos geográficos (Esri Maps for Office).

CartoDB

CartoDB fue creada por *Vizzuality* para trabajar con datos geoespaciales y grandes cantidades de datos (big data) de una manera más rápida y sencilla. Este servicio web es una base de datos geoespacial *open source*, con una interfaz para importar y visualizar datos geoespaciales creando hermosos mapas dinámicos. Las personas con conocimientos más profundos en programación, bases de datos o sistemas de información geográfica, pueden sacarle más provecho a la herramienta, teniendo en cuenta que funciona internamente con PostGIS y que posee una serie de APIs (*Application Programming Interface*) para acceder a los datos mediante sentencias SQL.

CartoDB permite arrastrar y soltar un set de datos, o cargar un nuevo conjunto de datos, personalizar la visualización, realizar consultas SQL y embeber el mapa en un sitio web o blog de *WordPress*.

Objetivo: Visualizar de manera dinámica un conjunto de puntos en un mapa a partir de coordenadas geográficas e identificar puntos inusuales en un conjunto de datos.

Enlace al programa: <http://cartodb.com/>

Caso Práctico

Un proyecto de caracterizaciones biológicas tiene todos sus conjuntos de datos estructurados y publicados a través de GBIF. Sin embargo, el proyecto quiere visualizar específicamente la abundancia de sus especímenes de Moluscos en un mapa y a su vez verificar que los puntos caen en las zonas correctas. Usted ha sido llamado para este reto!

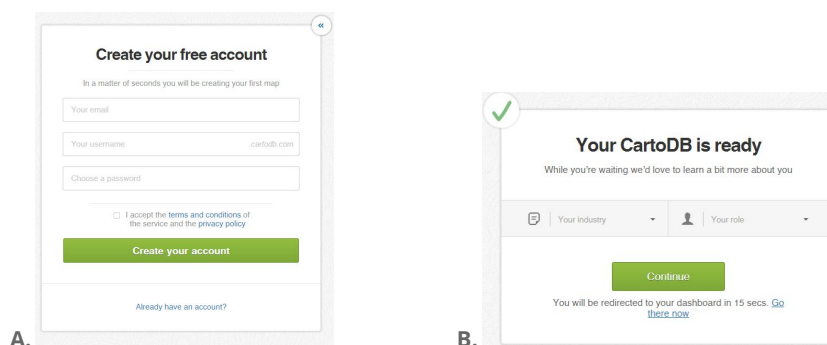
Nota: Para este ejercicio emplearemos el archivo "*Datos_ejemplo_geo.xlsx*", resultante de la práctica del Geovalidador SiB, es decir, con los campos **Latitud decimal** y **Longitud decimal** debidamente documentados y el Municipio y Departamento previamente validados.

Paso 1

Ingresa a la aplicación [CartoDB](#) y cree una cuenta de usuario. Para ello vaya a la pestaña superior derecha en **"SIGN UP"**. Diligencie la información solicitada y haga clic en **"Create your account"**.

Paso 2

Seguidamente aparecerá un aviso, que su cuenta ha sido creada exitosamente. Si lo desea, puede especificar su área de interés o de trabajo y su rol, de lo contrario haga clic en continuar.



A. **Create your free account**
In a matter of seconds you will be creating your first map
Your email
Your username: cartodb.com
Choose a password
☐ I accept the terms and conditions of the service and the privacy policy
Create your account
[Already have an account?](#)

B. **Your CartoDB is ready**
While you're waiting we'd love to learn a bit more about you
Your industry | Your role
Continue
You will be redirected to your dashboard in 15 secs. [Go there now](#)

Figura 63. Creación de cuenta en CartoDB. **A.** Datos personales a diligenciar **B.** Mensaje final de creación de cuenta.

Nota: La cuenta creada corresponde a una cuenta de usuario regular en la que se puede almacenar hasta 50 MB (Megabytes) de información y en este caso los mapas que comparta serán públicos.

Por ser un usuario nuevo, en la parte central encontrará en la página un video introductorio a la herramienta (para optimizar el tiempo de la práctica le sugerimos verlo en otro momento). En la parte superior encontrará el menú principal donde puede navegar entre sus tablas o datos almacenados (*tables*), la visualización geográfica de esa información (*visualizations*), o explorar y emplear otros datos compartidos por otras instituciones o usuarios (*common data*). Si en algún momento tiene alguna duda sobre el manejo o las múltiples opciones que ofrece la herramienta, CartoDB dispone toda la documentación necesaria (*documentation*). Si desea visualizar o modificar la información de su cuenta, vaya a su nombre, despliegue el submenú y seleccione “Account settings”.



Figura 64. Menú principal de CartoDB

Paso 3

En la página inicial de **CartoDB**, haga clic en el botón “**Create your first table**”, que se encuentra en la parte inferior de la página. De esta forma puede proceder a cargar la información que necesita visualizar.

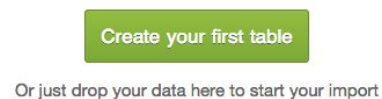


Figura 65. Botón de creación de su primera tabla en CartoDB.

Paso 4

Seguidamente se abre una ventana en la que podemos seleccionar nuestra fuente de información. En este ejercicio seleccione la opción “**Data file**”, ya que debe cargar el archivo “*Datos_ejemplo_geo.xlsx*” desde el equipo de computo. Haga clic en “**select a file**” para buscarlo y subirlo.

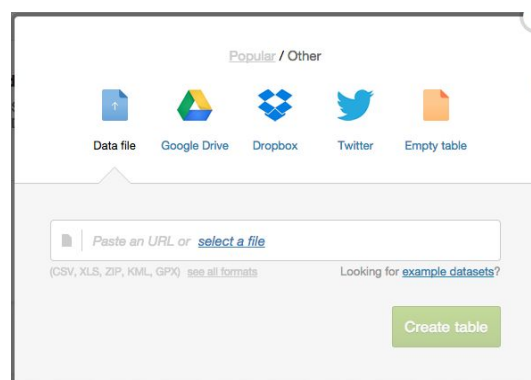


Figura 66. Ventana de selección de fuente de datos.

Nota: Se pueden cargar datos en diversos formatos, principalmente: CSV, XLS, KML, ZIP, GPX.

Paso 5

Una vez cargado el archivo “*Datos_ejemplo_geo.xlsx*”, se puede observar que la aplicación adiciona unas columnas de información por defecto: **cartodb_id**, un identificador para los datos de la tabla; **the_geom**, donde el programa debe reconocer el dato geográfico (latitud y longitud en este caso) y almacenarlo como tal (para ello de click sobre **GEO** en este encabezado y seleccione “*longitud_decimal*” y “*latitud_decimal*” como las columnas que almacenan sus coordenadas);

created_at, para la fecha de creación; y **updated_at**, para la fecha de actualización. Usted podrá editar la información de sus datos en cualquier momento haciendo doble clic sobre cualquiera de las celdas.



The screenshot shows the CartoDB interface with the 'DATA VIEW' tab selected. The table contains 7 rows of data with columns for cartodb_id, the_geom, categoria_del_taxon, clase, coordenadas_originales, departamento, epíteto_especifico, familia, and filo.

cartodb_id - number	the_geom - geometry	categoria_del_taxon - string	clase - string	coordenadas_originales - string	departamento - string	epíteto_especifico - string	familia - string	fillo - string
1	-73.1049, 5.8047	Género	Bivalvia	5° 48' 16.9" N, 73° 6' 17.6" W	BOYACÁ		Pteridae	Mollusca
2	-73.1557, 5.8223	Género	Bivalvia	5° 49' 20.2" N, 73° 9' 20.7" W	CASANARE		Ostreidae	Mollusca
3	-74.8928, 7.5145	Género	Bivalvia	7° 30' 52.29" N, 74° 53' 33.90" W	ANTIOQUIA		Pteridae	Mollusca
4	-75.1981, 7.9584	Género	Bivalvia	7° 57' 30.20" N, 75° 11' 52.99" W	ANTIOQUIA		Ostreidae	Mollusca
5	-75.7991, 9.2206	Género	Bivalvia	9° 13' 14.13" N, 75° 47' 56.94" W	CÓRDOBA		Pteridae	Mollusca
6	-75.6800, 9.4014	Género		9° 24' 05" N, 75° 40' 48" W	SUCRE		Gordiidae	Nematomorpha
7	-75.3306, 6.4375	Género	Hirudinea	6° 26' 15" N, 75° 19' 50" W	ANTIOQUIA		Glossiphoniidae	Annelida

Figura 67. Visualización de la tabla de datos en CartoDB

Paso 6

Para poder visualizar los datos que fueron cargados, haga clic en **"MAP VIEW"** que está en la parte superior de la página (puede moverse en todo momento entre sus datos **"DATA VIEW"** y el mapa **"MAP VIEW"**). En este caso la información mapeada corresponde a las coordenadas del archivo, las cuales se visualizan por defecto en color naranja.

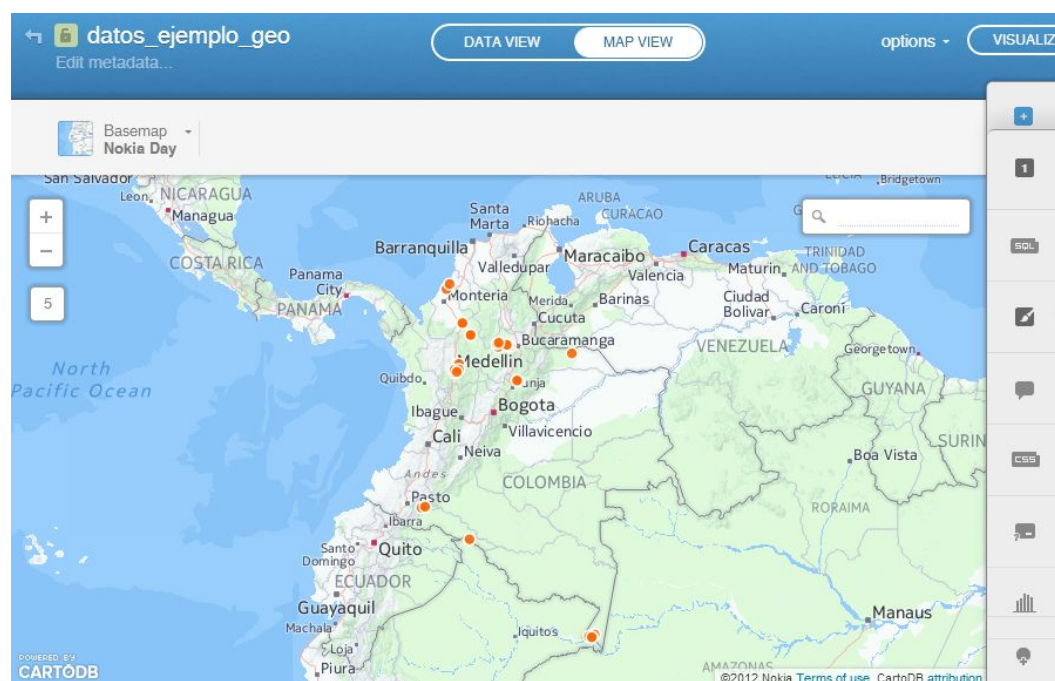



Figura 68. Visualización del mapa en CartoDB

Paso 7

En la parte superior izquierda puede identificar una opción para cambiar el mapa base. Por defecto está el mapa **"Nokia Day"**; Sin embargo, puede seleccionar el de su preferencia en cualquier momento.

Paso 8

Explore las opciones de la herramienta. Para esto, diríjase al menú gris de la parte derecha y haga clic en el número 1. Allí se despliega una serie de opciones, vaya al asistente de visualización  “Visualization wizard” en donde puede modificar de diversas maneras, la forma en que se ven los datos. Se presentan varias opciones para personalizar sus datos de la manera que desee:


- ❑ **Simple:** Tamaño, color y transparencia tanto del relleno (*Marker Fill*) como del borde (*Marker Stroke*). Se puede modificar la manera de composición de los puntos (*Composite operation*) y agregar también etiquetas a los puntos si se desea (*Label text*).
- ❑ **Cluster:** Permite agrupar sus datos dependiendo la escala de visibilidad y la cercanía de los puntos en el mapa. Útil en el caso de que se cuente con coordenadas repetidas o cercanas. Se puede seleccionar el rango de agrupación de los datos (*buckets*) e igualmente se pueden modificar las opciones de visualización, relleno, color, etc.
- ❑ **Choropleth:** Permite construir un mapa temático que representa la distribución espacial de un fenómeno mediante tramas o diferentes tonos de color o de gris. Se puede emplear con las columnas que presentan datos numéricos, y que se encuentran en ese formato. En su caso puede seleccionar la columna “*individualCount*” y modificar los rangos de agrupación, el método de cuantificación y las características de visualización.
- ❑ **Category:** Permite categorizar la visualización según la columna que se seleccione.
- ❑ **Bubble:** Representa la distribución espacial de la misma forma que lo hace *Choropleth*, pero empleando una variación en el tamaño de cada punto.

Encontrará otras opciones de visualización como Intensidad (*Intensity*) o densidad (*Density*), las cuales también puede explorar en detalle.

Paso 9

Para este ejercicio seleccione la opción de visualización **Bubble** y escoja la columna “número de individuos” para realizar la categorización visual según la cantidad de ejemplares. Cambie el color por defecto de los puntos (*Bubble fill*) y el tamaño de proporción (*Radius(min-max)*) a su gusto.

Paso 10

Ahora diríjase al símbolo de mensaje “infowindow” . Active en las ventanas de llamado las variables del conjunto de datos *Departamento*, *Familia* y *Género*; estas ventanas se activan al hacer clic sobre alguno de los puntos (si selecciona desde la pestaña “Click”) o al arrastrar el cursor sobre el punto (seleccionando desde la opción “Hover”). Explore y active los campos que desee.

Al abrir una de las ventanas de llamado usted observará (Figura 69) que tiene tres opciones en la parte inferior: puede (A) editar de la tabla alguno de los campos asociados al punto; (B) editar la geometría, le permitirá mover el punto a otro lugar del mapa y automáticamente modificará la coordenada; o (C) borrar por completo el dato de su tabla.

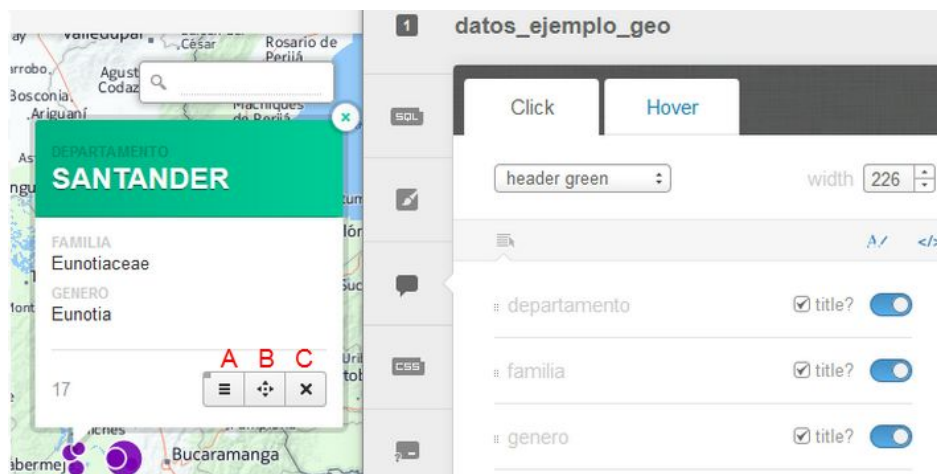



Figura 69. Visualización y opciones de las ventanas de llamado.

Paso 11

En el menú gris encontrará un ícono de barras  que corresponde al filtro de información. Filtre de sus datos la columna *Filo* de la lista desplegable; luego limpie la lista y seleccione **Mollusca**, si desea agregar más filtros dé clic sobre el símbolo (+); el filtro que seleccione se verá reflejado luego de unos segundos en el mapa.

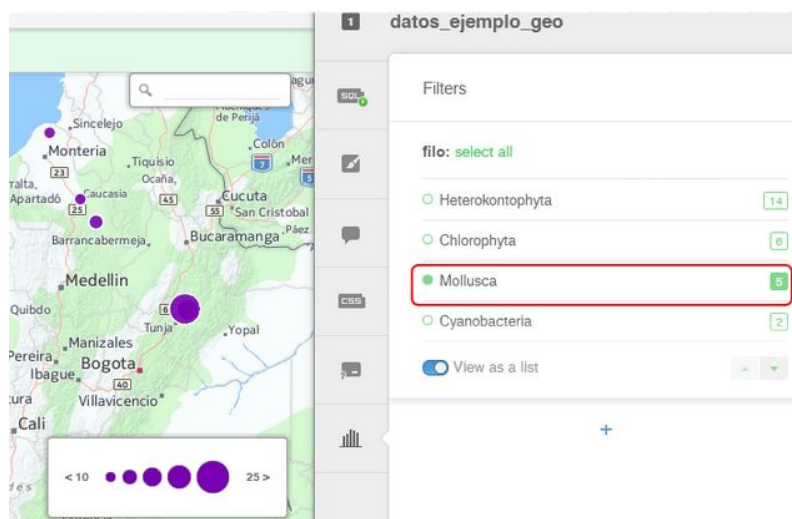


Figura 70. Visualización del filtro de información.

Paso 12

Cree una leyenda general al mapa haciendo clic sobre el ícono con el interrogante del menú derecho. Tenga en cuenta que la leyenda corresponda con los datos finales que haya filtrado. En este ejemplo: “Distribución y abundancia de Moluscos”.

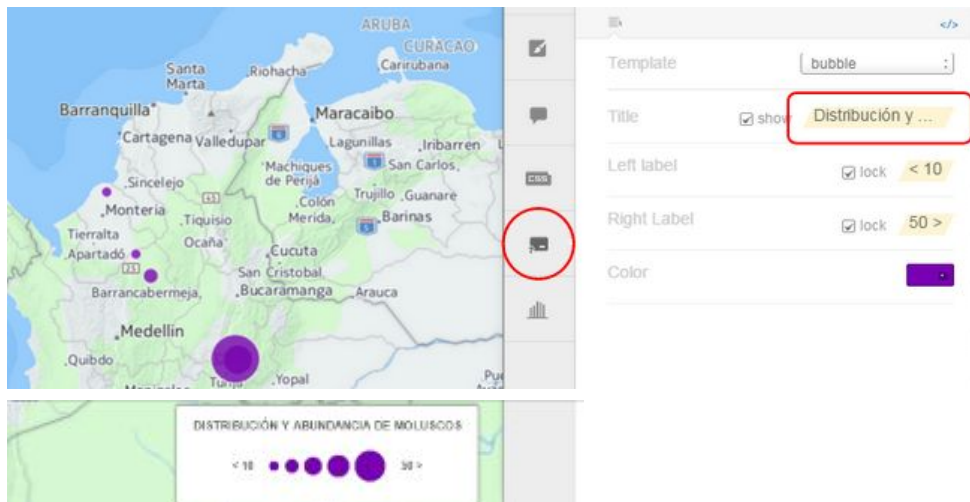


Figura 70. Creación de leyenda general para el mapa

Esta herramienta permite agregar capas de información para complementar la visualización y el trabajo con sus datos. En este caso agregaremos una capa en formato shape (.shp) que presenta los límites departamentales de Colombia (fuente: Instituto Geográfico Agustín Codazzi). Esto permitirá visualizar más claramente la validez de sus datos en cuanto al campo "Departamento".

Paso 13

Diríjase al símbolo azul (+) en la parte superior del menú gris. Haga clic en la opción "Data File" y luego en "select a file" para buscar el archivo "Deptos_Colombia.zip" ubicado en su equipo en la ruta **C:/Validaciones/Datos_Geo/Deptos_Colombia** y agregarlo a su visualización.

Nota: todos los archivos en formato shape deben haber sido comprimidos previamente en un archivo .ZIP con los archivos que lo componen (.dbf, .prj, .shx, .shp, etc.)

Paso 14

Se abrirá automáticamente luego de agregar el shape una ventana para salvar su visualización. Asigne un nombre a la visualización y haga clic sobre el botón verde para finalizar.

Nota: Aunque en este caso CartoDB despliega automáticamente esta opción al agregar una capa, usted puede para otros casos oprimir el botón "VISUALIZE", para crear la vista de su mapa en cualquier momento.

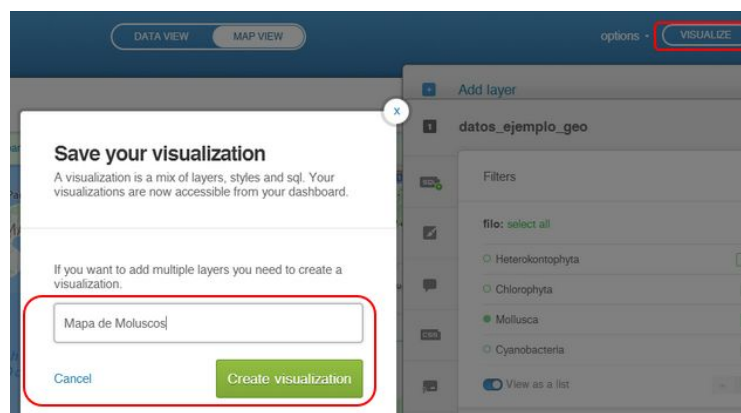


Figura 71. Guardado de la visualización del mapa en CartoDB.

Paso 15

Para ordenar las capas y que sus puntos se vean sobre la capa de Departamentos debe mover los números que encontrará en parte derecha arrastrando un número sobre el otro, ya que cada vez que ingresa una nueva capa esta va a quedar encima de las anteriormente agregadas. En este sentido, la capa más externa será la de mayor número (en este caso la número 2).

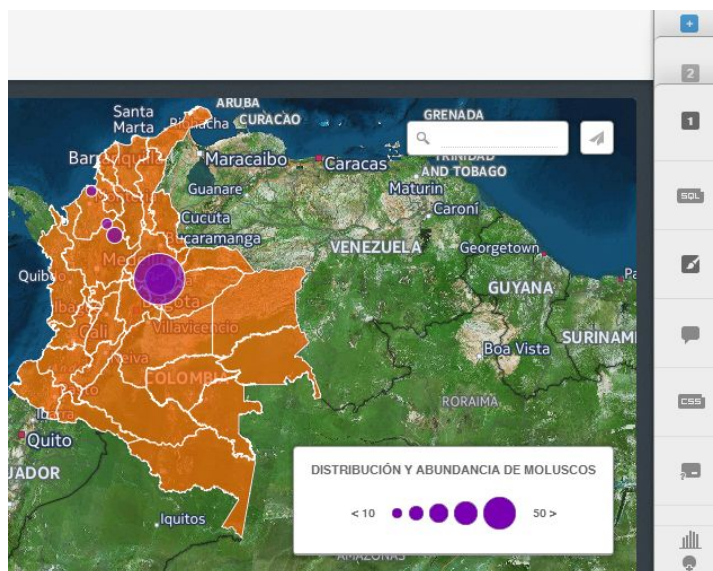


Figura 72. Organización y visualización de capas (menú a la derecha) en CartoDB.

Paso 16

Vaya a la opción Label text en el campo de visualización simple de la capa “deptos_colombia” y seleccione “departamen” para habilitar las etiquetas de cada Departamento y facilitar la ubicación de sus puntos dentro de la División político-Administrativa de Colombia. Ajuste igualmente el tamaño la fuente “Label font” de las etiquetas para que se ajuste al tamaño de visualización que desee.

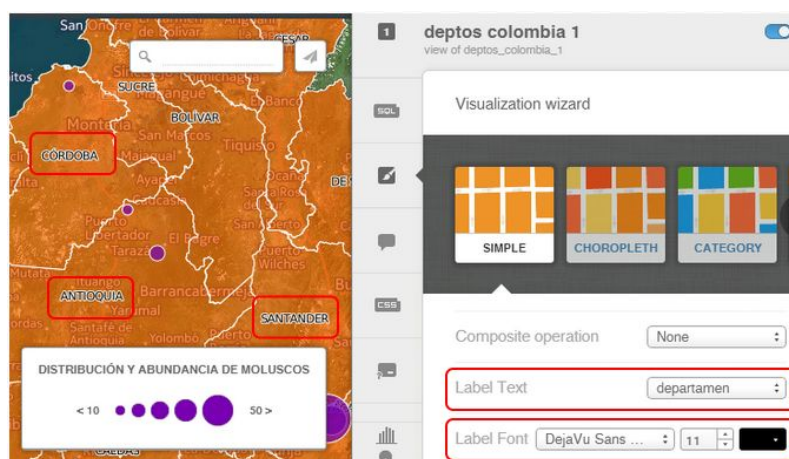


Figura 73. Configuración y visualización de etiquetas en CartoDB.

Paso 17

Configure los últimos detalles para finalmente compartir su mapa. Agregue un título o notas dentro del mapa y determine el mapa base que desea. Luego de esto, vaya a la parte superior derecha y seleccione “Share” para compartir su mapa.



Figura 74. Opciones finales para compartir y dar salida al mapa.

Nota: Usted puede compartir la visualización a través de una URL que genera la herramienta. Solo debe hacer clic en la opción *"Get the link"* y compartirlo a través de correo electrónico, redes sociales, etc. Igualmente si usted tiene un blog o *website*, puede embeber el mapa (*Embed it*) o agregarlo a una aplicación (*CartoDB.js*).

Paso 18

Pegue el enlace de su mapa en una pestaña nueva del explorador y observe el resultado de su trabajo. ¿Lo considera útil?

VeSPER

Vesper es una aplicación web que puede analizar el contenido de un archivo *Darwin Core Archive* utilizando varios criterios relacionados con la calidad. Proporciona resultados a través de gráficos y mapas que ayudan al usuario a entender mejor el contenido del archivo y donde pueden estar los problemas potenciales. El análisis se realiza en tres dimensiones fundamentales: taxonómica, geográfica y temporal, con una visualización dedicada a cada uno de estos aspectos.

Objetivo: Visualizar y analizar el contenido y distribución de un conjunto de datos en formato DwC, en términos geográficos, taxonómicos y temporales.

Enlace al programa:

<http://www.soc.napier.ac.uk/~cs22/vesperDemo/vesper/demoNew.html#small>

Caso Práctico

Usted es un biólogo al cual le han solicitado que investigue la cobertura taxonómica, geográfica y temporal de unos conjuntos de datos que se encuentran en formato *Darwin Core Archive*. Dentro de su investigación debe evaluar la calidad de los datos y qué tan adecuados pueden para el uso de un determinado usuario.

Paso 1

Descargue en el escritorio de su equipo un conjunto de datos del portal de GBIF. Para este ejercicio vamos a descargar un conjunto de datos llamado [Colección del Herbario PSO de la Universidad de Nariño](http://goo.gl/XPfjoo) disponible en: <http://goo.gl/XPfjoo>. Haga clic en el enlace [Darwin Core Archive](#) que está en la parte derecha de la página. Si lo desea puede descargarlo también del IPT del SiB Colombia: <http://ipt.sibcolombia.net/sib/resource.do?r=unar-001>.

Figura 74. Descarga de un archivo Darwin Core en GBIF

Nota 1: Lea rápidamente la descripción del conjunto de datos que aparece en la página de GBIF para que se haga una idea del conjunto de datos que va a analizar.

Nota 2: En este momento usted tiene en su escritorio un archivo tipo *Darwin Core Archive* (DwC-A) que contiene diferentes archivos de texto .txt y .xml con el conjunto de datos y los metadatos asociados.

Paso 2

Ingresa a la aplicación [VeSPER](#) y familiarícese con la herramienta. Identifique las pestañas de la siguiente imagen:

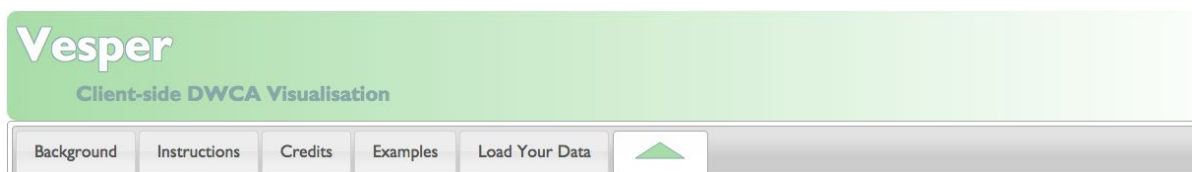


Figura 74. Vista del menú principal de Vesper. **Background:** Contienen información relacionada con los archivos DwC-A y una descripción de la herramienta. **Instructions:** Contiene las instrucciones básicas de como cargar, interactuar y visualizar la información. **Credits:** Contiene los detalles técnicos de la aplicación y los créditos para quienes la desarrollaron. **Examples:** Contiene cinco ejemplos con los cuales se puede explorar e interactuar. **Load Your Data:** Espacio para cargar un archivo DwC-A.

Paso 3

Ahora cargue el archivo DwC-A que descargó previamente en el punto 1 (dwca-unar-001.zip). Para esto, haga clic en la pestaña “**Load Your Data**” y luego en el botón “**Choose file**”.

Nota: En este paso el programa descomprime y consulta la parte meta.xml del DwC-A y establece si los campos necesarios están presentes para cada aspecto del análisis. De esta forma, ofrece las visualizaciones apropiadas como opciones antes de descomprimir los archivos de datos que quedan en el zip.

Paso 4

Luego de cargar el archivo, aparece una imagen con las secciones: **Label Records Using** y **Load Data For**.

En la sección **Label Records Using**, se escoge el tipo de etiqueta a utilizar en los registros, pueden aparecer opciones como: *acceptedNameUsage* (Nombre aceptado usado), *scientificName* (Nombre científico) o *vernacularName* (Nombre común), dependiendo de las columnas que tenga el archivo DwC.

En este ejercicio seleccione como etiqueta **scientificName**.

En la sección **Load Data For**, el investigador puede escoger si realiza el análisis en la parte taxonómica, geográfica y/o temporal.

En este ejercicio seleccione los tres aspectos **Denormalised Taxonomy**, **Map** y **Timeline**.

Figura 75. Opciones de carga de datos y configuración de etiquetas

A continuación se da una breve explicación de cada una de las opciones:

- ☐ **Denormalised Taxonomy:** Permite visualizar la taxonomía basada en los elementos del estándar DwC como Reino, Filo, Clase, Orden, Familia, entre otros.
- ☐ **Map:** Permite visualizar un mapa en forma de cluster o puntos de los datos proporcionados en los elementos de Latitud y Longitud Decimal.
- ☐ **Timeline:** Permite visualizar un gráfico de barras que agrupa las fechas en las cuales los especímenes fueron colectados. Esta sección tiene en cuenta los datos de la columna “Fecha del evento” del estándar DwC.

Para este ejercicio seleccione los tres aspectos mencionados y haga clic en el botón **Load**.

Nota: Para mayor claridad de los conceptos utilizados en este punto, puede colocar el cursor sobre las opciones de la aplicación y automáticamente se desplegarán las definiciones.

Paso 5

Luego de cargar los datos con las opciones escogidas aparece la siguiente pantalla con la visualización gráfica de los datos taxonómicos, espaciales y temporales.

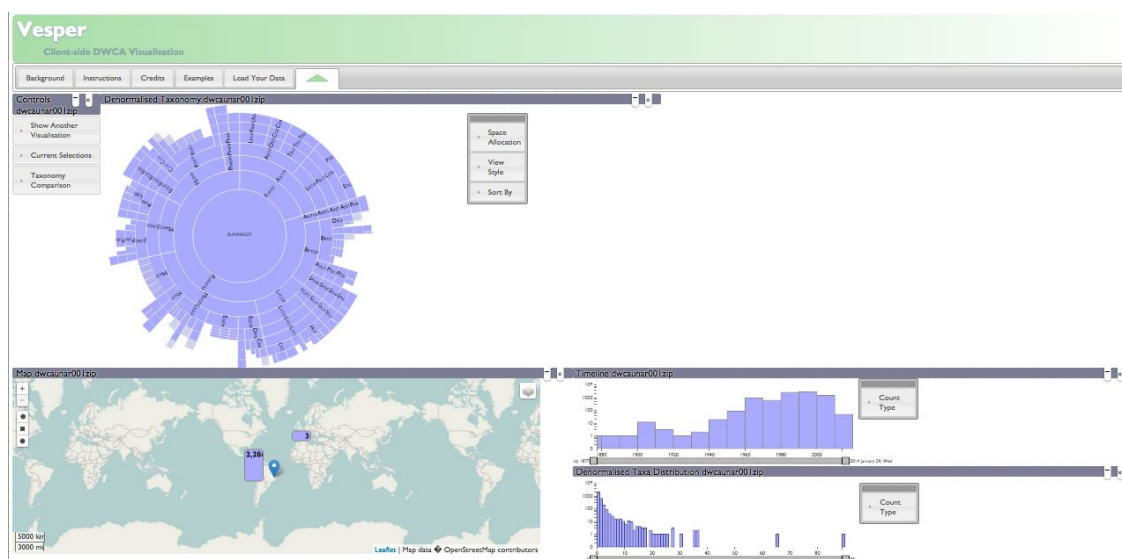


Figura 76. Captura de pantalla de los resultados de VeSPER. En la parte superior izquierda se encuentra la sección taxonómica, en la parte inferior izquierda la sección geográfica, en la parte inferior derecha arriba la sección temporal y finalmente, en la parte inferior derecha abajo la distribución taxonómica.

Visualización taxonómica

En la parte superior de la aplicación aparece la visualización taxonómica. Detalle la gráfica por niveles, desde el centro **SUPERROOT** hacia los extremos. A continuación encontrará algunas preguntas orientadoras para explorar la gráfica:

1. Coloque el cursor sobre el centro de la gráfica (**SUPERROOT**). Aparecerá un cuadro con información referente al conjunto de datos. ¿Cuántos especímenes tiene la Colección del Herbario PSO de la Universidad de Nariño?
2. Ahora desplace el cursor al primer nivel o anillo de adentro hacia afuera. ¿Cuántos y cuáles reinos visualiza? ¿Cuántos especímenes hay de cada uno?
3. Haga clic sobre **FUNGI**. ¿Cuántas clases de hongos están representadas en esta colección? Haga clic nuevamente sobre el centro de la gráfica para regresar a la raíz.
4. Ahora haga clic sobre **PLANTAE**. ¿Cuántos especímenes hay del phylum Magnoliophyta? ¿En cuántas clases se distribuyen?

Para cambiar la visualización de la distribución espacial (*Space Allocation*), estilo (*View Style*) y orden (*Sort By*) puede dirigirse a la siguiente ventana y escoger las opciones de acuerdo a sus necesidades.

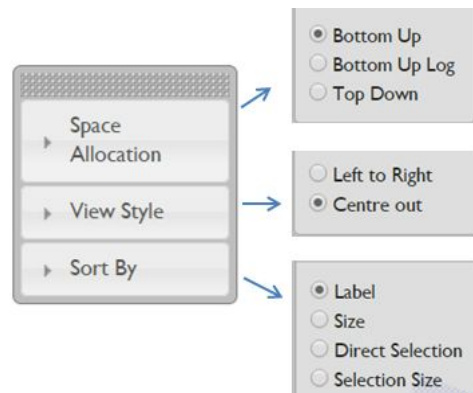


Figura 77. Opciones para cambiar la visualización de la gráfica "Denormalised Taxonomy"

Para este ejercicio resulta más informativo la distribución espacial en formato *Bottom up*, ya que muestra un mayor detalle de los registros desde la categoría taxonómica superior (Reino) hasta la inferior (Especie).

Visualización geográfica

La sección geográfica se encuentra en la parte inferior izquierda, el mapa tiene dos opciones de visualización: puntos agrupados (*Marker Group*) o puntos individuales (*Direct Plot*). Para escoger alguna de estas opciones o las dos en conjunto se despliega el botón que se encuentra en la parte superior derecha del mapa. El mapa también muestra puntos individuales azules que representan los puntos que no pudieron ser agrupados. A continuación encontrará algunas preguntas orientadoras para explorar el mapa.

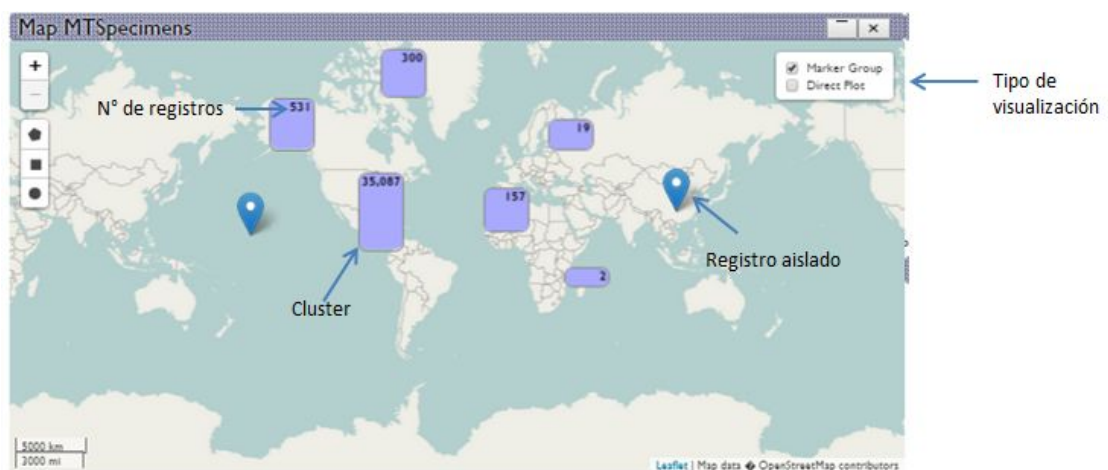


Figura 78. Visualización de la sección geográfica. En la parte superior derecha se encuentra el botón para escoger el tipo de visualización y en la parte superior izquierda se encuentran los botones + y - para establecer el zoom.

1. ¿De acuerdo al mapa hay registros aislados? Indique los IDs de aquellos especímenes y el país de procedencia.

Al ubicar el cursor en los puntos aislados se despliega una ventana con información acerca del espécimen, como el nombre científico, ID, latitud y longitud. De igual forma, al ubicar el cursor en alguno de los clusters, se puede observar el polígono agrupado y la ventana que se despliega proporciona información acerca del número de registros agrupados y la latitud y longitud del centro del cluster.

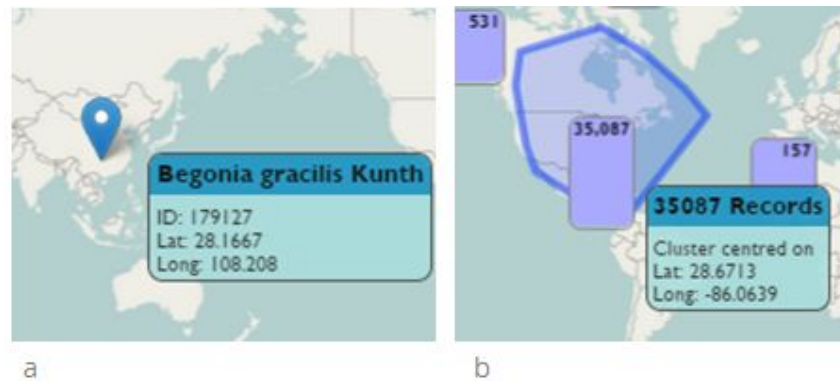


Figura 79. a. Acercamiento a uno de los puntos aislados con su información asociada. b. Acercamiento a uno de los clusters con su información asociada.

El mapa tiene opción de zoom en los botones de la parte superior izquierda o haciendo doble clic en cualquier parte del mapa, dependiendo del zoom se crean nuevos clusters según el grado de detalle que requiera el investigador. Por ejemplo, si se quiere tener más detalle acerca de los registros en Estados Unidos, haga doble clic en esa región. Adicionalmente, debajo de los botones del zoom se encuentran tres opciones para seleccionar registros (polígono, rectángulo y círculo).

2. ¿Cuántos registros georreferenciados tiene el conjunto de datos? ¿de estos registros cuantos se encuentran en el Chocó (Colombia)?

Nota: Chocó es un departamento de Colombia ubicado en la costa pacífica.

Visualización temporal

Continuando con la sección temporal, ésta se encuentra en la parte inferior derecha y muestra el rango de fechas que abarca el conjunto de datos. Es importante tener en cuenta que en la construcción del rango, el gráfico toma aproximadamente 10 años antes del primer registro y 10 años después del último registro, por ejemplo, si el registro más antiguo es de 1819 el primer rango empieza desde 1809.

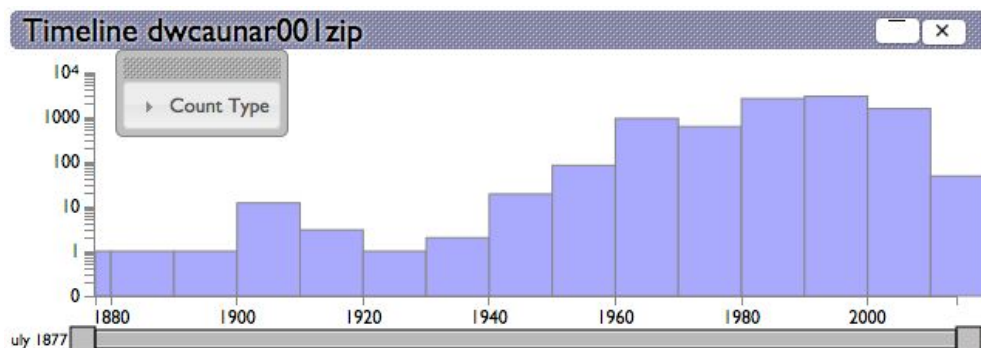


Figura 80. Visualización de la sección temporal. El gráfico de barras muestra el cambio en el número de registros documentados a lo largo del tiempo y cuenta con una barra en la parte inferior para escoger un rango menor de años de acuerdo al interés del investigador.

En la parte inferior se encuentra una barra que inicialmente muestra el rango completo de fechas. Si se requiere observar el comportamiento de los registros desde una fecha determinada, se mueve la barra de izquierda a derecha. Por ejemplo, al mover la barra para visualizar los registros a partir del 06 de junio de 1945 se obtiene la siguiente gráfica:

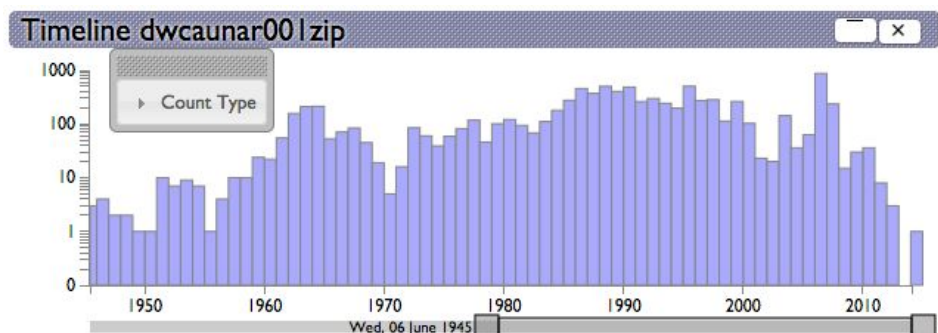


Figura 81. Visualización de la sección temporal al desplazar la barra a partir del 06 de junio de 1945.

Este tipo de gráfico permite analizar cuáles fueron los años de mayor actividad y determinar tendencias. Por ejemplo, en éste ejercicio se observa un incremento en el número de registros desde 1980 hasta 2000; luego se observa un leve descenso hasta la fecha.

El gráfico tiene la opción de desplegar cierta información al ubicar el cursor en alguna de las barras de tiempo. Por ejemplo, entre el 31 de diciembre de 1899 y el 31 de diciembre de 1909 hay 12 registros biológicos. Este tipo de información es útil en el momento de realizar análisis estadísticos. Por ejemplo, desde 1879-12-31 hasta 1929-12-31 hubo un incremento en el número de registros de 1220%.

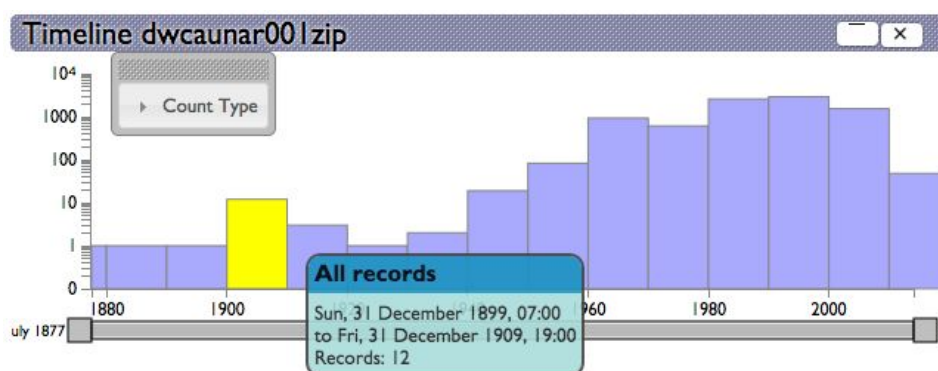


Figura 82. Información asociada a las barras temporales.

1. De acuerdo a la gráfica temporal ¿Cuáles son las fechas que reportan más especímenes de colecta? ¿Cuántos especímenes?
2. Al desplazar la barra de la gráfica temporal hasta el 5 de diciembre de 1945 ¿Cuál es el rango con mayor actividad y cuántos especímenes se registran?

Visualización taxonómica desnormalizada

Finalmente, la gráfica de distribución taxonómica desnormalizada, que se encuentra en la parte inferior derecha abajo, permite visualizar la taxonomía basada en los elementos del estándar Dwc como Reino, Filo, Clase, Orden, Familia, entre otros. A medida que nos movemos en el gráfico de izquierda a derecha, se encuentran los registros que tienen un mayor número de subtaxa asociados. Por ejemplo, al elegir en la gráfica una de las barra más lejanas, se observa que hay un registro que tiene 65 subtaxa asociados. Al hacer clic en la barra, cambia la gráfica de visualización taxonómica donde se observa a que registro(s) pertenecen la selección. En este caso, el género *Anthurium* tiene asociado 65 subtaxa (en este caso los subtaxa corresponden a especies).

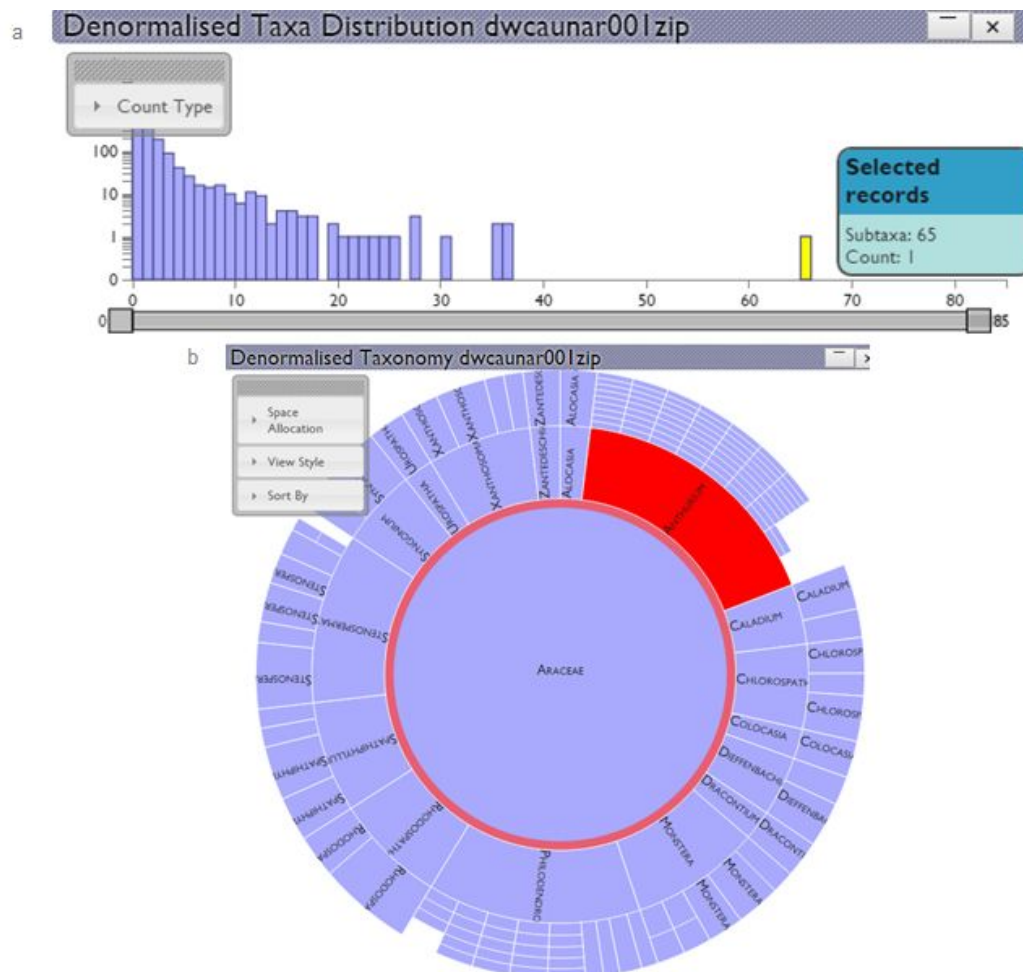


Figura 83. a. Gráfica de distribución taxonómica desnormalizada. Al elegir una de las barras más lejanas, ésta proporciona información acerca del número de registros (count) y el número de subtaxa asociados . **b.** Luego de dar clic en la barra, la gráfica de visualización taxonómica cambia mostrando el registro que tiene asociado los 65 subtaxa, en este caso, el género *Anthurium*.

1. ¿De acuerdo a la gráfica de distribución taxonómica desnormalizada, que registro tiene asociado el mayor número de subtaxa?

Respuestas

Visualización taxonómica

1. 8829
2. Dos; Plantae y Fungi; 8818 especímenes para plantae y 11 para fungi
3. Dos

4. 8416; 5

Visualización geográfica

1. PSO:42501 y PSO 42913; Alemania y Brasil
2. 2290; 11

Visualización temporal

1. 15 al 16 de enero de 1990
2. 31 de diciembre de 2005 al 01 de enero de 2007; 871

Visualización taxonómica desnormalizada

1. Familia Asteraceae con 89 subtaxa (en este caso 89 géneros)

Funciones adicionales de Vesper

Vesper cuenta con otras funciones adicionales a la visualización en los aspectos mencionados anteriormente, una de ellas es una sección que detalla el número de registros faltantes en varios campos utilizados para generar la visualización taxonómica, geográfica y temporal, de esta forma, se puede tener una idea general de la calidad de los gráficos, de acuerdo a la cantidad de registros que se utilizaron para generarlos vs los registros totales que posee el conjunto de datos.

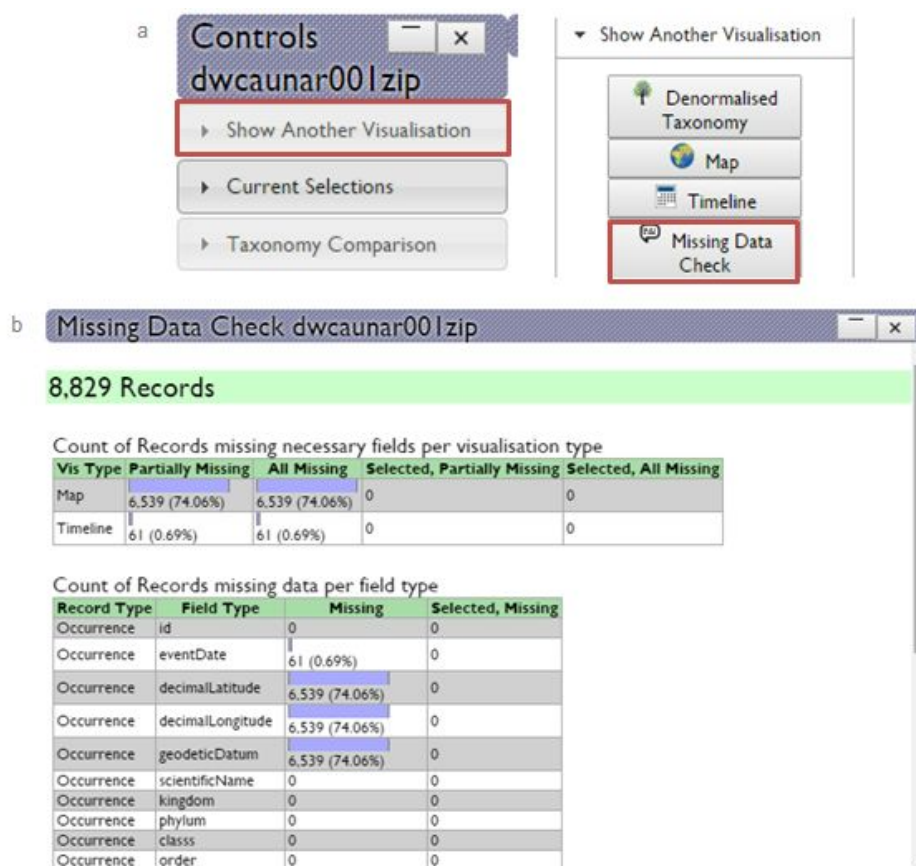


Figura 84. Chequeo de datos faltantes **a.** Controles del conjunto de datos donde se despliega la opción "Show another visualisation" y luego se da click en el botón "Missing Data Check". **b.** Tabla de resultados. En la parte superior de la tabla se encuentra el conteo de los registros faltantes en los campos necesarios para cada tipo de visualización y en la parte inferior se encuentra el conteo de registros faltantes más detallado.

En este ejemplo, el conjunto de datos presenta vacíos en la parte geográfica (6.539 registros, equivalentes al 74.06% del total) y en la parte temporal (61 registros, equivalentes al 0,69% del total)

Enlaces de interés: A través de los siguientes tres recursos web usted podrá ampliar la información sobre VeSPER.

- ❑ [VeSPER, Client-side DwC-A Visualisation](#)
- ❑ [Vesper: Visualising species archives](#)
- ❑ [Group discussion: New tool to analyze DwC-A: VeSPER](#)

Darwin Test

Es una aplicación diseñada para la validación y chequeo de los datos presentados en tablas en el formato estándar *Darwin Core* versión 1.2 y 1.4, empleado por GBIF para el intercambio de información de observaciones y especímenes conservados en colecciones de historia natural.

Se recomienda el uso de esta herramienta para la validación de los datos del formato *DwC*, antes de subir los datos al servidor para su visualización en la red de GBIF (<http://www.gbif.es/>). DARWIN TEST está desarrollado y apoyado por la Unidad de Coordinación del Nodo Nacional de GBIF en España.

Permite detectar errores en los campos mediante consultas predeterminadas configurables por el usuario. Ofrece la posibilidad de corregir los errores detectados en los datos desde los propios formularios de validación. Permite la activación/desactivación de las consultas existentes. Igualmente la creación de nuevas consultas de validación adaptadas a las necesidades de cada usuario. Realiza el chequeo de los nombres científicos con la Tabla de Nombres SCIENTIFIC_NAMES del consorcio The Catalogue Of Life: JTIS y Species 2000, con la tabla Archivos de Autoridad Taxonómica (AAT) del SIB de Colombia, o con otra base de datos con nombres científicos que cumpla unos requisitos mínimos. Dichas tablas de referencia están disponibles en la página de descarga de Darwin Test; entre muchas otras características y funcionalidades. Puede encontrar más información en la [página](#) de *GBIF España* y el código fuente queda accesible desde la página del proyecto en Sourceforge.net bajo licencia *Creative commons* [CC-BY-NC-SA 3.0](#).

Instalación de Darwin Test

El objetivo de esta práctica es instalar y configurar la aplicación. Durante el taller se trabajará con la aplicación proporcionada por los instructores preparada para funcionar con Office de 64 bits.

Paso 1


Copiar el archivo **Darwin_Test33-fuente-64** y las bases de datos de pruebas proporcionado por los instructores.

Paso 2

Ejecutar la aplicación haciendo doble click sobre el archivo **Darwin_Test33-fuente-64** que dará paso a un menú de inicio (Figura 85).

Paso 3

Desde la pestaña *Configuración y herramientas* - *Configurar validación de datos*, configure la ruta y nombre de la base de datos de Species2000.

Para ello pulse sobre el icono  que le permitirá navegar por los archivos de su equipo y localizar el

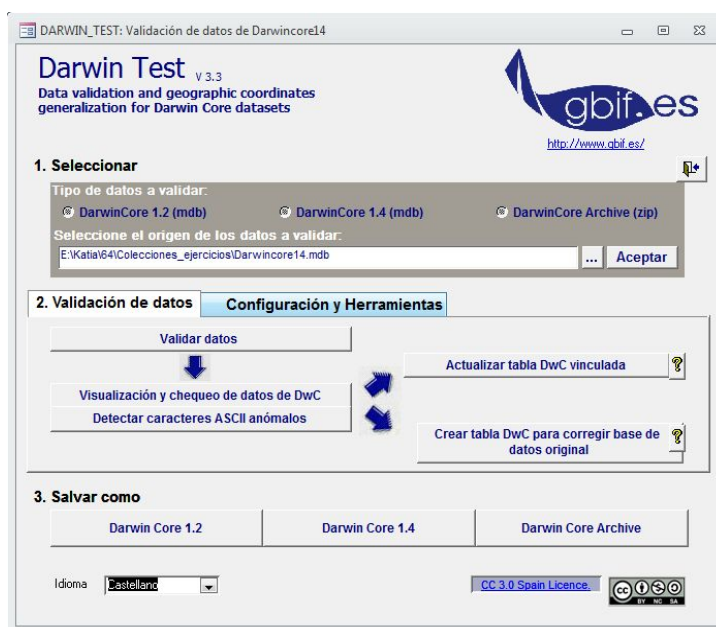


Figura 85. Menú inicial de Darwin Test

archivo con el nombre *Species2000_Scientific_names_2010* (incluido entre los materiales descargados) y una vez localizado pulse el botón *Vincular tabla ScientificNames etc.* . Mediante este proceso habremos vinculado la tabla con nombres de referencia que se utilizarán para posteriores chequeos.

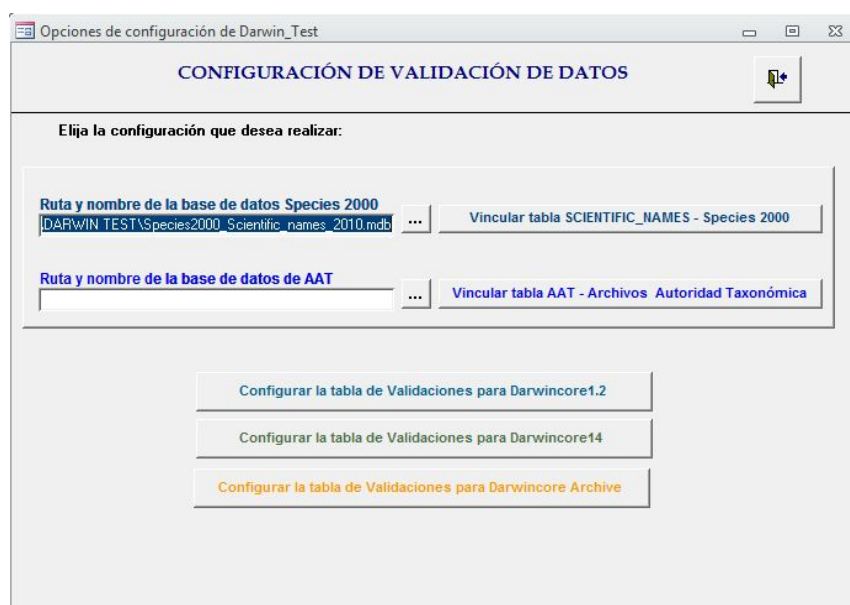


Figura 86. Configuración de Validación de datos en Darwin Test

Caso Práctico

Los casos prácticos están relacionados entre sí, es muy recomendable que se realicen de forma consecutiva y siguiendo la estructura establecida. Para realizar estos ejercicios usted debe descargar la aplicación *Darwin Test* y los materiales para los ejercicios disponibles de la carpeta **Casos prácticos Darwin Testin** en el siguiente vínculo: <http://tinyurl.com/practicaDQ>

CASO 1. VALIDACIÓN DE DATOS

El objetivo de esta práctica es utilizar Darwin Test para **comprobar la calidad de los datos** de la colección Darwin Core 1.4 de ejemplo (*Darwincore14.mdb*).

Paso 1

Desde el menú principal en la sección encabezada con el título *1. Seleccionar* escoja el formato de la tabla a validar Darwin Core 1.4, y busque en su equipo el archivo *Darwincore14.mdb*. A continuación pulse el botón *Aceptar* para que la tabla quede vinculada a la aplicación.



Figura 87. Selección del formato de tabla DwC a validar en Darwin Test

Paso 2

En el siguiente paso vamos a validar los datos pulsando sobre el botón *Validar datos* que se encuentra en la pestaña *Validación de datos*. En el momento de inicio se abrirán varias ventanas preguntando si desea realizar la validación siguiendo la configuración establecida (responda sí) si desea calcular el ICA (responda NO) y advertirá de que se están eliminando tablas de validaciones anteriores (Responda Sí). Diga a todo que sí para que continúe con el proceso.

Paso 3

Antes de finalizar, una nueva ventana nos preguntará si deseamos eliminar los caracteres ASCII anómalos, a lo que también responderemos que sí.

En este momento la aplicación habrá realizado todas las validaciones y habrá generado unas tablas con los resultados. En el siguiente caso veremos cómo consultar esos “posibles errores” y cómo corregirlos.

CASO 2. CORRECCIÓN DE CAMPOS VALIDADOS

Paso 1

El objetivo de este caso no es subsanar todos los errores, sino conocer el funcionamiento básico de la herramienta, así que para realizar este ejercicio sólo será necesario corregir los fallos detectados en el campo GENUS. Para ello, pulse en el botón *Visualización y chequeo de datos Darwin Core* que abrirá la siguiente pantalla:

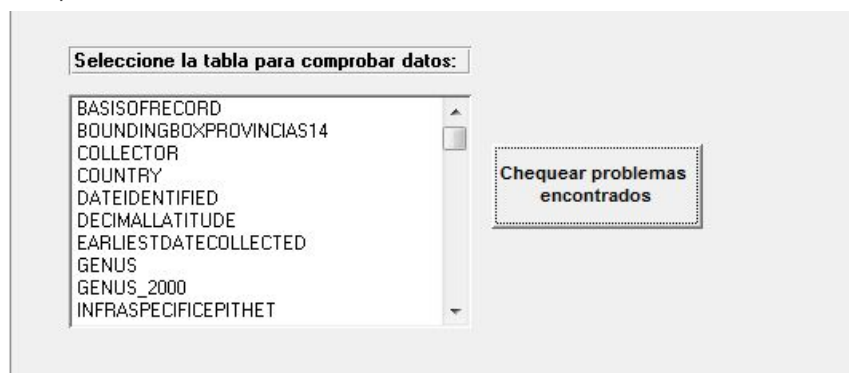


Figura 88. Selección de la tabla para comprobar datos en Darwin Test

Paso 2

Corrija los 4 casos detectados y para cada uno pulse el botón *Para validar*.

Paso 3

A continuación cierre estos formularios y desde el menú principal pulse *Actualizar tabla Darwin core Archive validada* que se encuentra en el menú principal. Mediante este proceso, todas las correcciones realizadas se guardarán en la tabla darwin Core 1.4.

CASO 3. GENERALIZACIÓN DE COORDENADAS

El objetivo de este ejercicio es generalizar a distintos grados las coordenadas de especies protegidas estableciendo los filtros necesarios.

Paso 1

Vincule la nueva tabla de origen Darwin Core 1.2 siguiendo las instrucciones del caso 1.

Paso 2

Desde la sección *Configuración y herramientas -Generalizar coordenadas* establezca las condiciones de consulta para que la aplicación busque todos los registros con sensibilidad *Extrema*, para ello haga como se muestra en la Figura 89.

Figura 89. Configuración de la generalización de coordenadas en Darwin Test

A continuación pulse *Seleccionar registros* y después *Ver como consulta* para comprobar los resultados.

DateLastModified	Institución	Colección	ScientificName	
2012-08-31T13:39:11,0Z	GBIF	GBIF	Peucedanum schottii Besser ex DC.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Peucedanum schottii Besser ex DC.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Peucedanum schottii Besser ex DC.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Peucedanum schottii Besser ex DC.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Peucedanum schottii Besser ex DC.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Peucedanum schottii Besser ex DC.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Solanum lidii Sunding	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Veronica chamaepithyoides Lam.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Veronica chamaepithyoides Lam.	Especie en peligro crítico (CR); Sensibilidad Extrema
2012-08-31T13:39:11,0Z	GBIF	GBIF	Veronica chamaepithyoides Lam.	Especie en peligro crítico (CR); Sensibilidad Extrema

Figura 90. Vista de consulta para comprobación de resultados en Darwin Test

Paso 3

A estos registros seleccionados aplíquese una eliminación de coordenadas, desde el botón *Eliminar coordenadas y radiopunto*.

Paso 4

Limpie las condiciones de la consulta que acaba de hacer pulsando el botón *Nuevo filtro* y aplique la generalización de coordenadas que corresponde a cada caso de sensibilidad expresado en la Tabla 4. En cada caso tendrá que realizar una consulta para buscar qué registros se encuentran bajo cada categoría de sensibilidad y aplicar la generalización de coordenadas que corresponda especificada en el campo Georreferenciación de la tabla.

Tabla 4. Generalización de coordenadas correspondiente a cada caso de sensibilidad.

Categoría	Sensibilidad	Especie	Georreferenciación	Nº de registros
1	Extrema	<i>Apium bermejoi</i>	Eliminar coordenadas	3
2	Alta	<i>Atropa baetica</i>	Generalizar a 0.1 grado	4
3	Media	<i>Phyllis viscosa</i>	Generalizar a 0.01 grado	9
4	Baja	<i>Sorbus hybrida</i>	Generalizar a 0.001 grado	5
No sensible	No sensible	Resto de especies	Georreferenciación sin restricción	1465

Paso 5

Compruebe en el campo Notes de la tabla Darwin Core 1.2 que se han cumplimentado los metadatos que informan sobre la generalización.

CASO 4. EXPORTACIÓN A DARWIN CORE ARCHIVE

En esta práctica el alumno debe exportar la tabla de trabajo al formato Darwin Core Archive. Para poder generar este formato es necesario rellenar o importar los metadatos (*eml*) que formarán parte de dicho paquete. Para este ejercicio se pueden importar los metadatos de ejemplo *eml.xml* incluidos en la carpeta de material de trabajo.

Paso 1

Abra configuración de *Herramientas - Metadatos de la colección* que dará paso al siguiente formulario, en el que debe seleccionar la opción *Importar desde Archivo eml* y a continuación buscar en su equipo el archivo con el nombre *eml.eml* que se encuentra entre el material para prácticas compartido.

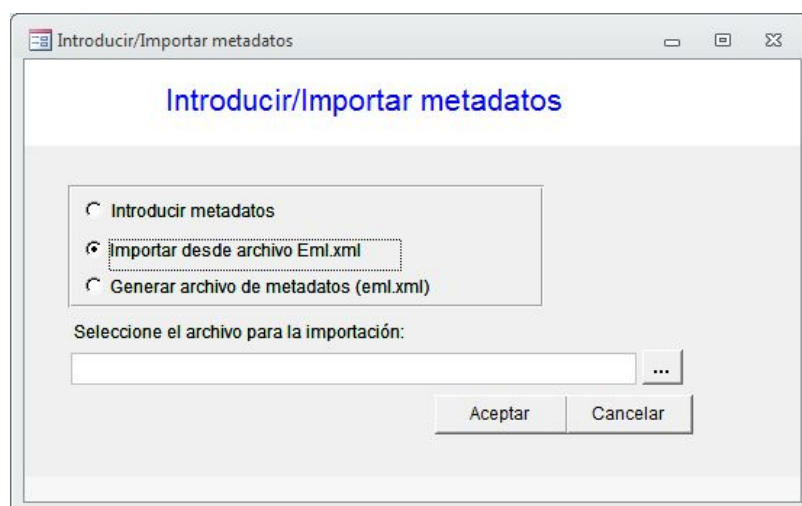


Figura 91. Ingreso o importación de metadatos en Darwin Test.

Paso 2

Pulse aceptar y compruebe que se han importado los metadatos llegando de nuevo a este formulario y seleccionando en esta ocasión la opción *Introducir metadatos*.

Paso 3

Una vez exportado, comprobar que el paquete se ha formado correctamente desde el Validador online de Darwin Core Archive: <http://tools.gbif.org/dwca-validator/> .

CASO 5. EXPORTACIÓN DE DARWIN CORE ARCHIVE A DARWIN CORE 1.2

OBJETIVO. Exportar al formato Darwin Core 1.2. el archivo en formato Darwin Core Archive.

Paso 1

Vincular el archivo dwca-guayacanal-parcelasmartos.zip en formato archive dentro de la aplicación.

Paso 2

Exportar al formato Darwin Core 1.2 desde *Salvar como - Darwin Core 1.2.*

ÍNDICE

[1. INSTALACIÓN](#)

[EN WINDOWS](#)

[EN MAC OS](#)

[2. CONVENCIONES DE LOS EJERCICIOS](#)

[3. USO BÁSICO](#)

[CARGA DE DATOS](#)

[FACETING](#)

[FILTERING](#)

[CLUSTERING](#)

[EXPORTACIÓN](#)

[4. USO AVANZADO](#)

[USO DEL API DE EOL](#)

[5. ENLACES ADICIONALES](#)

Open Refine

OpenRefine es una aplicación de escritorio de código abierto independiente para la limpieza y transformación de datos a otros formatos. Es similar a las aplicaciones de hojas de cálculo (como Excel). Sin embargo, se comporta más como una base de datos.

Opera en filas de datos que tienen celdas bajo columnas, muy similar a las tablas de bases de datos relacionales. El usuario puede filtrar las filas para mostrar el uso de “facets” que definen criterios de filtrado. A diferencia de las hojas de cálculo, la mayoría de las operaciones en OpenRefine se realizan en todas las filas visibles. Permite la creación de nuevas columnas basadas en los datos de las columnas existentes, etc. Todas las acciones realizadas en un conjunto de datos son almacenada en un proyecto y se pueden reproducir en otro conjunto de datos.

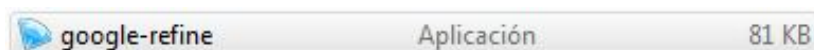
A diferencia de las hojas de cálculo, no hay fórmulas almacenadas en las celdas. Las fórmulas que se ingresan por columna se utilizan para transformar los datos y la transformación se realiza sólo una vez. Las expresiones de transformación se pueden escribir en Google Refine Expression Language (GREL), Jython (lenguaje Python) y Clojure.

INSTALACIÓN

Diríjase a la sección de descargas del sitio web de *OpenRefine* <http://openrefine.org> (anteriormente conocido como *GoogleRefine*) y seleccione su sistema operativo. Se recomienda observar los tres videos de introducción a la herramienta y no utilizar Internet Explorer.

EN WINDOWS

- Un archivo ZIP se descargará en su computador
- Extraiga el archivo a una carpeta nueva y nómbrela **Google_Refine**
- Ingrese en la carpeta y ejecute la aplicación:



Se abrirá automáticamente una ventana de línea de comandos y luego la interfaz del programa en el navegador de Internet.

EN MAC OS

- Un archivo DMG se descargará en su computador
- Abra el archivo y arrastre el icono del programa a la carpeta “Aplicaciones” (Applications):



- Ejecute la aplicación y se abrirá el programa en el navegador de Internet.

CONVENCIONES DE LOS EJERCICIOS

Fórmulas a usar en la herramienta (copiar y pegar).

cell.recon.match.id

Comandos en OpenRefine, rutas.

Edit column

Nombres de las columnas.

nombreRecon

Enlaces a sitios informativos.

www.sibcolombia.net

Menú Columna



USO BÁSICO

CARGA DE DATOS

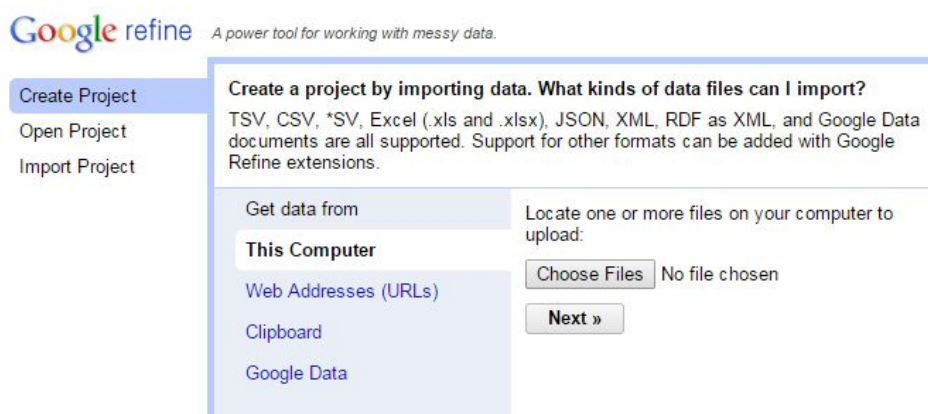
ANTES DE EMPEZAR

La carga de datos se puede hacer desde diversas fuentes de datos: TSV, CSV, SV, Excel (.xls y .xlsx), JSON, XML, RDF as XML y datos de Google Docs. La carga de datos implica dos etapas, la primera es la creación del proyecto y la segunda es el análisis de la fuente.

CREAR UN PROYECTO

Diríjase a [Guías Taller](#) en donde encontrará la carpeta **Archivos Ejemplo**.

1. Descargue el archivo **Datos_OR**
2. En *OpenRefine* la pestaña **Create Project** estará habilitada por defecto. Para cargar el archivo siga la ruta **Get data from > This Computer**, y haga clic en **Choose Files**:



3. Seleccione el archivo **Datos_OR** y haga clic en **Next**.
4. Un panel de selección aparecerá, este le permite especificar el tipo de datos que se cargan y configurar la manera en la que los datos son leídos. Deje las opciones tal y como se muestran en la imagen:

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

RDF/IN3 files

XML files

Open Document Format spreadsheets (.ods)

RDF/XML files

Character encoding **UTF-8**

Columns are separated by

☐ commas (CSV)

☒ tabs (TSV)

☐ custom \t

Escape special characters with \

Update Preview

☐ Ignore first 0 line(s) at beginning of file

☒ Parse next 1 line(s) as column headers

☐ Discard initial 0 row(s) of data

☐ Load at most 0 row(s) of data

☒ Parse cell text into numbers, dates, ...

☒ Quotation marks are used to enclose cells containing column separators

☒ Store blank rows

☒ Store blank cells as nulls

☐ Store file source (file names, URLs) in each row

- En la esquina superior derecha verá un cuadro de texto en el que puede cambiar el nombre del proyecto, haga clic en el botón **Create Project**:

Project name **Create Project »**


- Espere a que cargue el archivo, esto puede tomar un tiempo dependiendo del tamaño del mismo.

FACETING

ANTES DE EMPEZAR

Es un método para filtrar los datos en conjuntos más pequeños para facilitar el uso y análisis, puede hacerse para el texto, los números y las fechas.

EJERCICIO 1. **Faceting y correcciones masivas**

- Diríjase a la columna **Clase**, haga clic en el Menú Columna  y siga la ruta que se muestra en la imagen para hacer un **Text Facet**:

▼ **Clase**

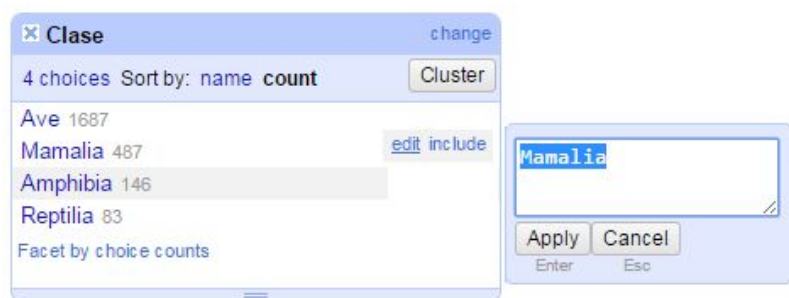
Facet	▶	Text facet
Text filter		Numeric facet
Edit cells	▶	Timeline facet
Edit column	▶	Scatterplot facet
Transpose	▶	Custom text facet...
Sort...		Custom numeric facet...
View	▶	Customized facets ▶
Reconcile	▶	

- A su izquierda aparecerá una ventana con el nombre de la columna y el Facet que se realizó:




Haga clic en **Count** para organizar las clases de la más a la menos abundante y en **Name** para organizarlas en orden alfabético.

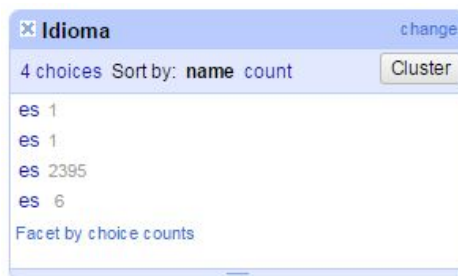
- Corrija las inconsistencias en los nombres de las Clases Aves y Mammalia. Para esto acerque el cursor al valor que desea corregir y haga clic en **Edit**, luego en el cuadro de texto que aparece corrija el error y haga clic en **Apply**:



Todos los valores serán corregidos de manera automática.

EJERCICIO 2. Faceting y espacios en blanco

- Diríjase a la columna **Idioma**, haga clic en el Menú Columna  y realice un **Text Facet**
- A su izquierda aparecerá la ventana con el nombre de la columna y el Facet que se realizó:




Aunque a simple vista los datos se encuentran sin errores, al realizar este procedimiento vemos que el programa ha detectado espacios extra y por eso nos muestra cuatro opciones diferentes.

- Corrija las inconsistencias desde el Menú Columna de **Idioma**, siguiendo la ruta **Edit Cells > Common transforms > Trim leading and trailing whitespace**, verá un mensaje de notificación:

Text transform on 8 cells in column Idioma: value.trim() Undo

2. Si observa la ventana del Facet de **Idioma**, notará que ahora solo existe una opción y que los espacios fueron eliminados.

EJERCICIO 3. Faceting y duplicados

1. Diríjase a la columna **Número de Catálogo**, haga clic en el Menú Columna  y siga la ruta **Facet > Customized facets > Duplicates facet**.
3. A su izquierda verá la ventana del Facet:




catalogNumber	
2 choices Sort by: name count	
false	2401
true	2

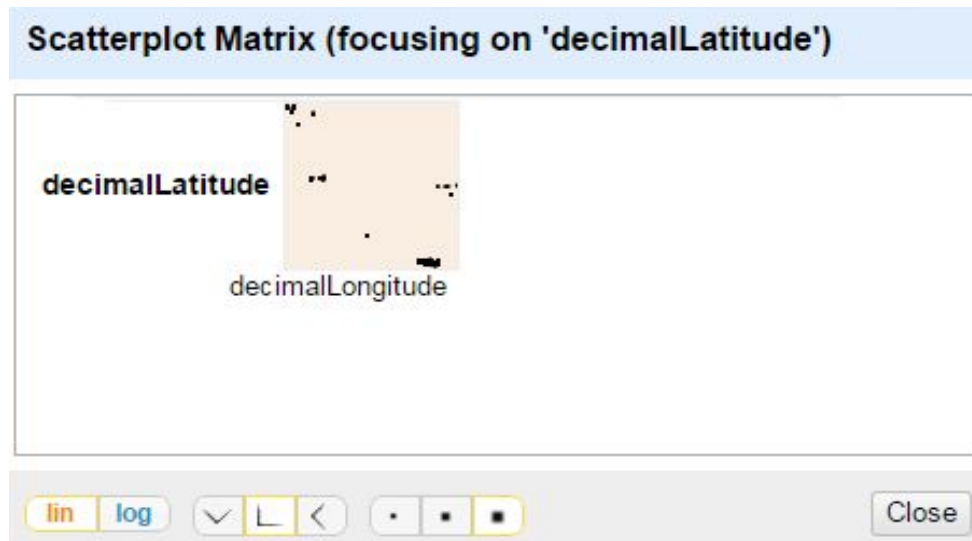
Facet by choice counts

Podemos ver que el programa ha detectado valores únicos (false) y valores duplicados (true).

4. Haga clic en **true** y verá los registros. De esta manera se pueden detectar los duplicados para un análisis posterior.

EJERCICIO 3. Scatterplot facet

2. Diríjase a la columna **Latitud Decimal**, haga clic en el Menú Columna  y siga la ruta **Facet > Scatterplot facet**.
4. Aparecerá una ventana con las posibles relaciones numéricas de la columna seleccionada y las demás columnas que contienen valores numéricos. Seleccionamos las opciones de los recuadros amarillos de la imagen y luego hacemos clic en la gráfica que se muestra: *Latitud Decimal (x) vs. Longitud decimal (y)*. De esta manera aparecerá la ventana del Facet en nuestro espacio de trabajo.



- Diríjase a la columna **Municipio**, haga clic en  y realice un **Text Facet**, de esta manera OpenRefine debe mostrar:



- Usted podría detectar coordenadas fuera de lugar en una gran cantidad de registros para realizar un análisis posterior. Haga clic en cada uno de los municipios para ver como estos se agrupan, de la misma manera puede seleccionar los grupos formados directamente en la gráfica. Al realizar esto se irán visualizando los registros que seleccione en la tabla de datos. Si desea ver la grafica mas grande aumente el zoom del navegador o haga clic en **export plot**.
- Selecione el municipio de Salaminas y observe como forma dos grupos en la gráfica.
- Selecione directamente en la gráfica cada grupo y observe como uno de los grupos presenta coordenadas completamente erróneas.

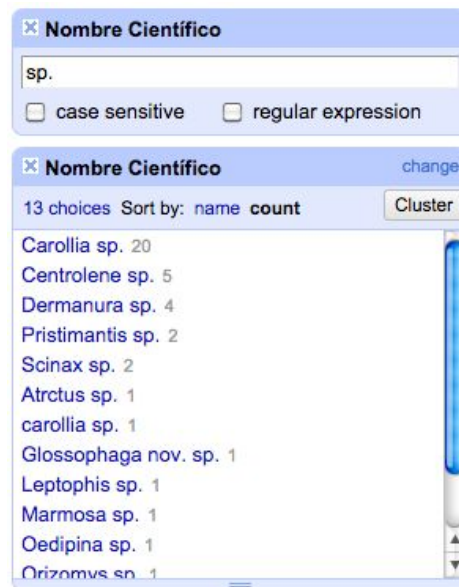
9. Para volver a ver todos los registros haga clic en **reset** en las dos ventanas,

FILTERING

EJERCICIO 4. Filtro básico

1. Diríjase a la columna **Nombre Científico**, haga clic en el Menú Columna y luego en **Text filter**, aparecerá la ventana del Filtro.

Escriba en el campo de texto **sp.** y realice un **Text Facet** en **Nombre Científico** para visualizar los registros con este valor:



Para corregir este error vaya nuevamente al Menú Columna de **Nombre Científico** y siga la ruta **Edit cells > Transform...**, luego ingrese la fórmula `value.replace(" sp.", "")` tal y como se muestra a continuación:

Custom text transform on column Nombre Científico

Expression Language Google Refine Expression Language (GREL)

`value.replace(" sp.", "")` No syntax error.

Preview [History](#) [Starred](#) [Help](#)

row	value	value.replace(" sp.", "")
11.	Centrolene sp.	Centrolene
12.	Centrolene sp.	Centrolene
13.	Centrolene sp.	Centrolene
14.	Centrolene sp.	Centrolene
15.	Centrolene sp.	Centrolene
94.	Oedipina sp.	Oedipina
99.	Pristimantis sp.	Pristimantis

On error ☒ keep original ☐ set to blank ☐ store error ☐ Re-transform up to times until no change

Haga clic en **OK** y verá el mensaje de confirmación de que los cambios se han realizado.

EJERCICIO 5. Filtro avanzado I

1. Diríjase a la columna **familia**, haga clic en el Menú Columna y luego en **Text filter**, aparecerá la ventana del Filtro.
2. Marque la casilla **regular expression**. Escriba en el campo de texto la expresión regular `.*(?:?!ae).$` esta expresión nos permite excluir todas las palabras de la columna que no terminan en "ae".
3. Realice un **Text Facet** en **familia** para visualizar los registros con este elemento:

family

`.*(?:?!ae).$`

☐ case sensitive ☒ regular expression

family [change](#)

2 choices Sort by: name count

Anura 1

Bolitoglossa 1

Facet by choice counts

Podrá observar como los registros que no corresponden a familias han sido filtrados, usted puede editarlos tal cual como en el punto 3 de Ejercicio 1. Para conocer más de las expresiones regulares haga clic [aquí](#).

EJERCICIO 6. Filtro avanzado II

- Diríjase a la columna **Nombre Científico**, haga clic en el Menú Columna y luego en **Text filter**, aparecerá la ventana del Filtro.
- Marque la casilla **regular expression**. Escriba en el campo de texto la expresión regular `[.]` y realice un **Text Facet** para visualizar los registros con este elemento:

Nombre Científico

[.]

☐ case sensitive ☒ regular expression

Nombre Científico [change](#)

value

4 choices Sort by: [name](#) [count](#) [Cluster](#)

Carollia cf. sowelli 39

Glossophaga cf. 1

Microsciurus cf. pucheranii 1

Proechimys cf. trinitatus 15

[Facet by choice counts](#)

Podrá observar los registros que cumplen este criterio, para editarlos masivamente haga clic en **change** en la ventana del Facet de **Nombre Científico**.

- En la ventana que aparece escriba la fórmula `value.replace(" cf.", "")` y haga clic en **OK**:

Edit Facet's Expression based on Column Nombre Científico

Expression Language [Google Refine Expression Language \(GREL\)](#)

value.replace(" cf.", "") No syntax error.

Preview [History](#) [Starred](#) [Help](#)

row	value	value.replace(" cf.", "")
1895.	Carollia cf. sowelli	Carollia sowelli
1896.	Carollia cf. sowelli	Carollia sowelli
1897.	Carollia cf. sowelli	Carollia sowelli
1898.	Carollia cf. sowelli	Carollia sowelli
1899.	Carollia cf. sowelli	Carollia sowelli
1900.	Carollia cf. sowelli	Carollia sowelli
1901.	Carollia cf. sowelli	Carollia sowelli

[OK](#) [Cancel](#)

Todos los registros se corregirán de manera automática.

CLUSTERING

EJERCICIO 7. Clustering básico

1. Diríjase a la columna **Colector**, haga clic en el Menú Columna y luego en **Text facet**, aparecerá la ventana del Facet:



2. En la parte superior derecha verá el botón **Cluster** haga clic, aparecerá la ventana de **Cluster & Edit** para la columna **Colector**.

Cluster & Edit column "identificationQualifier"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: **key collision** Keying Function: **fingerprint** 2 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	42	<ul style="list-style-type: none">ricardo g (40 rows)G. Ricardo (1 rows)Ricardo G. (1 rows)	<input type="checkbox"/>	ricardo g
2	2	<ul style="list-style-type: none">Ana gonzales (1 rows)Ana Gonzales (1 rows)	<input type="checkbox"/>	Ana_gonzales

Choices in Cluster

2 — 3

Rows in Cluster

2 — 42

Average Length of Choices

9.66 — 13.5

Length Variance of Choices

0.47000000000000003 — 1.5

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

3. Podrá ver la siguiente información:

Cluster size: la cantidad de las diferentes versiones que el algoritmo muestra como similares.

Row count: El número de registros por cluster.

Values in cluster: Los valores seleccionados por el algoritmo para esa agrupación y el número de registros por valor.

Merge?: En este cuadro se selecciona si los valores se fusionan en el valor por defecto que propone el algoritmo.

New cell value: En este campo de texto se puede escribir un valor completamente nuevo para el clúster. También se puede hacer clic en cualquier valor para asignarlo como valor por defecto.

4. Vaya a **Keying Function**, seleccione **ngram-fingerprint** y en **Ngram Size** escriba **1**.

Cluster & Edit column "identificationQualifier"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method **key collision** Keying Function **ngram-fingerprint** Ngram Size **1** 2 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	57	<ul style="list-style-type: none">ricardo g (40 rows)Ricardo G (15 rows)G. Ricardo (1 rows)Ricardo G. (1 rows)	<input type="checkbox"/>	ricardo g
3	3	<ul style="list-style-type: none">Ana gonzales (1 rows)Ana Gonzales (1 rows)GonzalesAna (1 rows)	<input type="checkbox"/>	Ana_gonzales

Choices in Cluster: 3 — 4

Rows in Cluster: 3 — 57

Average Length of Choices: 9.5 — 12.67

Length Variance of Choices: 0.5 — 1.7

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Para conocer más acerca de los algoritmos (altamente recomendado) haga clic [aquí](#).

5. Para el primer cluster asigne un valor nuevo, para esto vaya al cuadro de texto de **New cell value** y escriba **Ricardo Gonzales**. Luego haga check en el cuadro de **Merge?:** para ese cluster.
6. Para el segundo clúster haga clic en **Ana Gonzales**, automáticamente el valor en **New cell value** cambiará y la casilla **Merge?** se chequeará.
7. Vaya a **Merge Selected & close** para fusionar los valores y volver a la ventana principal. El resultado del proceso debería verse así:

identificationQualifier change

2 choices Sort by: name count Cluster

Ana Gonzales 3

Ricardo Gonzales 57

(blank) 2342 edit include

Facet by choice counts

EXPORTACIÓN

EJERCICIO 7. Exportar el archivo

Existen múltiples maneras de exportar los archivos en OpenRefine, la siguiente es la que ha mostrado funcionar en todos los casos

1. En la esquina superior derecha haga clic en el botón **Export**
2. Seleccione **Custom tabular exporter...** aparecerá la ventana de exportación:

The screenshot shows the 'Custom Tabular Exporter' dialog box. It has four tabs: 'Content', 'Download', 'Upload', and 'Option Code'. The 'Content' tab is active. On the left, under 'Select and Order Columns to Export', there is a list of columns with checkboxes: ID (checked), Idioma, Base del registro, catalogNumber, occurrenceRemarks, individualCount, samplingProtocol, eventDate, eventTime, and habitat. Below this list are 'Select All' and 'De-select All' buttons. On the right, under 'Options for ID', there are two sections. The first section, 'For reconciled cells, output', has three radio buttons: 'Matched entity's name' (selected), 'Matched entity's ID', and 'Cell's content'. There is also a checkbox for 'Link to matched entity's page' (checked) and a checkbox for 'Output nothing for unmatched cells'. The second section, 'For date/time values, use format', has five radio buttons: 'ISO 8601, e.g., 2011-08-24T18:36:10+08:00' (selected), 'Short locale format', 'Long locale format', 'Medium locale format', and 'Full locale format'. There is also a 'Custom' option with a 'Help' link. At the bottom, there are three checkboxes: 'Output column headers' (checked), 'Output blank rows', and 'Ignore facets and filters and export all rows'. A 'Cancel' button is at the bottom left.

3. En la pestaña **Content** puede seleccionar las columnas que quiere exportar, si selecciona **Ignore facets and filters and export all rows** todos los facets y filtros serán ignorados, esto sucede cuando solo queremos hacer visualizaciones de los datos.
4. Vaya a la pestaña **Download** y seleccione el separador de caracteres que desee, en este caso se recomienda dejar las opciones que vienen por defecto. Haga clic en **Download** y guarde su archivo.
5. También puede exportar el proyecto siguiendo la ruta **Export > Export project**. De esta manera puede descargar el proyecto para trabajarlo en otro equipo.

USO AVANZADO

USO DEL API DE EOL

ANTES DE EMPEZAR

En este ejercicio se utilizará el API de la *Encyclopedia of Life* (EOL) a través de OpenRefine para obtener las categorías taxonómicas de una lista de nombres determinada.

EOL agrupa las clasificaciones de los grupos de organismos de diversos proveedores de contenido, cada uno de las cuales es soportado por una comunidad de científicos. Para una lista completa de los proveedores y descripción de los mismos ingrese a:

http://eol.org/api/docs/provider_hierarchies

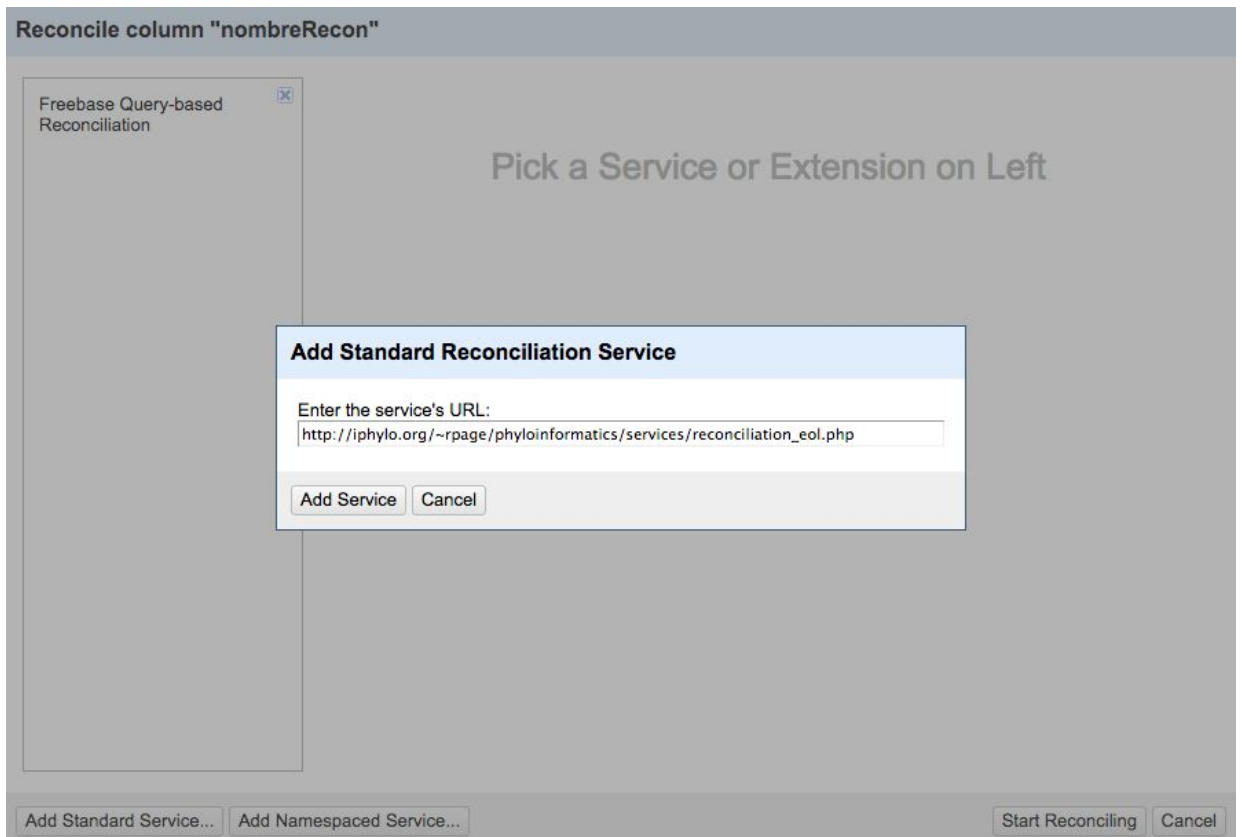
Entre los proveedores de clasificaciones se encuentran:

AntWeb	http://www.antweb.org/
Avibase	http://avibase.bsc-eoc.org/
CoL	http://www.catalogueoflife.org/
FishBase	http://www.fishbase.org/
Index Fungorum	http://www.indexfungorum.org/
ITIS	http://www.itis.gov/
IUCN	http://www.iucn.org/
Metalmark Moths of the World	http://choreutidae.lifedesks.org/
NCBI	http://www.ncbi.nlm.nih.gov/
The Paleobiology Database	http://www.paleodb.org/
The Reptile Database	http://www.reptile-database.org/
WoRMS	http://www.marinespecies.org/

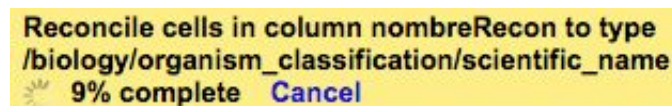
EJERCICIO 7. Obtención de las categorías

1. Vuelva a la venta principal de OpenRefine.
2. Vaya a la columna **Nombre Científico**. Es importante que estos nombres no contengan calificadores de como "cf.", "sp." o "spp.", de ser este el caso elimínelos y deje solamente como valor el nombre sin autoría.
3. Cree una columna llamada **nombreRecon** a partir de la columna **Nombre Científico**, haga clic en Menú Columna y luego siga la ruta **Edit column > Add column based on this column...** (solo modifique el campo **New column name** y deje los demás campos como están).
4. El siguiente paso es reconciliar los nombres con EOL, utilizando el servicio implementado por Roderic D.M. (<http://iphylo.blogspot.com>). Para esto haga clic en el Menú Columna de **nombreRecon**, y siga la ruta **Reconcile > Start reconciling...**
5. En la ventana que aparece haga clic en **Add Standard Service** (en la esquina inferior izquierda) y pegue la siguiente dirección:

http://iphylo.org/~rpage/phyloinformatics/services/reconciliation_eol.php



6. Haga clic en **Add Service** y espere un momento.
7. En la lista de servicios ahora podrá encontrar Encyclopedia of Life, selecciónelo y luego haga clic en **Start Reconciling**, después de un rato deberá ver una notificación:



8. Deberá ver un resultado como el siguiente:

▼ Nombre Científico	▼ nombreRecon
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match
Bolitoglossa phalarosoma	Bolitoglossa phalarosoma Wake and Brame, 1962 Choose new match

9. Como puede observar algunos nombres aparecen en color azul y otros en negro, los colores que están en azul han sido reconciliados automáticamente ya que solo había una opción disponible, los que están en negro tienen varias opciones para asignar. En este caso

- acepte todos los valores automáticamente, para esto en el Menú Columna de **nombreRecon** siga la ruta **Reconcile > Actions > Match each cell to its best candidate**
10. El siguiente paso es obtener el ID que asigna EOL para cada taxón, vaya al Menú Columna de **nombreRecon** y siga la ruta **Edit column > Add column based on this column...** en la ventana que aparece pegue en el campo de texto **Expression** la expresión **cell.recon.match.id** y nombre la columna como **EOL_ID**

Add column based on column nombreRecon

New column name:

On error: ☒ set to blank ☐ store error ☐ copy value from original column

Expression: Language: No syntax error.

Preview History Starred Help

row	value	cell.recon.match.id
1.	Bolitoglossa phalarosoma	330811
2.	Bolitoglossa phalarosoma	330811
3.	Bolitoglossa phalarosoma	330811
4.	Bolitoglossa phalarosoma	330811
5.	Bolitoglossa phalarosoma	330811
6.	Bolitoglossa phalarosoma	330811
7.	Bolitoglossa phalarosoma	330811

OK Cancel

11. Ahora se utilizará el API de EOL para obtener una lista de proveedores en los que se puede encontrar la taxonomía de cada taxón, vaya al Menú Columna de **EOL_ID** y en **Edit column** vaya a **Add column by fetching URLs...** e introduzca la expresión (tal y como aparece):

`"http://eol.org/api/pages/1.0/"+value+".json?images=2&videos=0&sounds=0&maps=0&text=2&iucn=false&subjects=overview&licenses=all&details=false&common_names=false&synonyms=false&references=false&vetted=0&cache_ttl="`

Cambie el nombre de la columna a **json_proveedores** y en el campo **Throttle delay** escriba 250, haga clic en **OK** y tenga paciencia ya que la consulta puede tomar un tiempo.

Add column by fetching URLs based on column EOL_ID

New column name
Throttle delay
 milliseconds

On error
☒ set to blank
☐ store error

Formulate the URLs to fetch:

Expression
Language
Google Refine Expression Language (GREL)

```
"http://eol.org/api/pages/1.0/"+value+".json?
images=2&videos=0&sounds=0&maps=0&text=2&iucn=false&subjects=overview&licen
ses=all&details=false&common_names=false&synonyms=false&references=false&ve
tted=0&cache_ttl="
```

No syntax error.

Preview
History
Starred
Help

row	value	"http://eol.org/api/pages/1.0/"+value+".json? images=2&videos=0&sounds=0&maps=0&text=2&iucn=false&subjects=overview&lic
1.	330811	http://eol.org/api/pages/1.0/330811.json? images=2&videos=0&sounds=0&maps=0&text=2&iucn=false&subjects=overview&licenses
2.	330811	http://eol.org/api/pages/1.0/330811.json? images=2&videos=0&sounds=0&maps=0&text=2&iucn=false&subjects=overview&licenses
3.	330811	http://eol.org/api/pages/1.0/330811.json? images=2&videos=0&sounds=0&maps=0&text=2&iucn=false&subjects=overview&licenses
4.	330811	http://eol.org/api/pages/1.0/330811.json?

OK
Cancel

12. Para obtener una lista legible agregue una columna basada en `json_proveedores` (**Edit column** > **Add column based on this column...**) usando la expresión:

```
forEach(value.parseJson().get('taxonConcepts'),entry,entry.get('nameAccording To')).uniques().join(", ")
```

Nómbrela `lista proveedores`:

Add column based on column json_proveedores

New column name:

On error: ☒ set to blank ☐ store error ☐ copy value from original column

Expression: Language: No syntax error.

Preview History Starred Help

row	value	forEach(value.parseJson().get('taxonConcepts'),entry)
1.	{ "identifier": 330811, "scientificName": "Bolitog phalarosoma Wake and Brame, 1962", "richness_score": 46.5845, "synonyms": [], "vernacularNames": [], "references": [], "taxonConcepts": [{ "identifier": 52658338, "scientificName": "Boli phalarosoma Wake and Brame, 1962", "nameAccordingTo": "Species 2000 & ITIS Catalogue of Life: April 2013, Integrating Red List (Species Assessed for Global Conservation), GE" }] }	

OK Cancel

13. EOL ofrece una serie de proveedores que se pueden consultar a través del API (http://eol.org/api/provider_hierarchies/1.0.json) al momento de realización de esta guía se encuentran:

"AntWeb (Ant Species)"
 "WORMS Species Information (Marine Species)"
 "FishBase (Fish Species)"
 "IUCN Red List (Species Assessed for Global Conservation)"
 "Index Fungorum"
 "Cephaloleia LifeDesk resource"
 "The Reptile Database"
 "GBIF Nub Taxonomy"
 "Avibase - IOC World Bird Names (2011)"
 "Integrated Taxonomic Information System (ITIS)"
 "Paleobiology Database"
 "Clements Checklist resource"
 "Taxonomic Hierarchy of COL-China 2012"
 "South African Vertebrates"
 "Continenticola Classification"
 "NCBI Taxonomy"
 "Species 2000 & ITIS Catalogue of Life: April 2013"
 "Metalmark Moths of the World"
 "Algeabase resource"
 "Algae"

14. El siguiente paso es obtener el ID del proveedor de interés. En este caso se utilizará *Catalogue of Life* ya que es el que más se presenta a lo largo de los taxa que se están usando, pero si lo desea se puede usar cualquiera de la lista que se obtuvo anteriormente. Para obtener el ID agregue una columna llamada **COL_ID** basada en **json_proveedores** utilizando la expresión:

```
forEach(value.parseJson().get('taxonConcepts'),entry,if(entry.get('nameAccordingTo')== "nombre_de_la_fuente",entry.get("identifier"),"")).uniques().join(",")
```


Tenga en cuenta que en "**nombre_de_la_fuente**" este valor debe reemplazarse por el nombre exacto que aparece en el punto anterior, en este caso "**Species 2000 & ITIS Catalogue of Life: April 2013**".

Add column based on column json_proveedores

New column name:

On error: ☒ set to blank ☐ store error ☐ copy value from original column

Expression: `forEach(value.parseJson().get('taxonConcepts'),entry,if(entry.get('nameAccordingTo')=="Species 2000 & ITIS Catalogue of Life: April 2013",entry.get("identifier"),"")).uniques().join(",")` Language: Google Refine Expression Language (GREL)

No syntax error.

Preview History Starred Help

row	value
1.	<pre>{ "identifier": "330811", "scientificName": "Bolitog 52658338 phalarosoma Wake and Brame, 1962", "richness_score": 46.5845, "synonyms": [], "vernacularNames": [], "references": [], "taxonConcepts": [{ "identifier": "52658338", "scientificName": "Boli phalarosoma Wake and Brame, 1962", "nameAccordingTo": "Species 2000 & ITIS Catalogue of Life: April" }] }</pre>

OK Cancel

15. Con el ID en la nueva columna ya se puede obtener la jerarquía taxonómica del proveedor seleccionado, para esto agregue una columna basada en **COL_ID** utilizando **Add column by fetching URLs...** y la expresión:

`"http://eol.org/api/hierarchy_entries/1.0/"+value+".json?common_names=false&synonyms=false&cache_ttl="`

Recuerde nombrar la columna como **COL_tax** y en el campo **Throttle delay** escribir 250, haga clic en **OK** y tenga paciencia ya que la consulta puede tomar un tiempo.

16. Finalmente se puede hacer esta información legible agregando una o varias columnas basadas en **COL_tax** se tienen dos opciones:

- a. Utilizar la expresión:

`forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank')=="family",entry.get("scientificName"),"")).uniques().join("")`

Y reemplazar **family** por la categoría taxonómica que deseemos, esto nos generara una columna para cada categoría.

- b. Concatenar las expresiones de reino, familia, filo, clase, orden y familia con un signo **+** después de cada expresión.

`forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank')=="kingdom",entry.get("scientificName"),"")).uniques().join("")+forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank')=="phylum",entry.get("scientificName"),"")).uniques().join("")`

```
Name"), "").uniques().join("")
```

Y así sucesivamente, esto nos arrojará una columna con todos los valores.

17. En este ejemplo se utilizará la opción b, vamos a **COL_tax** y en el *Menú Columna* agregamos una nueva columna llamada **COL** utilizando la expresión:

```
forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank') == "kingdom", entry.get("scientificName"), "")).uniques().join("")+  
", "+forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank')== "phylum", entry.get("scientificName"), "")).uniques().join("")+  
", "+forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank') == "class", entry.get("scientificName"), "")).uniques().join("")+  
", "+forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank')== "order", entry.get("scientificName"), "")).uniques().join("")+  
", "+forEach(value.parseJson().get('ancestors'),entry,if(entry.get('taxonRank') == "family", entry.get("scientificName"), "")).uniques().join("")
```

18. Para separar las categorías en varias columnas, en el Menú Columna de **COL** siga la ruta **Edit column > Split into several columns...**, haga clic en **OK** y verá los valores separados.

Split column COL into several columns

How to Split Column

☒ by separator
Separator ☐ regular expression
Split into columns at most (leave blank for no limit)

☐ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

☒ Guess cell type
☒ Remove this column

OK Cancel

Para cambiar los nombres de las columnas, ingrese al Menú columna y siga la ruta **Edit column > Rename this column** y asigne el nombre que corresponda.

ENLACES ADICIONALES

Name validation Tutorial:

https://docs.google.com/document/d/1tkDRXIYhmassYAk5T4v5oac5prF0jAiSMr_JEGTvhRo/edit

Higher Taxonomy Tutorial:

https://docs.google.com/document/d/1XZ_pM9gldQzHzl8wfUCVea-52yub5T_3tc-snBgPRa0/edit

Más servicios de reconciliación:

<http://iphylo.blogspot.com/2012/02/using-google-refine-and-taxonomic.html>

Documentación para usuarios del programa:

<https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users>

Api de EOL:

<http://eol.org/api>

Lista de recursos disponibles para OpenRefine:

<https://github.com/OpenRefine/OpenRefine/wiki/External-Resources>