

Guía de Uso Básico de



para la limpieza de datos sobre biodiversidad

Zermoglio P, Wieczorek J.
Versión 2.0 - Abril 2018

Comentarios Preliminares

La presente guía ha sido construida con fines únicamente pedagógicos. Los ejemplos presentados corresponden a un conjunto de datos de biodiversidad que ha sido específicamente modificado por los autores.

El conjunto de datos modificado puede obtenerse a través de: <https://tinyurl.com/yd8mekzw>. Descargar el archivo en formato .CSV, y no manipularlo en Excel, dado que ello puede cambiarle la codificación.

El conjunto de datos original utilizado fue:

Williams J (2011). Colección de Herbario. Facultad de Ciencias Naturales y Museo - U.N.L.P. Occurrence Dataset <https://doi.org/10.15468/i9bj5r> accessed via GBIF.org.

Los ejemplos de uso presentados en esta guía constituyen sólo algunas de las alternativas posibles para el tratamiento de datos en OpenRefine.

En el texto, los **nombres de los campos originales** se marcan con color verde claro.

Los autores agradecen a los colaboradores Anabela Plos (Museo Argentino de Ciencias Naturales y GBIF Argentina) y David Bloom (VertNet) por sus comentarios y sugerencias.

Índice

1. Carga de datos y creación de un proyecto	5
2. Limpieza de datos	8
A. Uso de Facetas	8
<i>Facetas de texto</i>	9
<i>Facetas y espacios en blanco</i>	10
<i>Facetas y duplicados</i>	12
<i>Facetas en múltiples columnas</i>	14
<i>Número de elecciones límite en las Facetas</i>	16
B. Deshacer y rehacer cambios	16
<i>Deshacer pasos</i>	16
<i>Guardar pasos para rehacer luego</i>	17
<i>Rehacer pasos guardados</i>	18
C. Uso de Filtros	19
<i>Filtros simples</i>	19
<i>Filtros con expresiones regulares</i>	22
D. Uso de Agrupamientos	23
<i>Agrupamientos simples</i>	23
E. Nuevos campos (columnas)	25
<i>Manejo básico de columnas</i>	25
<i>Nuevas columnas vacías</i>	27
<i>Nuevas columnas a partir transformaciones simples de otras columnas</i>	28
i. Concatenaciones.	28
ii. Divisiones.	29
F. Marcado de registros: banderas y estrellas	31
<i>Marcado con banderas y estrellas</i>	31
<i>Conservación de banderas y estrellas en la exportación</i>	32
<i>Uso de banderas y estrellas para eliminar registros</i>	33
3. Guardado y exportación de datos y proyectos	33
A. Guardado de datos y proyectos	33

B. Exportación de datos y proyectos	34
4. Consultas a servicios externos	36
A. Consultas externas a través de URLs	37
<i>Resolución de nombres científicos usando Global Names Resolver</i>	<i>37</i>
<i>Georreferenciación usando GeoLocate.....</i>	<i>41</i>
<i>Limpieza de fechas utilizando Canadensys Date Parsing</i>	<i>46</i>
B. Servicios de reconciliación	52
<i>Reconciliación de nombres científicos utilizando Encyclopedia of Life</i>	<i>52</i>

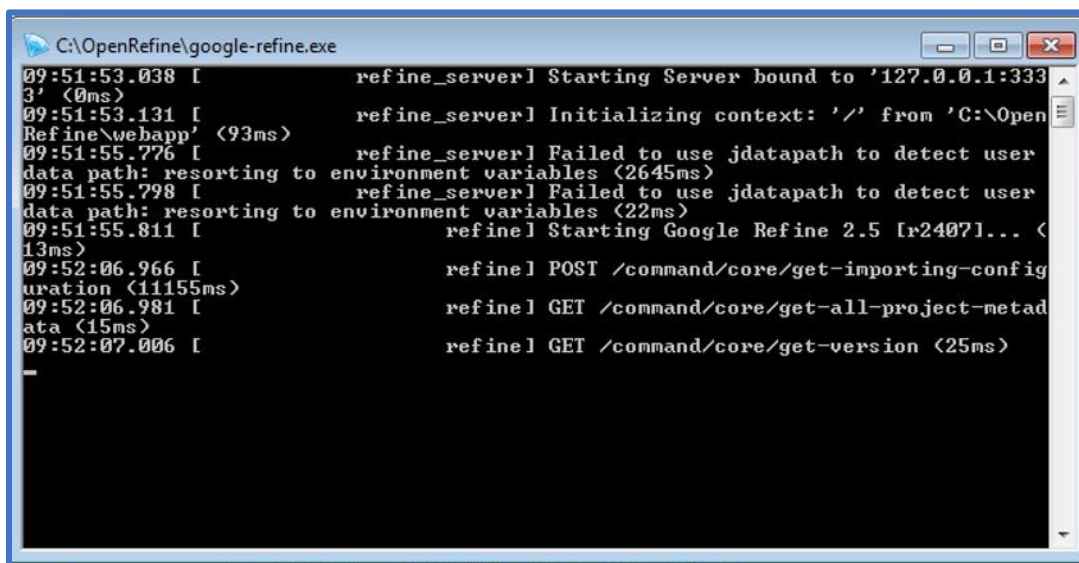
Guía de Uso Básico de **Refine**^{OPEN}

para la limpieza de datos sobre biodiversidad

1. Carga de datos y creación de un proyecto

Para comenzar a utilizar OpenRefine debe cargar sus datos en el programa y crear un proyecto. Para ello, siga los siguientes pasos:

1. **Abra la aplicación OpenRefine.** Si utiliza Windows, se abrirá una ventana de comandos que mostrará las acciones que se OpenRefine está realizando (Figura 1). No cierre esta ventana mientras esté trabajando con el programa.



```

C:\OpenRefine\google-refine.exe
09:51:53.038 [      refine_server] Starting Server bound to '127.0.0.1:3333' <0ms>
09:51:53.131 [      refine_server] Initializing context: '/' from 'C:\OpenRefine\webapp' <93ms>
09:51:55.776 [      refine_server] Failed to use jdatapath to detect user data path: resorting to environment variables <2645ms>
09:51:55.798 [      refine_server] Failed to use jdatapath to detect user data path: resorting to environment variables <22ms>
09:51:55.811 [      refine] Starting Google Refine 2.5 [r24071]... <13ms>
09:52:06.966 [      refine] POST /command/core/get-importing-config?duration <11155ms>
09:52:06.981 [      refine] GET /command/core/get-all-project-metadata <15ms>
09:52:07.006 [      refine] GET /command/core/get-version <25ms>

```

Figura 1.

OpenRefine se abrirá en el navegador que usted utilice por defecto inmediatamente después de ejecutar la aplicación (Figura 2). Si OpenRefine no abre, puede acceder manualmente ingresando la siguiente URL en su navegador:

<http://127.0.0.1:3333>

En el menú de la izquierda tiene las opciones para crear, abrir o importar proyectos y para cambiar la configuración de idiomas. (Si usted no tiene ningún proyecto aún, en la opción de “Abrir proyecto” verá una lista vacía.)

2. **Cargue los datos** (Figura 2). Dentro de la opción “Crear proyecto”, escoja el archivo que desea cargar. Note que hay varios formatos posibles de archivos que se pueden subir (tsv, csv, xls, json, etc). Haga click en “Next”.

NOTA: Si sube archivos con formato .xls o .xlsx, tenga en cuenta que no podrá modificar la codificación, y que pueden encontrarse algunos errores en los datos (ejemplo: los tildes en las palabras se verán como símbolos raros cuando cargue los datos). Para evitarse problemas, si trabaja con Excel es conveniente que exporte los datos como archivo .csv (de todas formas, tenga cuidado con la codificación, ver más abajo).

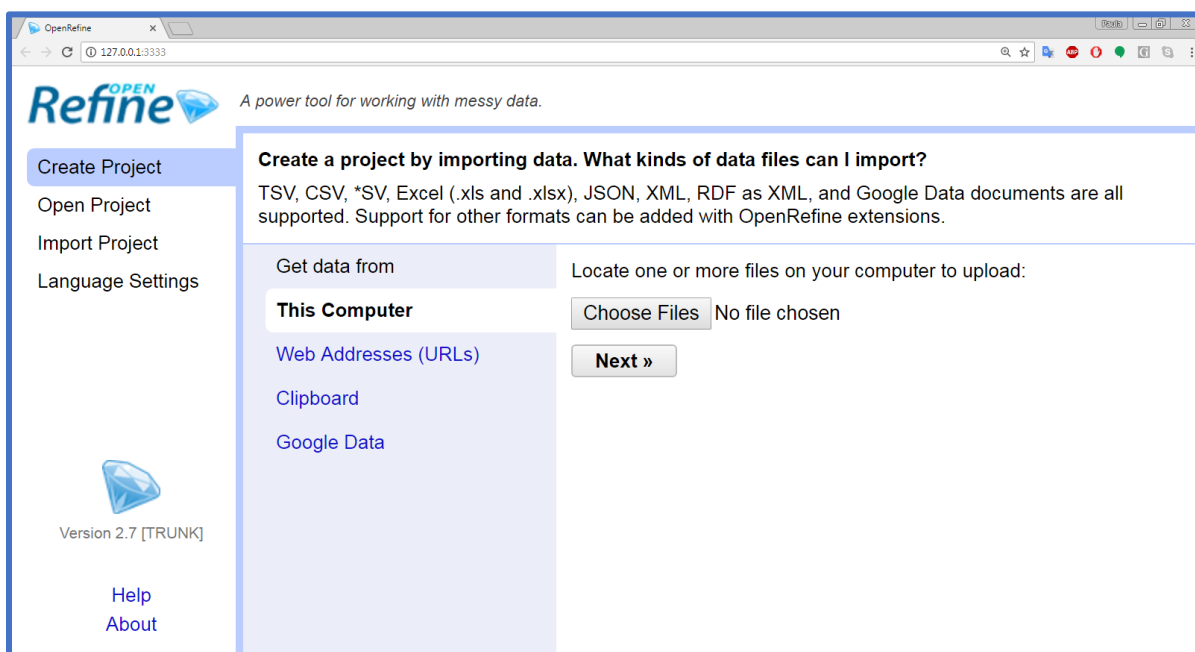


Figura 2.

Verá entonces una pantalla como la que se muestra en la Figura 3. Allí puede ver una muestra de sus datos y puede modificar varios aspectos de la carga de los datos al programa: codificación, criterio para la separación en columnas, inclusión o no de la primera fila, etc.

OpenRefine sugiere algunas de las codificaciones más utilizadas cuando se hace click en el cuadro de texto “Character encoding”. Asegúrese de escoger correctamente la codificación en el campo (UTF-8) para el archivo de ejemplo (Figura 4).

Asegúrese también de deseleccionar la opción “parse cell text into numbers, dates,...”.

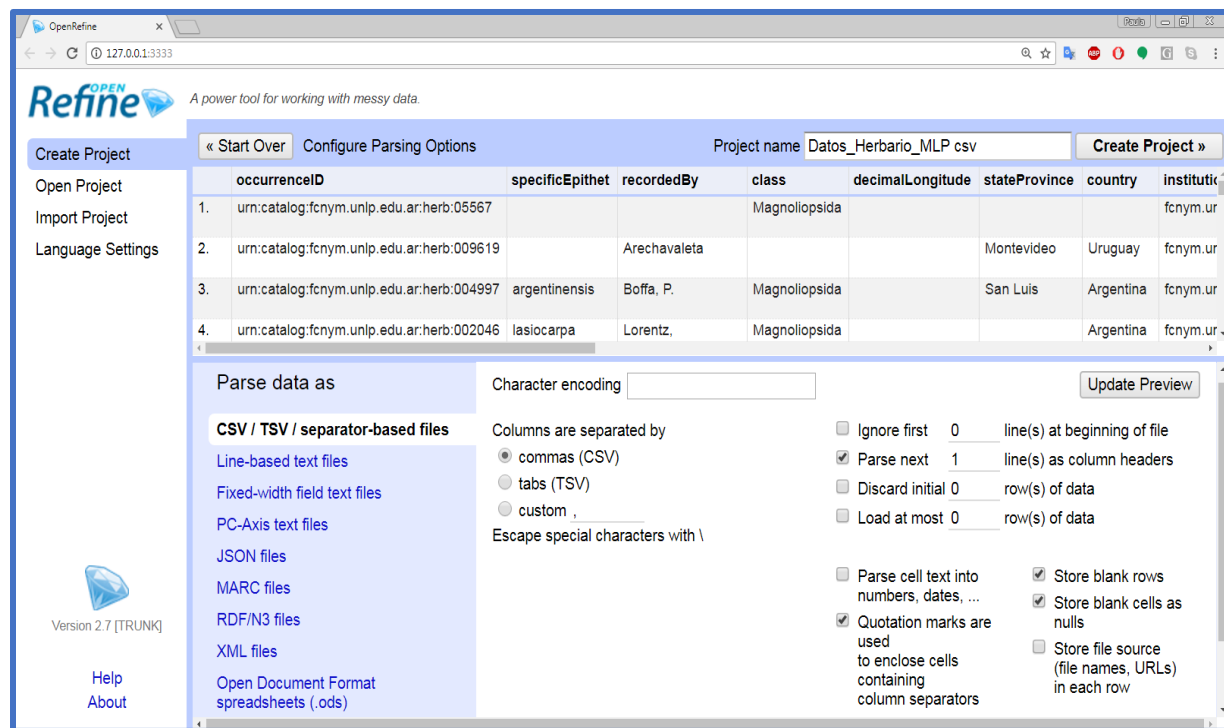


Figura 3.



Figura 4.

3. **Cree el proyecto.** Una vez que haya seleccionado las opciones de carga de datos, haga click en el botón “Create Project” arriba a la derecha.
4. **¡Felicitaciones!** Ya tiene un proyecto (lo verá como en la Figura 5).

24984 rows

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 10 next > last »

	occurrenceID	specificEpithet	recordedBy	class	decimalLongitude	stateProvince	country	institutionCode
1.	urn:catalog:fcnym.unip.edu.ar:herb:05567			Magnoliopsida				fcnym.unip.edu.ar
2.	urn:catalog:fcnym.unip.edu.ar:herb:009619		Arechavaleta			Montevideo	Uruguay	fcnym.unip.edu.ar
3.	urn:catalog:fcnym.unip.edu.ar:herb:004997	argentinensis	Boffa, P.	Magnoliopsida		San Luis	Argentina	fcnym.unip.edu.ar
4.	urn:catalog:fcnym.unip.edu.ar:herb:002046	lasiocarpa	Lorentz, Paul(Pablo) Günther	Magnoliopsida			Argentina	fcnym.unip.edu.ar
5.	urn:catalog:fcnym.unip.edu.ar:herb:002052	sprengeliana	Gardner, George	Magnoliopsida				fcnym.unip.edu.ar
6.	urn:catalog:fcnym.unip.edu.ar:herb:002048	doniana		Magnoliopsida				fcnym.unip.edu.ar
7.	urn:catalog:fcnym.unip.edu.ar:herb:002059	calicasana	Funck, Nicolas	Magnoliopsida				fcnym.unip.edu.ar
8.	urn:catalog:fcnym.unip.edu.ar:herb:002063	parviflora	Wright, John	Magnoliopsida			México	fcnym.unip.edu.ar
9.	urn:catalog:fcnym.unip.edu.ar:herb:001950	macdonaldii	Mac, Donald	Magnoliopsida				fcnym.unip.edu.ar
10.	urn:catalog:fcnym.unip.edu.ar:herb:002082	salicifolium	Valesquez, J.	Magnoliopsida				fcnym.unip.edu.ar

Figura 5.

NOTA: el número de líneas cargadas se muestra en este momento arriba de la tabla, aunque el número de filas mostradas en la tabla sea limitado. No desespere, OpenRefine sólo muestra hasta 50 líneas, pero las acciones que uno pueda tomar en la aplicación pueden tener efecto sobre filas aunque éstas no sean mostradas.

2. Limpieza de datos

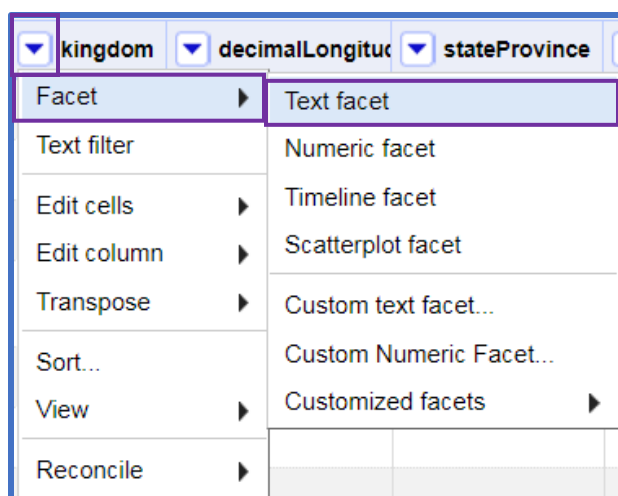
A. Uso de Facetas

La función “Facet” permite la visión general de los datos y el filtrado en bloque de grupos de registros. Es de gran ayuda cuando se quieren visualizar o modificar datos en varios registros a la vez. Las facetas se pueden aplicar a celdas que contengan cualquier tipo de texto, números o fechas.

A continuación llevaremos a cabo una serie de cambios en los datos de un archivo descargado del portal de datos de GBIF.org y modificado específicamente para mostrar las funcionalidades del programa.

Facetas de texto

Ubique la columna **kingdom** y haga click sobre la flecha azul. Dentro de “Facet”, escoja “Text facet”, como se muestra a continuación (Figura 6a). Se abrirá entonces a la izquierda una ventana con la faceta (Figura 6b).



a.



b.

Figura 6.

En dicha ventana de faceta, puede ordenar los valores alfabéticamente (haciendo click sobre “name”) o según el número de registros asociados a cada valor (haciendo click sobre “count”).

En la lista de valores podemos ver que hay algunos errores. Para corregirlos coloque el cursor sobre el valor que desea modificar y haga click en “edit”. Se abrirá entonces una pequeña ventana donde puede cambiar el valor (Figura 7). Para guardar el cambio haga click en “Apply”, ello aplicará el cambio a todos aquellos registros que tenían el valor dado.

Corrija los valores “Plante” y “Plants”. Cuando lo haga, habrá corregido todos los registros que contenían esos valores, y se modificará entonces el número de registros que tiene el valor “Plantae”.

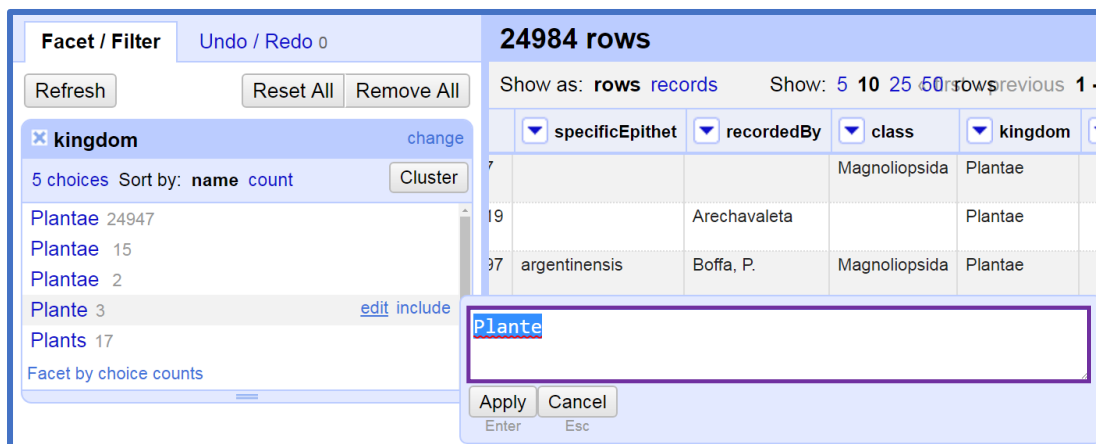


Figura 7.

Facetas y espacios en blanco

1. Espacios en blanco extra al principio o al final de una cadena de texto

Una vez que haya corregido los valores en el punto anterior, notará que aún aparecen 3 valores “Plantae”, aparentemente iguales (Figura 8). Sin embargo, estos valores sí son diferentes: tienen espacios adicionales al final del texto.



Figura 8.

Para corregir estos errores, asegúrese de que ninguno de los valores en la faceta están seleccionados y que el número de registros que se muestra arriba de la tabla es el total (24984). Sobre la columna **kingdom**, haga click sobre la flecha azul. Dentro de “Edit cells”, escoja “Common transforms”, y allí “Trim leading and trailing whitespace” (Figura 9). Esta función le permite eliminar espacios en blanco que puedan aparecer al principio y al final de cadenas de texto. Cuando termine este paso, los 24984 registros deberían tener el valor “Plantae” en la columna **kingdom**.

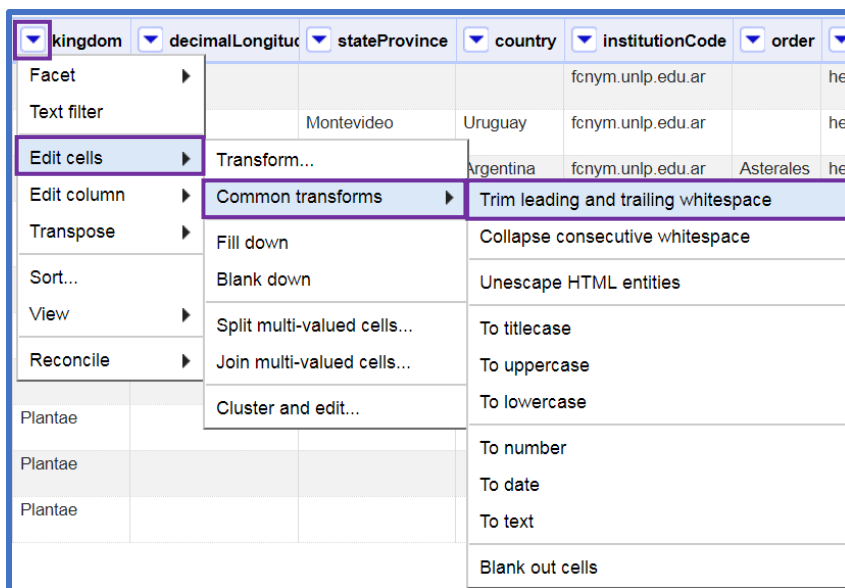


Figura 9.

2. Espacios en blanco extra entre palabras en una cadena de texto

A veces en campos que contienen cadenas de texto con varias palabras puede haber espacios en blanco extra entre palabras. Para ver un ejemplo, ubique la columna `stateProvince` en el conjunto de datos. Arme una faceta de texto para dicha columna (click sobre la flecha azul --> Facet --> Text facet). Luego, en la faceta, ordene los valores por número de registros asociados (seleccionando "count"). Verá entonces los valores que se encuentran en este campo como se muestra en la Figura 10.

Note que en primer y tercer lugar figura aparentemente el mismo valor, "Buenos Aires". La diferencia entre ambos valores es que uno de ellos tiene un doble espacio entre las palabras.



Figura 10.

Para corregir este error, sobre la columna **stateProvince**, haga click sobre la flecha azul. Dentro de “Edit cells”, escoja “Common transforms”, y allí “Collapse consecutive whitespaces” (Figura 11). Esta función le permite convertir múltiples espacios en blanco en un único espacio en blanco.

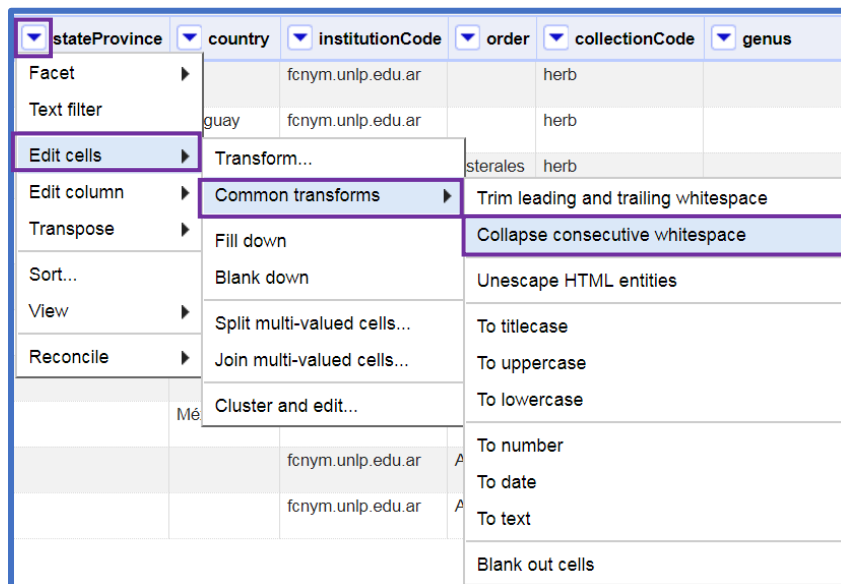


Figura 11.

Una vez que haya removido los espacios en blanco extra, en la faceta sólo verá un valor para “Buenos Aires”, con un número de registros que es la suma de los valores anteriores. Tenga en cuenta que si había otros valores con el mismo problema de dobles espacios entre palabras en esta misma columna, la modificación se aplicará a todos ellos, y no sólo a Buenos Aires. Puede comprobar cuántos valores se han modificado comprobando el número de valores disponibles en la faceta antes y después de la transformación.

Facetas y duplicados

Las facetas también permiten la detección y corrección de duplicados.

NOTA: cuando hablamos aquí de duplicados, nos referimos a valores duplicados dentro de una columna, no necesariamente a registros enteros duplicados, o a duplicados en el sentido biológico/de colecciones. Por ello, tenga especial cuidado a la hora de actuar sobre estos valores duplicados, pues podrían tener efectos a diferentes niveles.

Veremos un ejemplo de duplicados en la columna **catalogNumber**. Para ello, haga click en la flecha azul y luego escoja “Facet” --> “Customized facets” --> “Duplicates facets”. Verá una ventana como la mostrada en la Figura 12, donde “true” implica los valores duplicados.

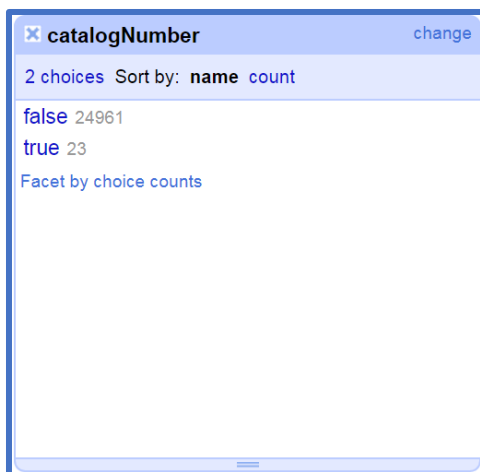


Figura 12.

Si hace click en “true”, la pantalla principal le mostrará los registros que tienen número de catálogo duplicado (Figura 13). Observe por ejemplo los siguientes registros:

- el primer y quinto registros tienen el mismo número de catálogo, 5567
- el tercer registro (y otros más abajo que no son visibles entre los diez primeros) no tiene número de catálogo (el valor nulo es lo que está duplicado).
- etc.

Facet / Filter		23 matching rows (24984 total)							Extensions:
Refresh Reset All Remove All catalogNumber change invert reset 2 choices Sort by: name count false 24961 true 23 exclude Facet by choice counts		Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 23 next > last »							
	locality	decimalLatitude	recordNumber	scientificNameA	infraspecificEpi	catalogNumber	taxonRank		
			5567			5567	subclassis		
			13305	Hieron.		13305	species		
	Hauteurs an-dessus d'Ushuaia		2246	Albov		2246	species		
	Emel-keik (Rio Chico) Patagonia		5567	Speg.		5567	species		
onaceae			4677	Viv.		4677	species		
onaceae	E. Sussex		4677	Viv.		4677	species		
ceae	Los Toldos, Finca "El Nogal" (de C. Burry), Río Huaico		4978	(Poir.) A.R.Sm. R.C.Moran		4978	species		
ceae	Los Toldos, Finca "El Nogal" (de C. Burry), Río Huaico		4978	(Poir.) A.R.Sm. R.C.Moran		4978	species		
	In horto		1697	Speg.	major	1697	forma		

Figura 13.

Corrija los números de catálogo. Para hacerlo, edite las celdas individualmente: sobre la celda haga click en el botón “edit”, modifique el valor y haga click en “Apply” (Figura 14).

NOTA: en la práctica la corrección de los números de catálogo sólo debe hacerse una vez que los números y los datos asociados han sido comprobados con las etiquetas de los especímenes.

decimalLatitude	recordNumber	scientificNameA	infraspecificEpit	catalogNumber
	5567	Speg.		5567
	4677	Viv.		4677
	4677	Viv.		4677
	4978	(Poir.) A.R.Sm. R.C.Moran		4978
	4978	(Poir.) A.R.Sm. R.C.Moran		4978
	1697	Speg.	major	1697

Figura 14.

Facetas en múltiples columnas

Las facetas también pueden aplicarse a combinaciones de campos (columnas), y actuar como filtros múltiples.

Por ejemplo, si quisiéramos seleccionar los registros que tienen “Buenos Aires” como valor en el campo `stateProvince` y “Argentina” como valor en el campo `country`, podemos hacer lo siguiente:

Sobre uno de los dos campos hacer click en la flecha azul --> Facet --> Custom text facet... (Figura 15).

Figura 15.

Se abrirá una ventana como la que se muestra en la Figura 16. En el cuadro de texto, pegue la siguiente expresión:

```
cells["stateProvince"].value=="Buenos Aires" and cells["country"].value=="Argentina"
```

Dicha expresión significa que se armará la nueva faceta con las siguientes condiciones: los valores del campo `stateProvince` (`cells["stateProvince"].value`) deben ser iguales a (`==`) "Buenos Aires" y (`and`) los valores del campo `country` (`cells["country"].value`) deben ser iguales a (`==`) "Argentina".

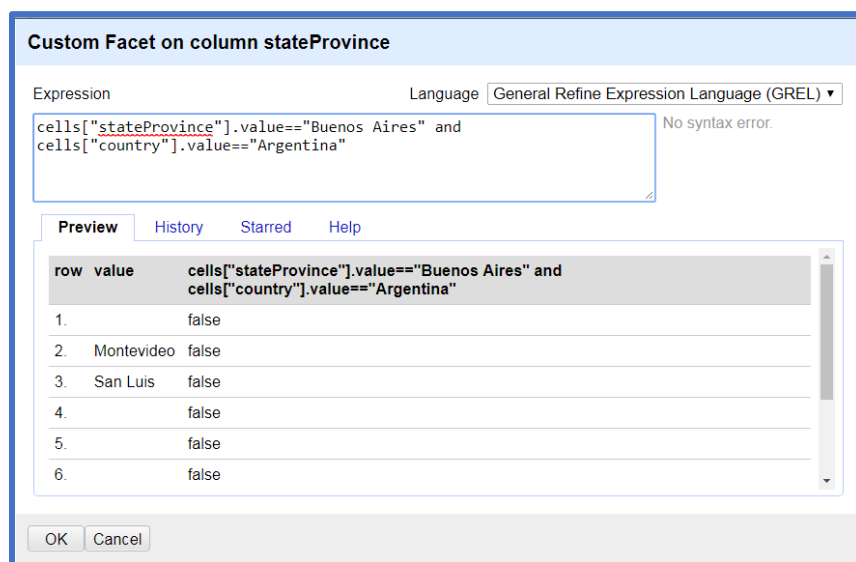


Figura 16.

El resultado de la faceta mostrará valores verdaderos y falsos para esa condición (Figura 17). Si se hace click sobre "True" se verán los registros para los cuales la provincia es Buenos Aires y el país es Argentina.

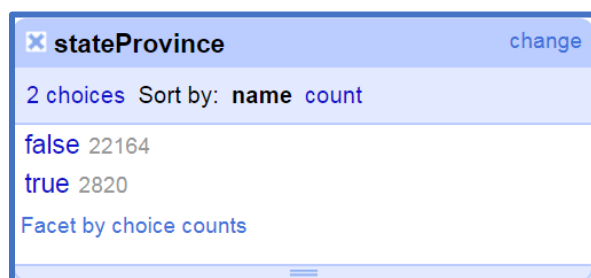


Figura 17.

Tenga en cuenta que como utilizamos "y" en la expresión, deben cumplirse las dos condiciones (provincia y país con los valores especificados) para que el resultado sea verdadero. En otras palabras, registros que tengan como país Argentina pero que no tengan Buenos Aires como provincia no serán incluidos, y viceversa.

Número de elecciones límite en las Facetas

En OpenRefine existe un límite para el número de elecciones de faceta que se muestran (“choices”). Muchas veces dicho número está pre-configurado a un valor de 2000. Ello quiere decir que sólo podrá ver 2000 opciones dentro de la faceta de interés.

Para modificar este límite, vaya a la siguiente dirección en su navegador web:

<http://127.0.0.1:3333/preferences>

Se abrirá entonces una ventana como la mostrada en la Figura 18. Allí, establezca el límite preferido para las facetas editando la clave "ui.browsing.listFacet.limit".

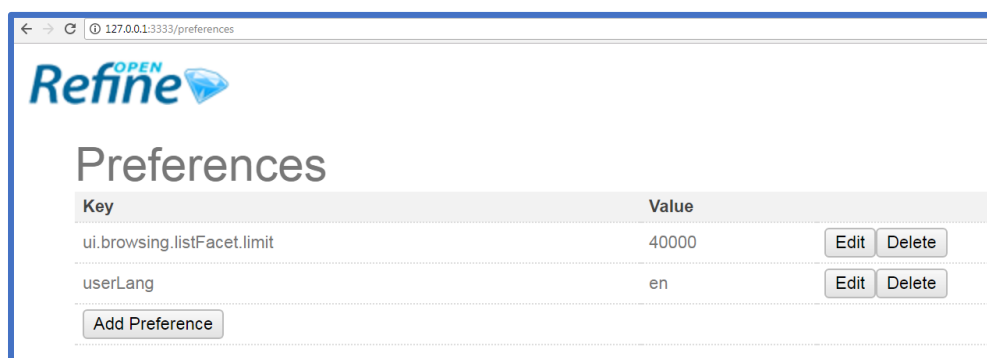


Figura 18.

B. Deshacer y rehacer cambios

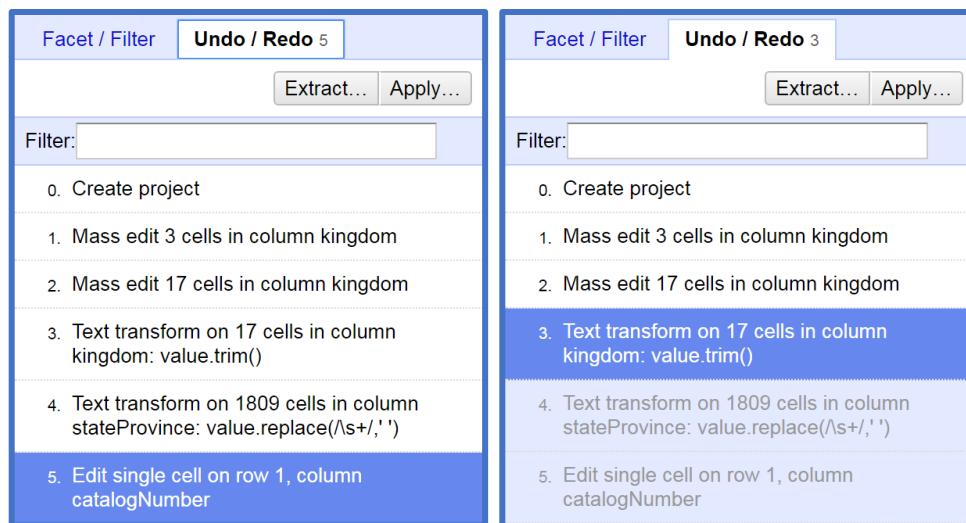
Ahora que ya ha acumulado una serie de modificaciones al conjunto de datos, veamos cómo se pueden deshacer y rehacer cambios.

En el menú de la izquierda, seleccione la pestaña “Undo/Redo”, que tiene un número al lado correspondiente al número de cambios acumulados hasta ahora. Verá entonces una lista de pasos realizados, como se muestra en la Figura 19a.

Note que el paso resaltado en azul es el que determina el estado de los datos. Todos los pasos hasta el resaltado, inclusive, han sido aplicados a los datos. Todos aquellos pasos ubicados después del paso resaltado no han sido aplicados.

Deshacer pasos

Si quiere deshacer todo lo posterior a algún paso, simplemente haga click sobre el paso inmediatamente anterior. Por ejemplo, si quiere deshacer los últimos dos pasos de cinco, haga click en el paso 3, y los dos últimos se revertirán automáticamente (Figura 19b).



a.

b.

Figura 19.

Para rehacer un paso luego de haberlo deshecho, simplemente haga click en ese paso.

IMPORTANTE:

El hacer y deshacer en OpenRefine trabaja sobre “estados”. Eso quiere decir que uno puede ir y volver a estados determinados, por ejemplo, el estado de los datos una vez que se han hecho ciertas modificaciones. Ello implica que si uno vuelve a un estado anterior y luego realiza una nueva modificación a partir de ese estado, entonces perderá los pasos originales y no podrá recuperarlos. En el ejemplo de la Figura 19, si uno vuelve al paso 3. y luego realiza sobre los datos alguna otra operación, no podrá volver a esos pasos 4. y 5. previos.

Guardar pasos para rehacer luego

Es importante entonces que guarde sus pasos, especialmente para pasos más complejos. Para ello, en la pestaña “Undo/Redo”, haga click en el botón “Extract”. Se abrirá una nueva ventana, como se muestra en la Figura 20, donde puede seleccionar los pasos que desea guardar. Los pasos están dados en formato JSON¹ en el panel de la derecha.

Copie las expresiones de los pasos de interés que se muestran a la derecha a un procesador de texto (e.g., Notepad, MS Word, etc.) y guárdelas para uso posterior (en caso de que no esté familiarizado con el formato JSON, recuerde tomar nota de qué cambios representan esas expresiones).

¹ JSON (Java Script Object Notation) es un formato que utiliza texto legible para los humanos para transmitir datos en la forma de pares de atributo:valor y de matrices de datos.

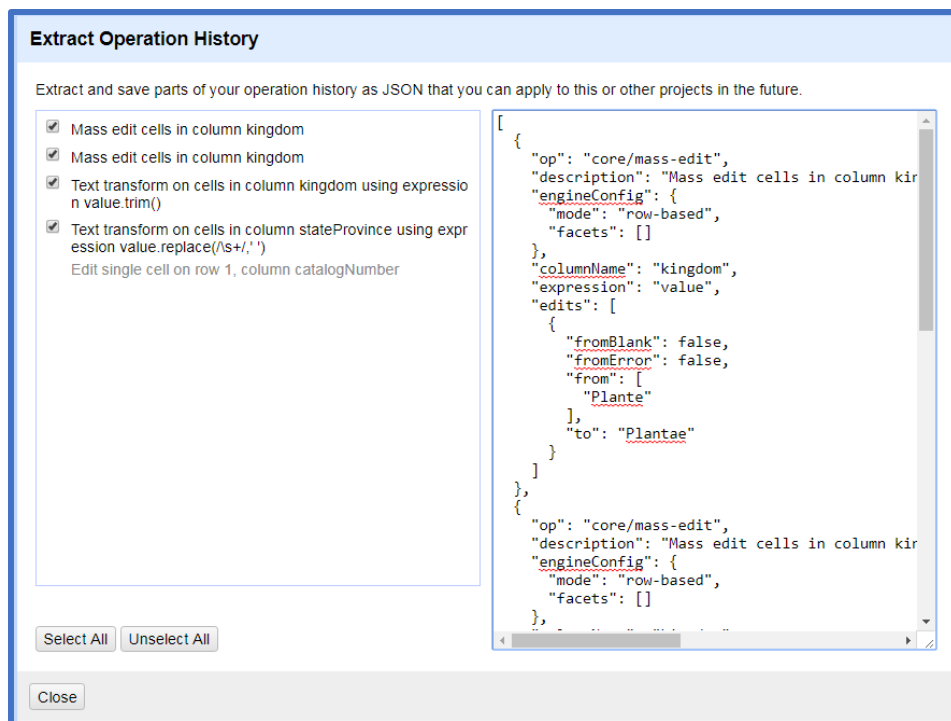


Figura 20.

IMPORTANTE:

Los cambios hechos a celdas particulares no tienen la opción de guardar expresiones. En el ejemplo anterior, Figura 20, note que el cambio en una celda única del número de catálogo figura en gris y no puede ser seleccionado. Esto es una limitación actual de OpenRefine, por lo que si va a deshacer un cambio de esta naturaleza pero quiere rehacerlo luego, deberá tomar nota usted mismo de cuál fue el cambio y en qué celda de forma separada (e.g, "Cambié el número de catálogo del registro X, de "1234" a "1236""; parece tonto, pero créame que si no lo anota en algún lado seguramente se va a olvidar).

Rehacer pasos guardados

Si desea rehacer pasos que tenga guardados (en formato JSON), dentro de la pestaña "Undo/Redo" haga click en el botón "Apply". Se abrirá entonces una ventana como la que se muestra en la Figura 21, pero vacía.

Pegue en el cuadro de texto la expresión deseada (copie y pegue lo que guardó en su procesador de texto en el apartado anterior) y haga click en "Perform Operations".

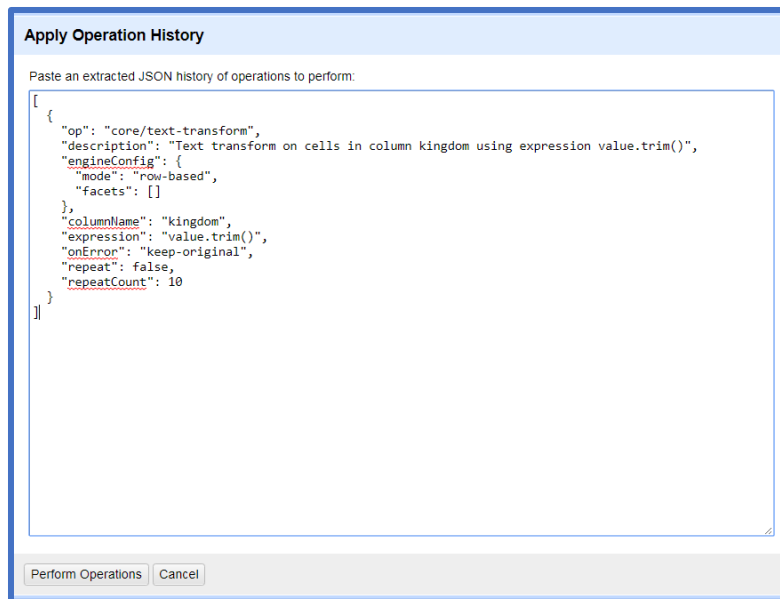


Figura 21.

C. Uso de Filtros

Filtros simples

OpenRefine permite el uso de filtros sobre campos particulares, función que puede ser muy útil para la limpieza de datos. Veremos un ejemplo a continuación.

Ubique el campo **specificEpithet** y cree una faceta de texto (haga click en la flecha azul, --> Facet --> Text facet). Luego vaya nuevamente a la flecha azul y cree un filtro de texto ("Text filter"). Sobre el menú de la izquierda se abrirá una ventana como la que se muestra en la Figura 22.

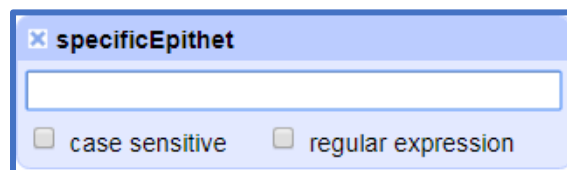


Figura 22.

En el cuadro de texto puede escribir el valor sobre el cual desea filtrar.

Por ejemplo, pruebe escribiendo "sp.".

En el menú de la izquierda, dentro de la faceta se mostrará el valor que usted buscó, y en la pantalla principal se mostrarán los registros asociados que tienen dicho valor en el campo **specificEpithet** (Figura 23).

Note que verá dos valores, uno en letra minúscula y otro en letra mayúscula. Si sólo deseara ver los valores escritos con minúscula, en el filtro debe seleccionar “case sensitive” (Figura 22), o seleccionar “sp.” en la faceta de **specificEpithet**.

The screenshot shows the Refine interface with a facet for **specificEpithet**. The facet has two choices: **sp.** (101) and **SP.** (1). The table displays 101 matching rows with columns: **occurrenceID**, **specificEpithet**, **recordedBy**, **class**, **kingdom**, and **decimalLongitude**. The first few rows are:

occurrenceID	specificEpithet	recordedBy	class	kingdom	decimalLongitude
1607	sp.	Lillo, Miguel	Liliopsida	Plantae	
1608	sp.		Liliopsida	Plantae	
2422	sp.		Liliopsida	Plantae	
2555	sp.	Spegazzini, C. L.	Liliopsida	Plantae	
2556	sp.	Spegazzini, C. L.	Liliopsida	Plantae	
2557	sp.	Spegazzini, C. L.	Liliopsida	Plantae	
3159	sp.	Spegazzini, Carlos Luigi (Carlos Luis)	Magnoliopsida	Plantae	
3275	sp.		Liliopsida	Plantae	
4083	sp.	Boffa, P.	Liliopsida	Plantae	
4166	sp.	Illin, Nicolás	Liliopsida	Plantae	
5182	sp.	Jorgensen, Pedro	Magnoliopsida	Plantae	
5300	sp.	Spegazzini, Carlos Luigi (Carlos Luis)	Liliopsida	Plantae	

Figura 23.

Corrija los valores “sp.” y “SP.” utilizando la función “edit” sobre los valores en la faceta (el valor correcto debería ser nulo).

Cierre el filtro y la faceta de **specificEpithet**.

Abra una faceta de texto y un filtro para el campo **scientificName**. En el filtro, busque el valor “sp.”. Verá entonces varios valores para ese campo que incluyen “sp.”, como se muestra en la Figura 24.

The screenshot shows the Refine interface with a facet for **scientificName**. The facet has 7 choices: **Aegiphila sp.** (2), **Croton sp.** (1), **Ctenitis sp.** (1), **Muhlenbergia sp.** (18), **Nassella sp.** (1), **Phleum sp.** (22), and **Sporobolus sp.** (46). The table displays 91 matching rows with columns: **collectionCode**, **genus**, **scientificName**, **basisOfRecord**, **phylum**, **family**, **locality**, and **decimalLatitude**. The first few rows are:

collectionCode	genus	scientificName	basisOfRecord	phylum	family	locality	decimalLatitude
herb	Muhlenbergia	Muhlenbergia sp.	Preserved Specimen	Magnoliophyta	Poaceae		
herb	Muhlenbergia	Muhlenbergia sp.	Preserved Specimen	Magnoliophyta	Poaceae		
herb	Sporobolus	Sporobolus sp.	Preserved Specimen	Magnoliophyta	Poaceae		
herb	Phleum	Phleum sp.	Preserved Specimen	Magnoliophyta	Poaceae		
herb	Phleum	Phleum sp.	Preserved Specimen	Magnoliophyta	Poaceae		
herb	Phleum	Phleum sp.	Preserved Specimen	Magnoliophyta	Poaceae		
herb	Croton	Croton sp.	Preserved Specimen	Magnoliophyta	Euphorbiaceae		
herb	Muhlenbergia	Muhlenbergia sp.	Preserved Specimen	Magnoliophyta	Poaceae	Buen Orden	
herb	Sporobolus	Sporobolus sp.	Preserved Specimen	Magnoliophyta	Poaceae	Resistencia	
herb	Phleum	Phleum sp.	Preserved Specimen	Magnoliophyta	Poaceae	Río Carren-leofu	
herb	Sporobolus	Sporobolus sp.	Preserved Specimen	Magnoliophyta	Poaceae	Pampa Grande	
herb	Phleum	Phleum sp.	Preserved	Magnoliophyta	Poaceae		

Figura 24.

Debemos corregir esos nombres, sacando “sp.” y dejando solamente el nombre del género. Para no tener que hacerlo uno por uno, puede seguir los siguientes pasos.

Haga click sobre la flecha azul en **scientificName**, seleccione “Edit cells” y allí “Transform...” (Figura 25).

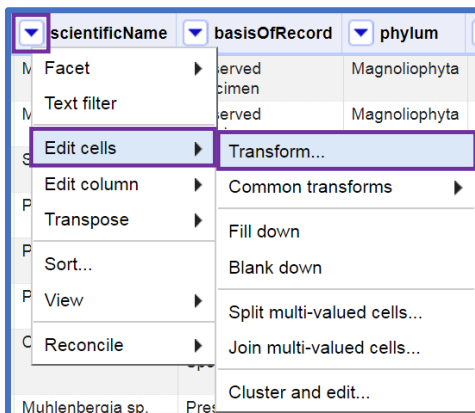


Figura 25.

Se abrirá entonces una ventana como la mostrada en la Figura 26. En el cuadro de texto, pegue la siguiente expresión:

```
value.replace(" sp.", "")
```

Dicha expresión tiene la función de reemplazar lo que está entre las primeras comillas por aquello que está entre las segundas comillas, es decir, la porción “ sp.” ([espacio]sp.) por “” ([nada]).

En la Figura 26 puede observar cómo se vería el resultado del cambio en la pestaña “Preview”.

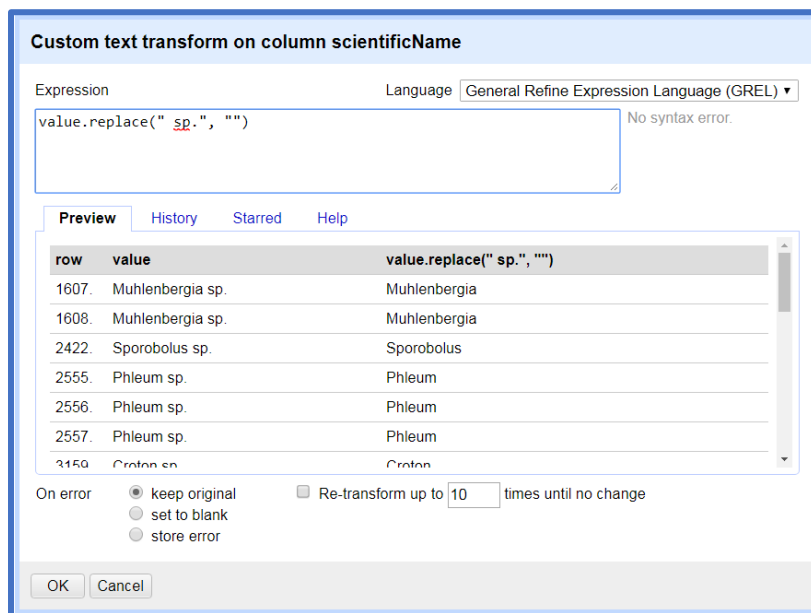


Figura 26.

Oprima “OK” para ejecutar la transformación, y verá que en el filtro ya no hay registros que contengan “sp.” como parte del valor en el campo `scientificName`.

Cierre la faceta y el filtro del campo `scientificName`.

Filtros con expresiones regulares

Abra una faceta y un filtro de texto para el campo `genus`. En el filtro (Figura 27), seleccione las opciones “case sensitive” y “regular expression” y utilice la siguiente expresión en el cuadro de texto: `^[a-z]`. Con dicha expresión se pueden buscar los valores en los que la primera letra es minúscula.

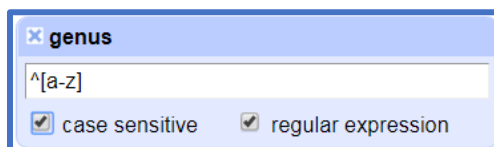


Figura 27.

Corrija los valores filtrados, dado que el género debe comenzar con mayúscula.

OpenRefine acepta un lenguaje de expresiones regulares Java, que puede consultar aquí: <http://docs.oracle.com/javase/tutorial/essential/regex/index.html>. Algunas expresiones que pueden ser útiles como filtros para diversos campos son:

- `^[A-C]`
Busca las cadenas de texto que comienzan (^) con mayúscula de la A a la C ([A-C])
- `^[^a-d]`
Busca las cadenas de texto que comienzan (^) con cualquier caracter en minúscula salvo de la a a la d ([^a-d]) – el ^ dentro del [] indica negación.
- `^\w`
Busca las cadenas de texto que comienzan (^) con una letra (\w) –de la a a la z, mayúscula o minúscula.
- `^\s`
Busca las cadenas de texto que comienzan (^) con un espacio en blanco (\s).
- `^\d`
Busca las cadenas de texto que comienzan (^) con un dígito (\d).
- `^\D`
Busca las cadenas de texto que comienzan (^) con un caracter no dígito (\D). Equivalente a la expresión con negación `^[^0-9]`.
- `\d{4}`
Busca cadenas de texto que contengan dígitos (\d), en particular 4 dígitos ({4}).
- `^\w.*\d$`
Busca las cadenas de texto que comiencen (^) con una letra (\w), sigan (.) cualquier caracter (*) y terminen (\$) con un dígito (\d).
- `^[A-Z].*\s[A-Z]`
Busca las cadenas de texto que comienzan (^) con mayúscula ([A-Z]) –cualquier mayúscula de la A a la Z– seguidas de (.) cualquier caracter (*), luego un espacio (\s), luego otra letra mayúscula ([A-Z]).

Pruebe el uso de algunas de esas expresiones en distintos campos.

Para más ejemplos y usos, puede consultar el repositorio de OpenRefine en GitHub <https://github.com/OpenRefine/OpenRefine/wiki>.

D. Uso de Agrupamientos

Agrupamientos simples

Los agrupamientos permiten, como su nombre lo indica, agrupar celdas de acuerdo a diferentes criterios. Por ejemplo, pueden agruparse celdas de acuerdo al grado de similitud de sus valores en cuanto a las letras que los componen o en cuanto a la fonética asociada. Esta función es muy útil para corregir errores de ortografía y variaciones en los datos.

Ubique el campo `stateProvince` y arme una faceta de texto para este campo.

En la ventana de la faceta, haga click en el botón “Cluster”. Se abrirá entonces una ventana como la mostrada en la Figura 28. Allí verá que algunos valores que son similares han sido agrupados por el algoritmo. La ventana también muestra el tamaño del clúster (“Cluster Size”, el número de valores agrupados), cuántos registros hay por cluster (“Row Count”) y por valor (entre paréntesis junto a los valores en “Values in Cluster”).

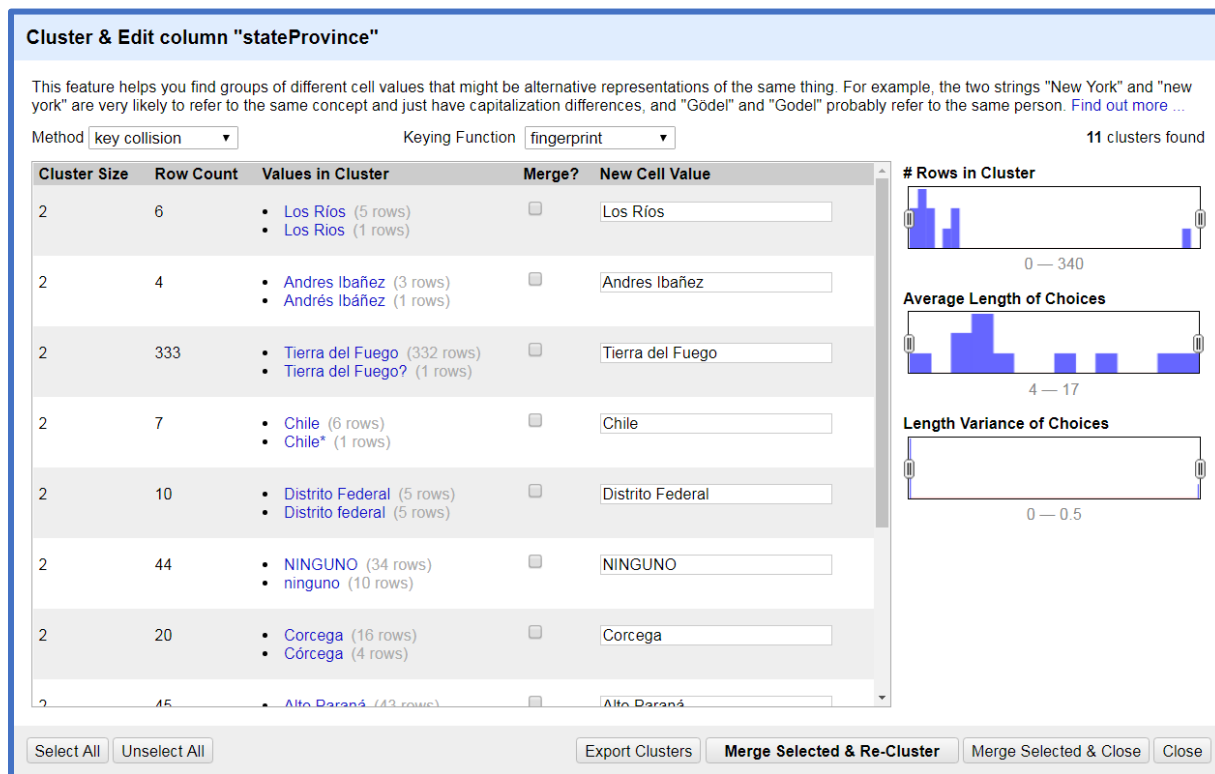


Figura 28.

Además, para cada cluster verá una opción para fusionar los valores (“Merge?”) y el nuevo valor que se asignará a todos los registros del cluster. OpenRefine asigna de forma predeterminada como nuevo valor aquel que presenta mayor número de registros asociados. Esto no necesariamente está bien, y puede modificarse haciendo click en el valor deseado (una ventana pequeña “Use” se abrirá), o editando el campo “New Cell Value”.

Corrija los valores a los que considere apropiados. Fusiones los valores de los agrupamientos (haga “merge”) haciendo click sobre “Merge Selected & Re-Cluster”.

NOTA IMPORTANTE: Cuando se agrupan valores se debe tener mucho cuidado a la hora de corregir registros. Esto es particularmente cierto para los nombres científicos, dado que variaciones en los nombres que podrían verse como aparentes errores (por ejemplo, si estamos mirando el campo epíteto específico, podemos tener dos palabras iguales con diferente terminación –um, –us), no necesariamente lo sean (por ejemplo, si uno mira también el campo género podría encontrar que esos epítetos se aplican a géneros distintos, y que ambos son válidos). Por esto, si tiene dudas, consulte los registros completos. Y si aún tiene dudas, consulte en la colección.

Una vez resueltos los agrupamientos, la ventana dirá que no se encontraron más agrupamientos utilizando el método seleccionado. Puede cambiar el método y el algoritmo que se utiliza para agrupar escogiendo entre las opciones del menú, como se muestra en la Figura 29.

Cluster & Edit column "stateProvince"

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: **nearest neighbor** (dropdown menu with options: nearest neighbor, key collision, nearest neighbor) | Distance Function: **levenshtein** (dropdown menu) | Radius: 1.0 | Block Chars: 6 | 7 clusters found

Cluster	Values in Cluster	Merge?	New Cell Value
2	13 • Mato Grosso (12 rows) • Matto Grosso (1 rows)	<input type="checkbox"/>	Mato Grosso
2	2 • Matto Grosso (1 rows) • Matto Grosso (1 rows)	<input type="checkbox"/>	Matto Grosso
2	12 • Valle de Cauca (8 rows) • Valle del Cauca (4 rows)	<input type="checkbox"/>	Valle de Cauca
2	7 • North Caroline (5 rows) • North Carolina (2 rows)	<input type="checkbox"/>	North Caroline
2	7 • Faroe Island (6 rows) • Faroe Islands (1 rows)	<input type="checkbox"/>	Faroe Island
2	2 • Ruichang (1 rows) • Raichang (1 rows)	<input type="checkbox"/>	Ruichang
2	535 • Catamarca (515 rows) • Cajamarca (20 rows)	<input type="checkbox"/>	Catamarca

Rows in Cluster: 0 — 540

Average Length of Choices: 8 — 14.5

Length Variance of Choices: 0 — 0.5

Select All | Unselect All | Export Clusters | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Figura 29.

Pruebe agrupamientos con distintos métodos para limpiar los datos.

Para conocer los detalles de cada método de agrupamiento, puede consultar el repositorio de OpenRefine en GitHub: <https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth>.

E. Nuevos campos (columnas)

Muchas veces no queremos modificar los datos directamente en los campos (columnas) en que se presentan, dado que queremos mantener los valores originales y/o queremos proveer información adicional basada en ciertos campos. Por ejemplo, podríamos tener como campos individuales el género y el epíteto específico y queremos agregar el campo nombre científico como concatenación de los dos; o viceversa: tenemos un único campo nombre científico y queremos proveer ese campo y otros dos campos para género y epíteto, a partir de la división del anterior pero sin perderlo.

Para estos casos es útil crear nuevos campos en nuevas columnas.

Manejo básico de columnas

Veamos primero algunas funciones básicas que se pueden hacer sobre los campos:

1. **Renombrar un campo.**
Hacer click en la flecha azul del campo --> Edit --> Rename this column
2. **Eliminar un campo.**
Hacer click en la flecha azul del campo --> Edit --> Remove this column
3. **Mover un campo.**
Hacer click en la flecha azul del campo --> Edit --> Move column to beginning
--> Move column to end
--> Move column left
--> Move column right

Estas tres opciones pueden verse en la Figura 30.

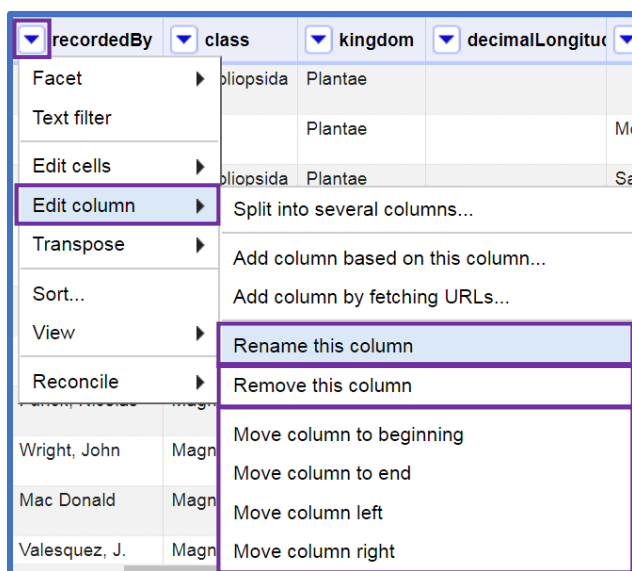


Figura 30.

4. **Reordenar o eliminar varios campos.** (Figura 31).

Hacer click en la flecha azul en campo "All" --> Edit columns --> Reorder/remove columns...

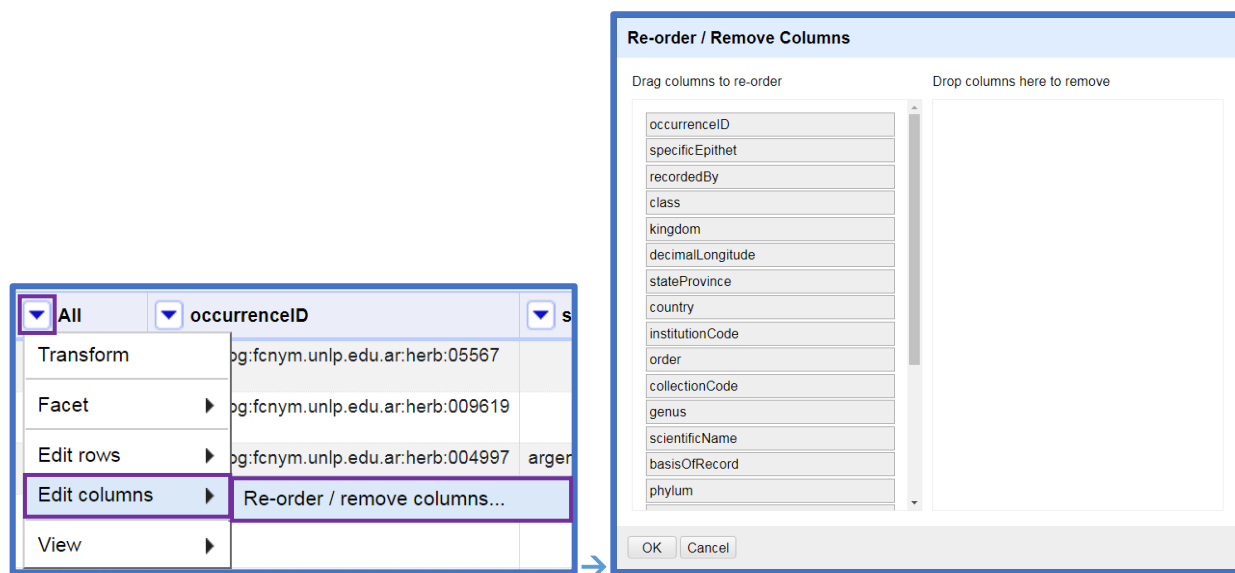


Figura 31.

En OpenRefine se considera que cualquiera de los cuatro cambios descritos anteriormente son cambios a los datos, y por ende se registran como tales en el historial hacer/deshacer (Undo/Redo).

Nuevas columnas vacías

Se pueden crear nuevos campos en base a cero, uno o más campos preexistentes.

Para crear un nuevo campo de cero, sobre cualquier columna preexistente siga la ruta: Edit column --> Add column based on this column. Se abrirá una ventana como la que se muestra en la Figura 32.

Arriba de todo, coloque el nombre del nuevo campo.

NOTA IMPORTANTE: Debe tener extremo cuidado al escoger los nombres que dará a las nuevas columnas. Considere que el nombre sea indicativo de lo que contiene (e.g., no le ponga “Columna 1” o “Transformación 3”). OpenRefine no le dejará utilizar nombres que ya sean utilizados para nombrar otros campos. Considere qué otros campos tiene en su base de datos original y no utilice nombres que ya hayan sido utilizados, se evitará así importar datos a columnas equivocadas al volver a su base de datos.

Luego, en el cuadro de texto “Expression” escriba: “null”. Ello quiere decir que se creará un campo con valores nulos. Luego oprima “OK”.

Add column based on column occurrenceID

New column name:

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

Expression: Language: General Refine Expression Language (GREL) ▼

No syntax error.

Preview History Starred Help

row	value	null
1.	urn:catalog:fcnym.unlp.edu.ar:herb:05567	null
2.	urn:catalog:fcnym.unlp.edu.ar:herb:009619	null
3.	urn:catalog:fcnym.unlp.edu.ar:herb:004997	null
4.	urn:catalog:fcnym.unlp.edu.ar:herb:002046	null
5.	urn:catalog:fcnym.unlp.edu.ar:herb:002052	null
6.	urn:catalog:fcnym.unlp.edu.ar:herb:002048	null
7.	urn:catalog:fcnym.unlp.edu.ar:herb:002059	null

OK Cancel

Figura 32.

Su nuevo campo, con el nombre que le haya dado, aparecerá a la derecha de aquel a partir del cual fue generado, y todos sus valores serán nulos.

NOTA IMPORTANTE: Tenga en cuenta que las columnas nuevas que cree en la aplicación no estarán en su base de datos original. Al importar nuevamente los datos que han sido limpiados a su base de datos, dependiendo de cómo esté estructurada esa base de datos, puede que estas nuevas columnas no sean importadas o que reciba un mensaje de error de importación porque el número de campos del archivo no coincide con el de la base de datos. En estos casos, debe asegurarse de agregar los nuevos campos en su base de datos si desea importar todos los campos nuevos.

Nuevas columnas a partir transformaciones simples de otras columnas

Veamos ahora como crear nuevas columnas con datos modificados a partir de columnas preexistentes.

- i. **Concatenaciones.** Si desea crear un campo que sea la concatenación de otros dos campos separados, siga la siguiente ruta:
Utilizaremos como ejemplo la concatenación de los campos **genus** y **specificEpithet**.

Click en la flecha azul del campo **genus** --> Edit column --> Add column based on this column.

Se abrirá una nueva ventana. Puede llamar al nuevo campo “concat_scientificName”, para indicar que se trata de la concatenación (note que ya hay un campo **scientificName** en los datos).

En el cuadro de texto, pegue la siguiente expresión:

Expresión ejemplo: `cells["genus"].value + " " + cells["specificEpithet"].value`
 Expresión general: `cells["campo1"].value + " " + cells["campo2"].value`

La expresión ejemplo concatena (+) los valores del campo **genus** (`cells["genus"].value`) y los del campo **specificEpithet** (`cells["specificEpithet"].value`), con un espacio entre los valores (" ").

Note que algunos de los valores de los campos **genus** y **specificEpithet** son nulos ("null", y no "blank"), y en la ventana se mostrará una lista de errores (Figura 33). Esto se debe a que no se puede operar sobre valores nulos.

Add column based on column genus

New column name

concat_scientificName

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

Expression

cells[genus].value + cells[specificEpithet].value

No syntax error.

Language

General Refine Expression Language (GREL) ▼

Preview

History

Starred

Help

row	value	cells[genus].value + cells[specificEpithet].value
1.	null	Error: Cannot retrieve field from null
2.	null	Error: Cannot retrieve field from null
3.	null	Error: Cannot retrieve field from null
4.	Filago	Error: Cannot retrieve field from null
5.	Flotovia	Error: Cannot retrieve field from null
6.	Flotovia	Error: Cannot retrieve field from null
7.	Galinsona	Error: Cannot retrieve field from null

OK

Cancel

Figura 33.

En este caso, puede sortear el problema utilizando en cambio la siguiente expresión:

```
if(isBlank(cells["genus"].value), "", cells["genus"].value) + " " +
if(isBlank(cells["specificEpithet"].value), "", cells["specificEpithet"].value)
```

Lo que dicha expresión significa es: concatenar (+) dos partes, cada una proviene de una sub-expresión "if", separadas por un espacio (+ " " +). Cada una de estas sub-expresiones indica: si (if) el valor del campo dado es nulo (isBlank(cells["genus"].value)), colocar un blanco (""), si no (,), colocar el valor del campo (cells["genus"].value). La otra sub-expresión es lo mismo pero para epíteto específico.

NOTA: Para evitarse de modo más general este problema de celdas nulas, cuando importa el conjunto de datos para crear su proyecto al principio del proceso, puede asegurarse de NO seleccionar la opción "Store blank cells as nulls" (ver Figura 2).

- ii. **Divisiones.** Si desea crear campos separados a partir de los valores en un único campo, siga la siguiente ruta:
Utilizaremos como ejemplo la división del campo `eventDate` para agregar tres campos: año, mes y día (`year`, `month` y `day`)

Click en la flecha azul del campo `eventDate` --> Edit column --> Split into several columns.

Se abrirá una nueva ventana (Figura 34). Allí debe escoger si se dividirá por separador o por longitud de caracteres, y en el primer caso qué tipo de separador se utilizará (puede ser espacio –tab–, coma, punto y coma, guión, etc.).

En este caso, si exploramos los datos del campo original veremos que año, mes y día están separados por guiones, de modo que elegiremos el guión como separador.

IMPORTANTE: Deseleccione la opción “Remove this column” a la derecha. Si la deja seleccionada, perderá el campo original y sólo tendrá los tres nuevos campos.

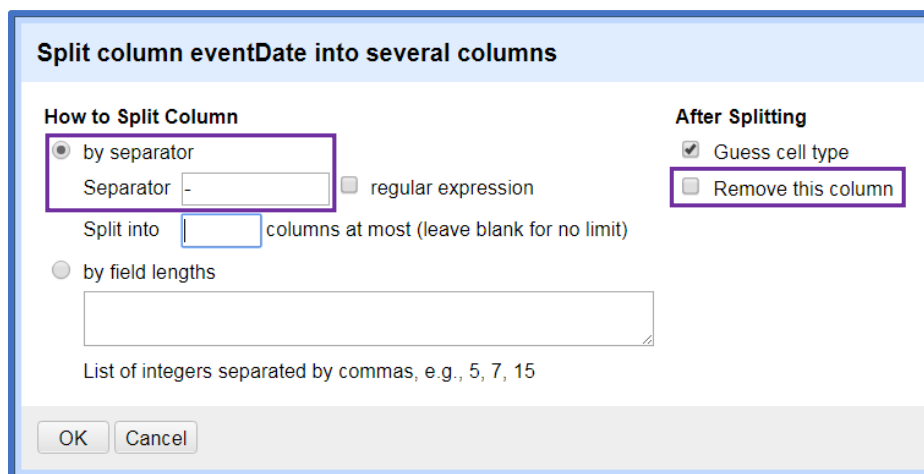


Figura 34.

Una vez que oprima OK, se crearán las nuevas columnas a la derecha del campo **eventDate**. OpenRefine las nombra automáticamente agregando números al final del nombre (en este caso: eventDate1, eventDate2 y eventDate3). Cambie los nombres de las columnas por los que corresponda (flecha azul-->Edit column --> Rename this column).

NOTA IMPORTANTE: Cuando efectúe este tipo de divisiones de campos utilizando como criterio o bien separadores o bien longitud de caracteres, asegúrese de que en el campo original no haya distintos formatos para diferentes registros. Vea el siguiente ejemplo:

Se quiere separar un campo nombrado “coordenadas” que contiene datos de latitud y longitud separados por coma, del tipo: “-32.04588990, -54.98789901”, para obtener dos campos distintos, latitud y longitud.

Si todos los campos tienen el mismo formato, obtendrá dos campos nuevos de la siguiente forma:

campo 1: -32.04588990

campo 2: -54.98789901

En cambio, si en algún registro los valores dentro del campo coordenadas no están en formato decimal, entonces tendrá problemas al dividir el campo. Suponga como ejemplo que uno o más registros tienen valores con formato “34° 20’ 15,2” S, 54° 49’ 13” O”. En ese caso, la separación le dará 3 campos en vez de dos, con la latitud incorrectamente separada:

campo 1: 34° 20’ 15

campo 2: 2” S

campo 3: 54° 49’ 13” O

F. Marcado de registros: banderas y estrellas

OpenRefine ofrece la opción de marcar los distintos registros con banderas (flags) y/o estrellas (stars). Esta opción es a veces muy útil para reconocer registros o grupos de registros rápidamente.

Las banderas y estrellas NO forman parte de los datos. Son solamente una herramienta que facilita el trabajo dentro del programa. Por ello, cuando exporte los datos NO verá las columnas que corresponden a estas funciones. Es decir, si usted marcó algún registro con una bandera, por ejemplo, no verá esa bandera ni ninguna otra marca indicadora de su existencia en los datos exportados.

Marcado con banderas y estrellas

Las banderas y estrellas se encuentran dentro del campo “All”. Para marcar un registro con una bandera o estrella, simplemente haga click sobre el ícono correspondiente en ese registro (que se pondrá de color amarillo).

Para desmarcar el registro, haga click nuevamente sobre el ícono (que volverá a su color blanco original).

También puede marcar o desmarcar conjuntos de varios registros.

Para ello escoja algún criterio que los agrupe. Por ejemplo, si quiere marcar todos los registros del género Acacia, arme una faceta sobre el campo **genus** (haga click sobre la flecha azul del campo --> Facet --> Text facet).

En la faceta, seleccione el valor “Acacia” haciendo click en el valor (verá que en la ventana principal sólo se mostrarán esos registros).

Para marcar todos esos registros con una bandera, haga click en la flecha azul del campo “All” --> Edit rows --> Flag rows, como se muestra en la Figura 35.

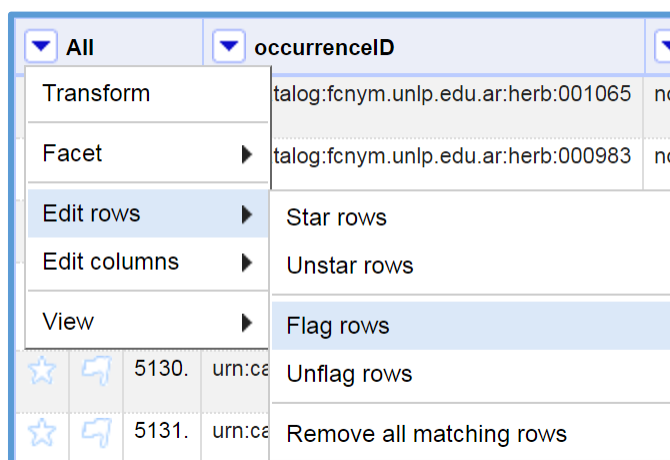


Figura 35.

Una vez que lo haga, verá que todos los registros seleccionados tienen ahora marcada una bandera.

Para desmarcar todos esos registros, puede seguir la ruta: haga click en la flecha azul del campo "All" --> Edit rows --> Unflag rows.

Para poner y sacar estrellas, siga el mismo procedimiento con "stars" en lugar de "flags".

Conservación de banderas y estrellas en la exportación

Si desea marcar los registros de modo que al exportar se vean las marcas, deberá crear un nuevo campo que capture esa información. Puede por ejemplo hacer lo siguiente:

Cree un nuevo campo: sobre cualquier campo haga click en la flecha azul --> Edit column --> Add column based on this column.

Se abrirá una ventana como la mostrada en la Figura 36. Asigne un nombre al campo. Por ejemplo, si sus banderas significan que ha detectado errores en los registros, puede llamarlo "hasError".

En el cuadro de texto pegue la siguiente expresión:

```
if(row.flagged, "yes", "no")
```

Esta expresión hará que el campo nuevo tenga como valor "yes" si usted ha asignado una bandera al registro y "no" si no ha asignado una bandera.

Al apretar "OK" su campo se habrá creado. Verifique los valores que toma asignando a algunos registros una bandera.

Add column based on column occurrenceID

New column name:

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

Expression: Language: General Refine Expression Language (GREL) ▼

No syntax error.

Preview History Starred Help

row	value	if(row.starred, "yes", "no")
1.	urn:catalog:fcnym.unlp.edu.ar:herb:05567	no
2.	urn:catalog:fcnym.unlp.edu.ar:herb:009619	no
3.	urn:catalog:fcnym.unlp.edu.ar:herb:004997	no
4.	urn:catalog:fcnym.unlp.edu.ar:herb:002046	no
5.	urn:catalog:fcnym.unlp.edu.ar:herb:002052	no
6.	urn:catalog:fcnym.unlp.edu.ar:herb:002048	no
7.	urn:catalog:fcnym.unlp.edu.ar:herb:002059	no

OK Cancel

Figura 36.

Puede repetir el proceso creando otro campo para las estrellas, usando la expresión:

```
if(row.starred, "yes", "no")
```


Para ver los pasos de exportación de datos, vea la sección de Exportación de proyectos.

Uso de banderas y estrellas para eliminar registros

Las banderas y estrellas se pueden utilizar para eliminar grupos de registros. Para ello, siga los siguientes pasos:

1. Marque con una bandera (o estrella) los registros deseados. Puede hacerlo uno por uno o en grupos a través del marcado dentro de facetas (ver más arriba).
2. Cree una faceta para la bandera. Haga click en la flecha azul sobre el campo "All" --> Facet --> Facet by flag (Figura 37a).
3. En esta nueva faceta, a la izquierda, seleccione la opción "true" haciendo click. Ello le mostrará los registros a los que se ha asignado una bandera.
4. Haga click nuevamente sobre la flecha azul del campo "All" --> Edit rows --> Remove all matching rows (Figura 37b).

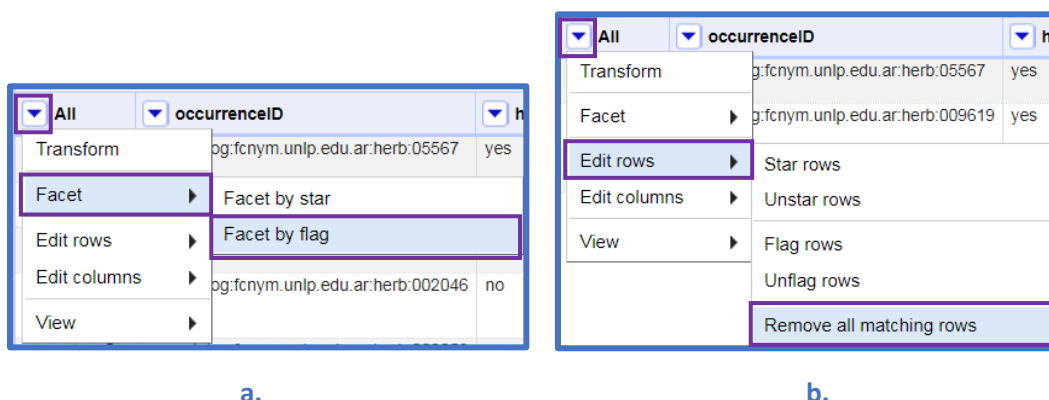


Figura 37.

De esta forma habrá eliminado todos los registros que fueron marcados con una bandera.

3. Guardado y exportación de datos y proyectos

Debe tener en cuenta que lo que guarda al usar el programa es el proyecto, y que ello no implica en ningún caso que los cambios que realice se reflejen automáticamente en su base de datos original. Para ello, deberá exportar los datos desde OpenRefine e importarlos nuevamente a su base de datos.

A. Guardado de datos y proyectos

Los proyectos con los que trabaja usando OpenRefine son guardados en su propia computadora de forma automática. En otras palabras, no existe un botón o un comando "Guardar".

Los directorios en que se guardan los proyectos se listan a continuación:

Windows: dependiendo de la versión de Windows que utilice, los datos se encontrarán en uno de estos directorios:

- C:\Documents and Settings\{user id}\Local Settings\Application Data\OpenRefine
- C:\Users\{user id}\AppData\Roaming\OpenRefine
- C:\Users\{user id}\AppData\Local\OpenRefine
- C:\Users\{user id}\OpenRefine

MacOSX:

- ~/Library/Application Support/OpenRefine/
- ~/Library/Application Support/Google/Refine/ (versions de Google Refine más antiguas)
- Ingreso a través de /var/log/daemon.log - grep para com.google.refine.Refine

Linux:

- ~/.local/share/openrefine/

B. Exportación de datos y proyectos

OpenRefine ofrece varias opciones para exportar los datos y proyectos. Se puede acceder a estas opciones en la esquina superior derecha de la ventana del programa, haciendo click en el botón “Export” (Figura 38).

Note que la primera opción, “Export project”, permite exportar el proyecto, mientras que las otras permiten exportar los datos.

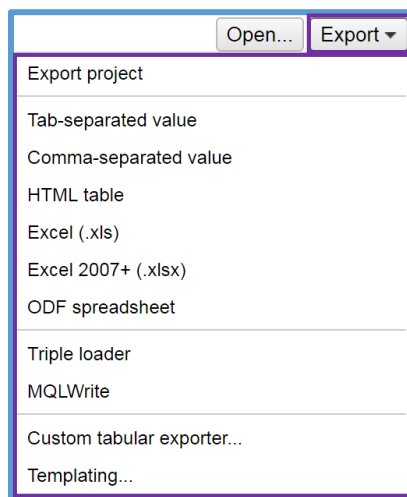


Figura 38.

La exportación de proyectos es útil cuando uno quiere abrir el mismo proyecto en OpenRefine en otra computadora. En este caso se descarga un archivo .GZIP, que sólo puede ser abierto por el programa (no se descarga un archivo de datos que pueda abrir en un procesador de textos ni en una planilla de cálculo).

Para exportar los datos y poder abrirlos en otro programa, puede seguir cualquiera de las otras opciones, que resultarán en un archivo con uno de los formatos disponibles.

NOTA IMPORTANTE: La exportación se realizará teniendo en cuenta las facetas y filtros aplicados. Esto implica que si usted tiene abierta por ejemplo una faceta, sólo los datos correspondientes a dicha faceta serán exportados. Por lo tanto, para exportar todos los datos, recuerde cerrar todos los filtros y facetas antes de hacer la exportación.

Para una exportación más personalizada, en “Export” escoja “Custom tabular exporter...”. Se abrirá una ventana como la mostrada en la Figura 39.

The screenshot shows the 'Custom Tabular Exporter' window with the 'Content' tab selected. On the left, under 'Select and Order Columns to Export', a list of columns is shown with checkboxes: occurrenceID, hasError, specificEpithet, recordedBy, eventDate, class, kingdom, decimalLongitude, and stateProvince. Below this list are 'Select All' and 'De-select All' buttons. On the right, 'Options for occurrenceID' are displayed, including radio buttons for 'Matched entity's name', 'Matched entity's ID', and 'Cell's content', as well as checkboxes for 'Link to matched entity's page' and 'Output nothing for unmatched cells'. There are also options for date/time formats (ISO 8601, Short, Long, Medium, Full locale format, Custom) and checkboxes for 'Use local time zone' and 'Omit time'. At the bottom, there are checkboxes for 'Output column headers', 'Output empty rows (ie all cells null)', and 'Ignore facets and filters and export all rows'. A 'Cancel' button is at the bottom left.

Figura 39.

En la pestaña “Content” puede elegir qué campos exportar y modificar ciertos parámetros para cada campo individualmente (como los formatos o los valores reconciliados –para esto último vea la sección de Reconciliación).

Observe que en esta pestaña también puede escoger ignorar todas las facetas y filtros al exportar, lo cual es muy útil en caso de que haya olvidado cerrar alguna.

Para descargar los datos, vaya a la pestaña “Download”, como se ve en la Figura 40.

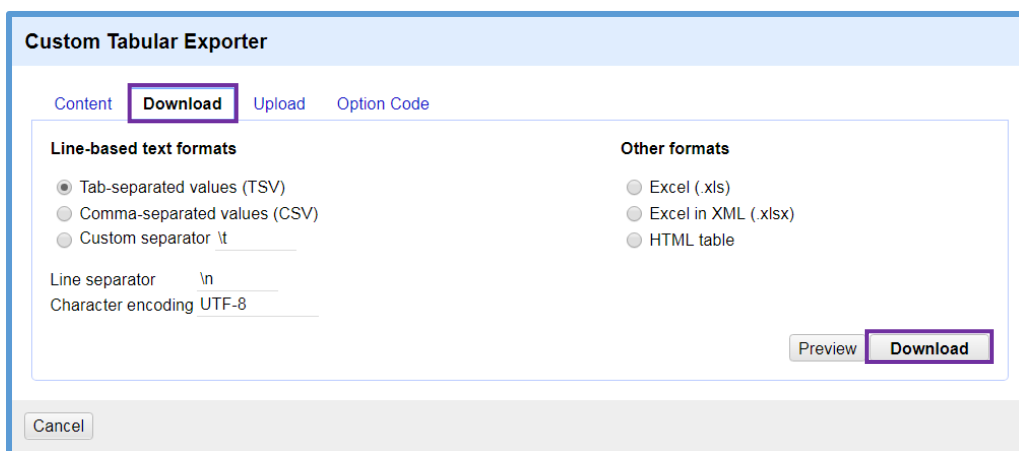


Figura 40.

En esta pestaña puede seleccionar el formato de los datos para la descarga. Escoja el que prefiera y haga click en “Download”. Inmediatamente comenzará la descarga de los datos.

También, para ver una vista previa de los datos que descargará, puede hacer click en “Preview”, y se abrirá otra ventana en su navegador web donde podrá ver una muestra de los datos a descargar.

4. Consultas a servicios externos

OpenRefine ofrece la posibilidad de consultar fuentes externas, una función que es muy útil cuando se intenta mejorar la calidad de los datos. Permite, por ejemplo, validar nombres taxonómicos y geográficos contra fuentes de información que se consideren confiables, completar rangos taxonómicos y campos geográficos superiores, georreferenciar, incorporar enlaces a imágenes almacenadas en sitios web, entre otros.

En OpenRefine las consultas externas pueden realizarse por dos vías: a través de URLs, o a través de servicios de reconciliación.

NOTA: Debe recordarse que para poder realizar consultas a servicios que se encuentran en línea se requiere conexión a internet.

NOTA: La velocidad a la que se obtienen los resultados de las consultas depende de la velocidad de respuesta del servicio en particular. De esta forma, si se quieren comparar muchos registros, el tiempo de la operación será prolongado. Para acortar tiempos, se pueden hacer comparaciones de registros contra el servicio deseado dentro de una faceta, es decir, en una fracción particular de los registros.

A. Consultas externas a través de URLs

Nos referimos a consultas a través de URLs cuando el proceso implica proveer a OpenRefine con la dirección web (URL) de un determinado servicio y ciertos parámetros mínimos para obtener de dicho servicio un resultado.

Resolución de nombres científicos usando Global Names Resolver

En el ejemplo siguiente, compararemos los nombres científicos (contenidos en el campo **scientificName**) contra el servicio Global Names Resolver (<http://resolver.globalnames.org>, de aquí en más GNR).

Para acortar tiempos, cree una faceta para el campo **genus** y dentro de ella escoja el género *Cinna*. En el conjunto de datos utilizado *Cinna* parece tener 3 especies asociadas: *C. lateralis* (1 registro), *C. arundinacea* (6 registros) y *C. latifolia* (3 registros).

Proceda a crear la faceta (en el campo **genus**, haga click en la flecha azul --> Facet --> Text facet) y escoger el género indicado.

Para comparar los nombres contra el GNR, haremos un llamado al servicio y capturaremos los resultados en un nuevo campo:

A partir del campo **scientificName**, cree una nueva columna a partir de una dirección URL (como se muestra en la Figura 41) siguiendo la ruta: haga click en la flecha azul del campo --> Edit column --> Add column by fetching URLs...

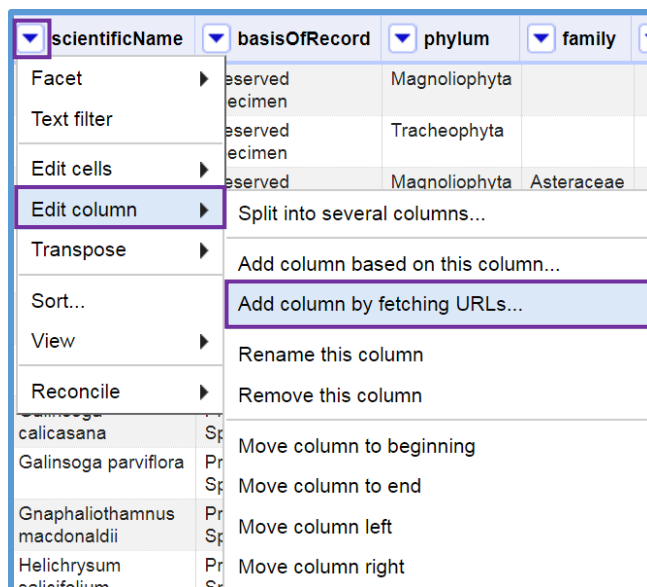


Figura 41.

Se abrirá una ventana como la mostrada en la Figura 42. Allí, dé un nombre al nuevo campo (por ejemplo, GNR_json_sciName), y en el cuadro de texto coloque la siguiente expresión:

"http://resolver.globalnames.org/name_resolvers.json?names="+escape(cells["scientificName"].value,"url")

Dicha expresión indica se hará una consulta en el GNR utilizando como valores de comparación aquellos que se encuentran en el campo **scientificName**.

Add column by fetching URLs based on column scientificName

New column name: Throttle delay: milliseconds

On error: ☒ set to blank ☐ store error ☒ Cache responses

Formulate the URLs to fetch:

Expression:

Una vez que haya creado el nuevo campo con la expresión general, verá que contiene, en formato JSON, los resultados de la consulta en GNR para cada nombre, con todos los parámetros y valores que GNR reporta.

Para poder trabajar con esto más cómodamente, debemos extraer de allí los valores de interés.

Dado que GNR consulta varias fuentes de nombres taxonómicos, nos interesa saber cuál es el nombre científico que figura en cada fuente. Algunas fuentes pueden tener listado el nombre pero considerarlo inválido y proveer el nombre correcto. Entonces, extraeremos del resultado en JSON, en un nuevo campo, los siguientes valores:

- Fuente consultada: "data_source_title"
- Nombre encontrado en la fuente: "name_string"
- Nombre aceptado por la fuente: "current_name_string"

Para ello, a partir del campo en Json (en el ejemplo, GNR_json_sciName), cree un nuevo campo siguiendo: haga click en la flecha azul --> Edit column --> Add column based on this column (Figura 43).

Dé un nombre al nuevo campo (por ejemplo, GNR_sciName_options) y en el cuadro de texto, coloque la siguiente expresión:

```
forEach(value.parseJson().get("data")[0].get("results"),v,v.get("data_source_title") + ";" +  
v.get("name_string") + ";" + if(isBlank(v.get("current_name_string")), "",  
v.get("current_name_string"))).join(" | ")
```

Dicha expresión analiza la cadena en formato Json, que tiene dentro de su estructura secciones “data” y dentro de esta “results” –un “result” proveniente de cada fuente consultada (por ejemplo, un “result” de Catalogue of Life). Dentro de cada sección “results” extrae los valores de interés (“data_source_title”, “name_string” y “current_name_string”) y los separa con un “;”. Como no todas las fuentes proveen un nombre aceptado (“current_name_string”), la expresión “if” especifica que si ese parámetro es nulo debe dejarse el espacio vacío (“”), y si no, colocar el valor extraído. Por último, une los grupos de valores extraídos en una única cadena de texto, separados por un “ | ”.

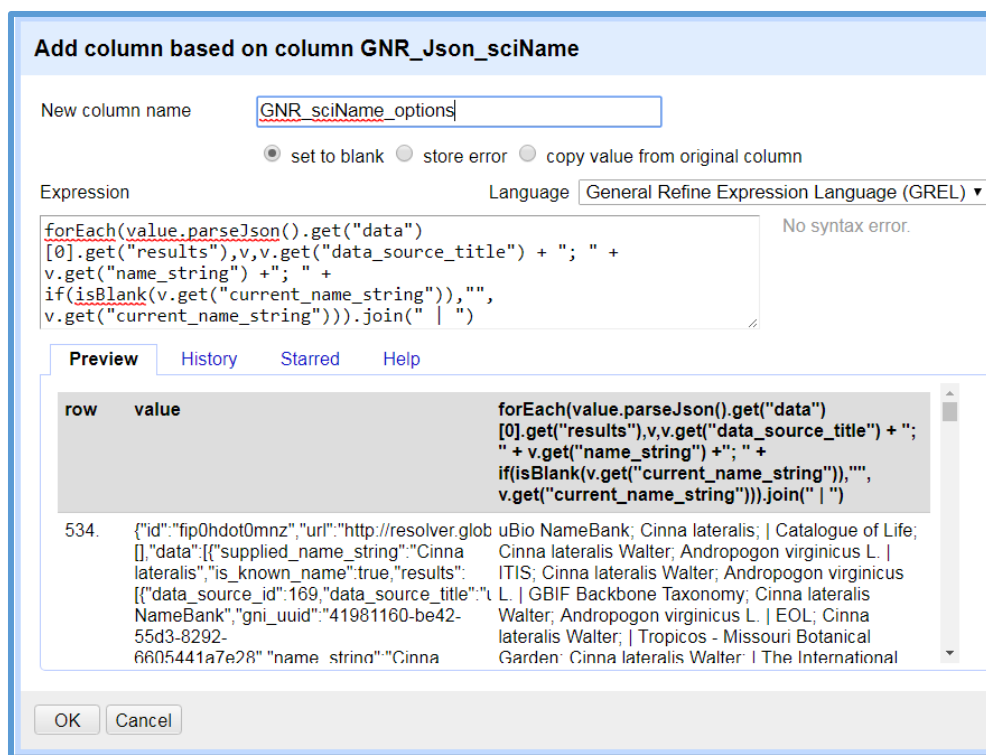


Figura 43.

Una vez que haya creado el campo, verá que contiene, aun en formato JSON, los valores de interés extraídos de GNR. Por ejemplo:

uBio NameBank; Cinna lateralis; | Catalogue of Life; Cinna lateralis Walter; Andropogon virginicus L. | ITIS; Cinna lateralis Walter; Andropogon virginicus L. | GBIF Backbone Taxonomy; Cinna lateralis Walter; Andropogon virginicus L. | EOL; Cinna lateralis Walter; | Tropicos - Missouri Botanical Garden; Cinna lateralis Walter; | The International Plant Names Index; Cinna lateralis Walter; | uBio NameBank; Cinna lateralis Walter; | uBio NameBank; Cinna lateralis Walter, 1788; | Arctos; Cinna lateralis Walter;

Note que algunas fuentes encuentran el nombre pero no proveen un nombre aceptado, por ejemplo:

uBio NameBank; Cinna lateralis;

no tiene un valor en el tercer lugar, mientras que:

Catalogue of Life; Cinna lateralis Walter; Andropogon virginicus L.

provee el nombre encontrado y el nombre válido.

Note además que algunas fuentes tienen más de una variante asociada al nombre, por ejemplo:

uBio NameBank; Cinna lateralis;

uBio NameBank; Cinna lateralis Walter;

uBio NameBank; Cinna lateralis Walter, 1788;

NOTA IMPORTANTE: No todos los nombres serán necesariamente encontrados en todas las fuentes consultadas, por lo que el número de fuentes variará de un nombre al otro. En consecuencia, la ubicación de las fuentes en la cadena de texto no será homogénea de un registro al otro. Una consecuencia de esto es que si usted quiere luego separar el contenido en campos distintos de acuerdo a la fuente consultada (e.g., un campo para ITIS, uno para Catalogue of Life, etc.), no podrá hacerlo de modo que cada nuevo campo tenga los datos de una misma y única fuente.

En este caso, le conviene en cambio hacer varios llamados a GNR separados, cada uno especificando una fuente determinada. Como se menciona más arriba, si quiere por ejemplo sólo consultar los valores dados por Catalogue of Life, use la expresión siguiente:

```
"http://resolver.globalnames.org/name_resolvers.json?names="+escape(cells["scientificName"].value,"url")+"&data_source_ids=1"
```

Georreferenciación usando GeoLocate

En este ejemplo, para facilitar la explicación y reducir el tiempo de consulta al servicio, construiremos previamente dos facetas. La primera sobre el campo **country**, dentro de la cual seleccionaremos el valor "Argentina". La segunda faceta será sobre el campo **genus**, dentro de la cual seleccionaremos el valor "Acacia". Una vez aplicadas ambas facetas y escogidos los valores, verá que en la ventana principal sólo se muestra un subconjunto de registros que cumplen estas condiciones.

Llevaremos a cabo la georreferenciación a partir del campo **locality**. Para ello, cree un nuevo campo a partir de éste siguiendo: haga click en la flecha azul --> Edit column --> Add column by fetching URLs...

Se abrirá una nueva ventana. Allí dé un nombre al nuevo campo, por ejemplo "GeoLocate_Json_georref", y pegue en el cuadro de texto la siguiente expresión:

```
"http://www.museum.tulane.edu/webservices/geolocatesvcv2/glcwrap.aspx?Country=Argentina&fmt=json&Locality="+escape(value,'url')
```

En esta expresión, "fmt" indica el formato en el que el resultado será devuelto por el servicio. GeoLocate ofrece dos posibles formatos, json y geojson.

Una vez que haya creado el nuevo campo con la expresión general, verá que contiene, en formato JSON, los resultados de la consulta en GeoLocate para cada localidad, con todos los parámetros y valores que este servicio reporta.

En los resultados puede tener varios casos:

Caso 1) Ningún resultado encontrado. Ello quiere decir que GeoLocate no ha podido ubicar la localidad de interés. En la celda correspondiente verá lo siguiente:

```
{ "engineVersion" : "GLC:5.21|U:1.01374|eng:1.0", "numResults" : 0, "executionTimems" : 171.6003 }
```

Caso 2) Un único resultado encontrado. En la celda correspondiente verá, por ejemplo, lo siguiente:

```
{ "engineVersion" : "GLC:5.21|U:1.01374|eng:1.0", "numResults" : 1, "executionTimems" : 171.6003,
  "resultSet" : { "type": "FeatureCollection", "features": [ { "type": "Feature", "geometry": { "type": "Point",
    "coordinates": [-64.471941, -23.643418] }, "properties": { "parsePattern" : "YUTO", "precision" : "High",
    "score" : 79, "uncertaintyRadiusMeters" : 3036, "uncertaintyPolygon" : "Unavailable",
    "displacedDistanceMiles" : 0, "displacedHeadingDegrees" : 0, "debug" :
    ":GazPartMatch=False|:inAdm=True|:Adm=JUJUY|:NPExtent=5040|:NP=YUTO|:KFID=|YUTO" } } ], "crs":
  { "type" : "EPSG", "properties" : { "code" : 4326 } } }
```

Allí puede observar varios parámetros de interés:

- Las coordenadas: "coordinates": [-64.471941, -23.643418]
- Las localidad original que consultó: "parsePattern" : "YUTO"
- El radio de incerteza en metros: "uncertaintyRadiusMeters" : 3036
- El polígono de incerteza asociado: "uncertaintyPolygon" : "Unavailable", en este caso no disponible.
- Los desplazamientos: distancia en millas y grados en una dirección: "displacedDistanceMiles" : 0, "displacedHeadingDegrees" : 0, en este caso con valores "0" porque no se especifica desplazamiento de ningún tipo en la localidad (e.g., 45km de Yuto, o 45km N Yuto).
- La correspondencia en el gacetero consultado: GazPartMatch, y en éste la división administrativa bajo la cual se encontró la localidad: |:Adm=JUJUY|.

Caso 3) Varios resultados encontrados para un mismo valor de localidad. Esto sucede comúnmente cuando no se especifican en la consulta niveles administrativos por debajo de país (e.g., podría haber en un mismo país varios lugares con el mismo nombre). Un ejemplo sería:

```
{ "engineVersion" : "GLC:5.21|U:1.01374|eng:1.0", "numResults" : 3, "executionTimems" : 187.2004,
  "resultSet" : { "type": "FeatureCollection", "features": [
    { "type": "Feature", "geometry": { "type": "Point", "coordinates": [-64.158097, -26.21252] },
    "properties": { "parsePattern" : "TARTAGAL", "precision" : "High", "score" : 83, "uncertaintyRadiusMeters" : 301,
    "uncertaintyPolygon" : "Unavailable", "displacedDistanceMiles" : 0, "displacedHeadingDegrees" : 0,
    "debug" : ":GazPartMatch=False|:inAdm=True|:Adm=SANTIAGO DEL ESTERO|:NPExtent=500|:NP=TARTAGAL|:KFID=|TARTAGAL" } },
    ...
  ] }
```

```
{ "type": "Feature", "geometry": { "type": "Point", "coordinates": [-59.846115, -28.671732] },
"properties": { "parsePattern": "TARTAGAL", "precision": "High", "score": 83, "uncertaintyRadiusMeters":
: 3036, "uncertaintyPolygon": "Unavailable", "displacedDistanceMiles": 0, "displacedHeadingDegrees":
0,
"debug"
:
":GazPartMatch=False|:inAdm=True|:Adm=SANTA
FE|:NPExtent=5040|:NP=TARTAGAL|:KFID=|TARTAGAL" } },
```

```
{ "type": "Feature", "geometry": { "type": "Point", "coordinates": [-63.801314, -22.516365] },
"properties": { "parsePattern": "TARTAGAL", "precision": "High", "score": 83, "uncertaintyRadiusMeters":
: 3036, "uncertaintyPolygon": "Unavailable", "displacedDistanceMiles": 0, "displacedHeadingDegrees":
0,
"debug"
:
":GazPartMatch=False|:inAdm=True|:Adm=SALTA|:NPExtent=5040|:NP=TARTAGAL|:KFID=|TARTAGAL"
} }
```

```
], "crs": { "type": "EPSG", "properties": { "code": 4326 } } }
```

Note que los tres resultados del ejemplo corresponden a provincias distintas en las que se encuentra una localidad “Tartagal”, puede comparar las coordenadas para cada una.

NOTA: Para visualizar la estructura de los resultados en JSON de modo más amigable, puede probar copiando el resultado de alguna celda en un analizador de JSON en línea. Existen muchas opciones, una de ellas es <http://json.parser.online.fr/>. Allí, seleccionando distintas opciones arriba a la derecha podrá distinguir mejor la estructura, cuáles son los objetos, los arreglos y las cadenas de texto y cómo están relacionados unos con otros (Figura 42). Esto puede ser muy útil a la hora de armar expresiones para desglosar el contenido de los campos en nuevos campos sin perder información.

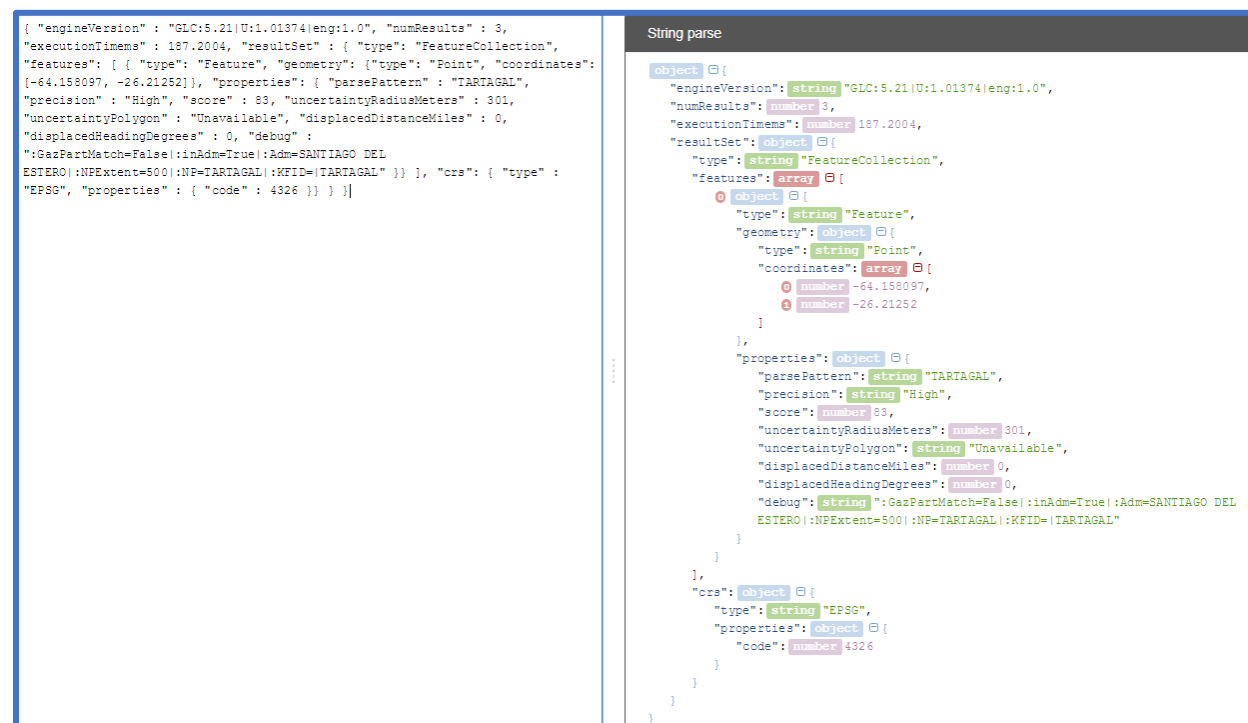


Figura 42.

NOTA: La expresión utilizada es muy simple y sólo le pide al servicio que resuelva la georreferenciación en base al campo localidad y teniendo como valor fijo “Argentina” para el campo país, pero sin especificar los valores de otros campos geográficos. Sin embargo, todos los campos se pueden incluir en la expresión para obtener resultados más específicos. Ello puede hacerse de dos maneras:

a. Establecer los valores de los campos como valores fijos, como hicimos con el país, agregando luego por ejemplo: `&state=VALOR`

donde VALOR es el valor fijo que uno establece (e.g., Córdoba). Esto restringirá los resultados en función de esos parámetros.

b. Incluir los campos como valores a consultar, en cuyo caso para cada campo hay que incluir como valor: `+escape(cells.NOMBREDELCAMPO.value,'url')+”`

La expresión con todos los campos se verá entonces como:

```
"http://www.museum.tulane.edu/webservices/geolocate/v2/glcwrap.aspx?country=Argentina&state="+escape(cells.stateProvince.value,'url')+ "&locality="+escape(cells.locality.value,'url')
```

Note que el nombre del campo será el que tiene en su base de datos. Note también que en la base de datos dada para este ejercicio no hay un campo correspondiente a **county**, pero GeoLocate permite incluirlo si lo hubiera.

Para poder trabajar con estos resultados más cómodamente, debemos extraer de allí los valores de interés. En este paso debe tener cuidado. Debido a que no especificamos todos los campos geográficos en la consulta a GeoLocate, recuerde que los registros pueden tener más de un resultado posible, y que cada resultado tiene sus propios parámetros de georreferenciación.

A modo de ejemplo, extraeremos en nuevos campos los valores de las coordenadas. (El conjunto de datos provisto para realizar los ejercicios de esta guía contiene campos originales de latitud y longitud provistos por la fuente, puede utilizarlos para contrastar los resultados obtenidos utilizando GeoLocate).

Para extraer las coordenadas puede seguir dos métodos: 1) extraer latitud y longitud conjuntamente y luego separar; o 2) extraer latitud y longitud de modo independiente.

Método 1: extraer latitud y longitud conjuntamente

Haga click en la flecha azul del campo GeoLocate_Json_georref --> Edit column --> Add column based on this column.

De un nombre al nuevo campo, por ejemplo, GeoLocate_parseCoord, y en el cuadro de texto pegue la siguiente expresión:

```
forEach(filter(value.parseJson().resultSet.features,v,isNotBlank(v.geometry)),w,w.geometry.coordinates
.join("; ")).join("|")
```

Esta expresión es un poco más compleja que las que hemos estado utilizando, debido a que se requiere extraer información de una estructura Json particular Objeto --> Arreglo --> Objeto --> Arreglo. (Puede visualizar la estructura en Json como se menciona en la nota de la Figura 42).

El nuevo campo tendrá valores como los siguientes, por ejemplo, para un registro cuya consulta devolvió tres resultados:

```
-64.158097; -26.21252|-59.846115; -28.671732|-63.801314; -22.516365
```

IMPORTANTE: Note que GeoLocate provee como primer valor de coordenadas la longitud y como segundo valor la latitud.

Dividiremos ahora este campo en tres partes, una para cada resultado:

Haga click en la flecha azul del campo --> Edit column --> Split into several columns.

Escoja como separador "|". Deseleccione la opción "Remove this column" si quiere mantener el campo original (esto es recomendable, siempre puede eliminar los campos después).

Tendrá entonces ahora una serie de campos con valores del tipo: -64.158097; -26.21252. Sobre cada uno, puede realizar una nueva separación utilizando como separador ";".

Método 2: extraer latitud y longitud independientemente

Haga click en la flecha azul del campo GeoLocate_Json_georref --> Edit column --> Add column based on this column.

De un nombre al nuevo campo, por ejemplo, GeoLocate_parseLong, y en el cuadro de texto pegue la siguiente expresión:

```
forEach(filter(value.parseJson().resultSet.features,v,isNotBlank(v.geometry)),w,w.geometry.coordinates
[0]).join("; ")).join("|")
```

Esta expresión es diferente a la usada anteriormente en que se especifica qué valor del arreglo coordenadas se desea obtener: [0]. En OpenRefine, el primer valor se indica con 0, el segundo con 1, y así sucesivamente. Dado que en los resultados de la consulta se indica primero la longitud, ésta será el valor [0], y la latitud será el valor [1] dentro del arreglo “coordinates”.

El nuevo campo creado tendrá valores como los siguientes: -64.158097; -59.846115; -63.801314

cada uno correspondiente a una longitud de uno de los resultados obtenidos de la consulta a GeoLocate para un determinado registro.

Puede repetir el proceso para obtener las latitudes, cambiando en la expresión anterior [0] por [1], y luego separar los campos por resultado, utilizando como separador “;”.

NOTA IMPORTANTE: Debe tener en cuenta que, como se mencionó antes, cuantos más datos se provean al servicio de GeoLocate en la consulta más sencillo será desglosar los resultados después. El proceso de desglose puede ser muy engorroso y requiere que sea muy meticuloso a la hora de nombrar campos y separar contenido. Si no está familiarizado con el uso de Json, es preferible que realice el desglose “pasito a pasito” para evitar perder o mezclar información. Por ejemplo, puede crearse un documento con el flujo de trabajo donde enumere los pasos a seguir con todos los detalles necesarios (incluya allí el tipo de resultados que espera ver y cómo se verían en los campos).

NOTA IMPORTANTE: A la hora de agregar datos de georreferenciación, contraste siempre los resultados contra los campos geográficos que tiene. En el caso de tener varios resultados posibles, no siempre el primer resultado es el correcto. Recuerde reportar cuál fue el proceso de georreferenciación utilizado y todos los parámetros posibles asociados. Para consultar en qué campos de Darwin Core se reporta cada parámetro, puede referirse a: <http://rs.tdwg.org/dwc/terms/index.htm#locationindex>, y consultar: <https://github.com/tdwg/dwc-qa/wiki/Georeferences>.

Limpieza de fechas utilizando Canadensys Date Parsing

a. Breve introducción

Uno de los campos sobre el que se puede corroborar la calidad de los datos es el campo de fecha: **eventDate**.

Recordemos primero la **definición de eventDate en el estándar Darwin Core** <http://rs.tdwg.org/dwc/terms/index.htm#eventDate>:

The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context. Recommended best practice is to use an encoding scheme, such as ISO 8601:2004(E).

Si piensa en un ejemplar de museo, **eventDate** refiere a cuándo fue colectado el ejemplar. Si piensa en una observación, **eventDate** refiere a cuándo fue realizada esa observación.

Darwin Core sugiere que se utilice para capturar la información de fecha el estándar **ISO 8601:2004(E)** (https://en.wikipedia.org/wiki/ISO_8601). Para fechas únicas, este estándar tiene el siguiente formato:

AAAA-MM-DDTHH:mmX

Donde:

AAAA: año, con cuatro dígitos.

MM: mes, con dos dígitos. E.g.: mayo sería 05.

DD: día, con dos dígitos. E.g.: segundo día de un mes sería 02.

T: indica que lo que viene a continuación es la hora.

HH: horas, con dos dígitos, en formato de 24 hs.

mm: minutos, con dos dígitos.

X: indica la zona horaria. La zona horaria se determina tomando como base UTC (Coordinated Universal Time). Si uno está justo sobre la zona horaria UTC, X se reemplaza por "Z". Si uno está en otra zona horaria, debe reemplazarse X por la diferencia horaria correspondiente.

Por ejemplo, Argentina es UTC-3, o sea, 03horas00minutos al oeste (-) de UTC, por lo cual X debe reemplazarse por "-0300".

NOTAS:

- De este formato, uno puede utilizar tanto el formato completo (incluyendo la hora) como sólo la primera parte, AAAA-MM-DD.

- Este formato también puede utilizarse para expresar rangos de fecha de manera estandarizada. Para ello, se usa el mismo formato y se separan las fechas con barras "/", ver ejemplos abajo.

EJEMPLOS:

FECHA ORIGINAL	FECHA ESTANDARIZADA
12 Feb 1809	1809-02-12
12/02/1809	1809-02-12
Jun 1906	1906-06
1971	1971

20 Feb 2009 8:40am UTC	2009-02-20T08:40Z
8 Mar 1963 2:07pm, en la zona horaria 6 horas más temprano que UTC	1963-03-08T14:07-0600
13-15 Nov 2007	2007-11-13/15
1 Mar 2007 1pm UTC – 11 May 2008 3:30pm UTC	2007-03-01T13:00:00Z/2008-05-11T15:30:00Z

a. Limpieza de fechas

Muchas veces, a pesar de lo que indica el estándar Darwin Core, encontramos en el campo **eventDate** fechas que no siguen el formato sugerido. Para limpiarlas, puede hacer uso de la herramienta que ofrece Canadensys: Date Parsing (<http://data.canadensys.net/tools/dates>).

Esta herramienta permite interpretar fechas, devolviéndolas en formato estándar. Ejemplos de los tipos de valores que puede interpretar son:

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

Algunas fechas, sin embargo no las interpreta, veamos el siguiente ejemplo (Figura 43):

Date parsing results				
original	year	month	day	ISO 8601
2-4-1980				
2/4/1980				
2/13/1980	1980	2	13	1980-02-13
13/2/1980	1980	2	13	1980-02-13

Figura 43.

En las dos líneas inferiores, “13” sólo puede referir a días, pues no hay un mes “13”.

En las dos líneas superiores, en cambio, “2” y “4” pueden ambos referir a mes y día. Como en distintas partes del mundo se utilizan sistemas distintos (primero se pone día y luego mes, o viceversa), la herramienta no puede determinar inequívocamente cuál es cuál, y por ende no hace la interpretación.

Debe tener esto en cuenta cuando utilice la herramienta para limpiar los datos.

Ahora sí, invoque Date Parsing desde OpenRefine. Para ello:

1. Sobre la columna **eventDate**: Edit column --> Add column by fetching URL... (Figura 44)
Con ello creará una nueva columna con los resultados que indique Canadensys.
2. En la ventana que aparece, nombre la nueva columna y pegue en el cuadro de texto la siguiente expresión:

```
"http://data.canadensys.net/tools/dates.json?data="+escape(cells["eventDate"].value,"url")"
```

Lo que hace esta expresión es pedirle a herramienta que evalúe los valores del campo **eventDate** y que nos envíe los resultados en formato JSON.

Add column by fetching URLs based on column eventDate

New column name: **Canadensys_eventDate** Throttle delay: **300** milliseconds

On error: ☒ set to blank ☐ store error

Formulate the URLs to fetch:

Expression: `"http://data.canadensys.net/tools/dates.json?data="+escape(cells["eventDate"].value,"url")"` Language: Google Refine Expression Language (GREL) ▼

No syntax error.

Preview History Starred Help

row	value	
		<code>"http://data.canadensys.net/tools/dates.json?data="+escape(cells["eventDate"].value,"url")"</code>
1.	9/16/1967	<code>http://data.canadensys.net/tools/dates.json?data=9%2F16%2F1967</code>
2.	6/4/1968	<code>http://data.canadensys.net/tools/dates.json?data=6%2F4%2F1968</code>
3.	5/27/1968	<code>http://data.canadensys.net/tools/dates.json?data=5%2F27%2F1968</code>
4.	2/9/1973	<code>http://data.canadensys.net/tools/dates.json?data=2%2F9%2F1973</code>
5.	2/1/1996	<code>http://data.canadensys.net/tools/dates.json?data=2%2F1%2F1996</code>
6.	2/1/1996	<code>http://data.canadensys.net/tools/dates.json?data=2%2F1%2F1996</code>

OK Cancel

Figura 44.

3. La limpieza puede tomar bastante tiempo, incluso horas, sea paciente... váyase a almorzar, o incluso a dormir y lo revisa al día siguiente... Cuando vuelva, encontrará el nuevo campo con los valores estandarizados! En formato JSON... (Figura 45).

<input type="checkbox"/> eventDate	<input type="checkbox"/> Canadensys_eventDate
9/16/1967	<pre>{"data":{"results":[{"originalValue":"9/16/1967","year":1967,"month":9,"day":16,"iso8601":"1967-09-16","partial":false}]}}</pre>
6/4/1968	<pre><u>{"data":{"results":[{"originalValue":"6/4/1968","error":"The date [6-4-1968] could not be precisely determined.","partial":true}]}}</u></pre>

Figura 45.

Fíjese que en el primer caso de la figura, Canadensys ha podido resolver la fecha, mientras que en el segundo caso no ha podido, dado que no puede interpretar inequívocamente “6” y “4” como día y mes o viceversa (como se explica más arriba).

- Ahora que tiene el JSON, extraerá de allí los valores de interés. Podría extraer sólo la fecha en formato ISO, o también año, mes y día en campos separados. Para ello, a partir de la columna que tiene el JSON, cree nuevas columnas: Edit column --> Add column based on this column (Figura 46).

Para extraer sólo la fecha en formato ISO, en la ventana nombre la nueva columna y en el cuadro de texto pegue la siguiente expresión:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("iso8601"))[0])
```

Add column based on column Canadensys_eventDate

New column name

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

Expression

Language
Google Refine Expression Language (GREL)

```
forEach(value.parseJson().get("data").get("results"),v,v.get("iso8601"))[0])
```

No syntax error.

Preview
History
Starred
Help

row	value	forEach(value.parseJson().get("data").get("results"),v,v.get("iso8601"))
1.	{ "data": { "results": [{"originalValue": "9/16/1967", "year": 1967, "month": "09-16", "partial": false}] }}	1967-09-16
2.	{ "data": { "results": [{"originalValue": "6/4/1968", "error": "The date [6-4-1968] could not be precisely determined.", "partial": true}] }}	null
3.	{ "data": { "results": [{"originalValue": "10/27/1968", "year": 1968, "month": "10-27", "partial": false}] }}	1968-10-27

OK
Cancel

Figura 46.

Para extraer el año, mes o día, pegue en cambio una de las siguientes expresiones:

Año:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("year"))[0])
```

Mes:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("month"))[0])
```

Día:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("day"))[0])
```

Verá que algunos de los resultados serán nulos, éstos corresponden a los casos que Canadensys no ha podido resolver (como se explica más arriba) (Figura 47).

eventDate	Canadensys_eventDate	ISO_eventDate
9/16/1967	{"data":{"results":[{"originalValue":"9/16/1967","year":1967,"month":9,"day":16,"iso8601":"1967-09-16","partial":false}]}}	1967-09-16
6/4/1968	{"data":{"results":[{"originalValue":"6/4/1968","error":"The date [6-4-1968] could not be precisely determined.","partial":true}]}}	

Figura 47.

- Para terminar de limpiar las fechas, entonces, tendrá que revisar los valores que no hayan sido estandarizados por la herramienta. Para ello, sobre el campo ISO_eventDate puede armar una faceta y seleccionar el valor "blank". Luego, arme una faceta sobre el campo eventDate (el que tenía los valores originales) y si estos son pocos, puede hacer un chequeo manual y completar el campo ISO_eventDate.

B. Servicios de reconciliación

Nos referimos a reconciliación cuando se realiza una consulta a un servicio externo a través de una interfaz de reconciliación. Es decir, es diferente a las consultas vía URL en que las reconciliaciones requieren que exista un código que vincule OpenRefine y el servicio dado, y en que los resultados de las reconciliaciones se presentan en OpenRefine de manera más amigable y más sencilla de utilizar de forma directa.

A pesar de que esta herramienta es muy útil, sólo algunos servicios cuentan con la opción de reconciliación, que ha sido construida especialmente por algún usuario interesado. Algunos ejemplos de tales servicios son los creados por Rod Page (University of Glasgow, UK), que puede consultar en el siguiente enlace: <http://iphylo.blogspot.com.ar/2012/02/using-google-refine-and-taxonomic.html>.

OpenRefine permite incorporar tantos servicios de reconciliación como se quiera, y tiene ya incorporado uno que resuelve los nombres científicos consultando las bases de datos de Encyclopedia of Life (de aquí en más EOL). Utilizaremos éste último para el siguiente ejemplo.

Reconciliación de nombres científicos utilizando Encyclopedia of Life

Como en los ejemplos anteriores, para acortar los tiempos de consulta, trabajaremos sobre facetas. Cierre las facetas que tenga abiertas y cree una nueva faceta para el campo genus (en el campo genus, haga click en la flecha azul --> Facet --> Text facet); dentro de ella escoja el género Acacia.

En el ejemplo siguiente, reconciliaremos los nombres científicos. La reconciliación sobrescribirá los valores del campo sobre el que se realice. Por lo tanto, para no perder los datos originales, crearemos un nuevo

campo que sea una copia del campo **scientificName**, sobre el que haremos la reconciliación. Para ello siga: haga click en la flecha azul --> Edit column --> Add column based on this column.

Dé un nombre al nuevo campo, por ejemplo "Reconciled_scientificName", y en el cuadro de texto deje "value". Ello hará que el nuevo campo tenga exactamente los mismos valores que el campo **scientificName**.

Llevaremos a cabo la reconciliación sobre este nuevo campo. Para ello, haga click en la flecha azul del campo --> Reconcile --> Start reconciling...

Se abrirá una ventana como la mostrada en la Figura 48. Cuando OpenRefine es instalado por primera vez, el único servicio registrado es "Wikidata Reconciliation for OpenRefine (en)". Si el servicio para Encyclopedia of Life no está en el menú, agréguelo haciendo click en "Add Standard Service..." abajo a la izquierda. Se abrirá una ventana como la mostrada en la Figura 49, en la cual puede colocar la URL correspondiente: http://iphylo.org/~rpage/phyloinformatics/services/reconciliation_eol.php. Haga click en "Add Service".

En el menú de la izquierda, escoja el servicio Encyclopedia of Life. Puede también en esta ventana agregar otros servicios ("Add Standard Service...", abajo a la izquierda).

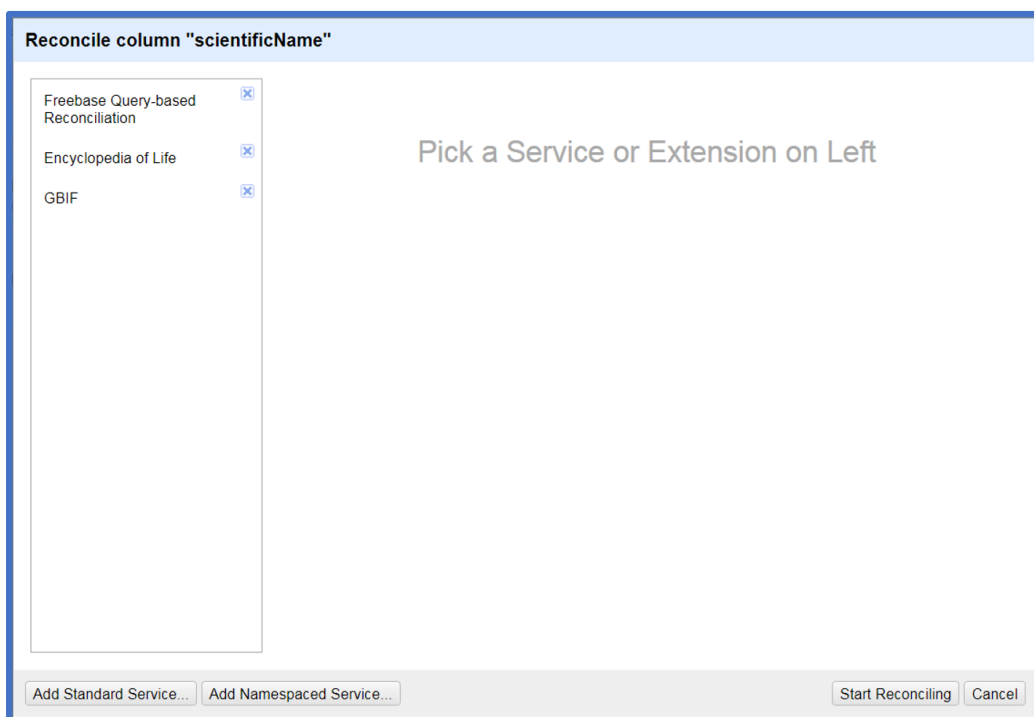
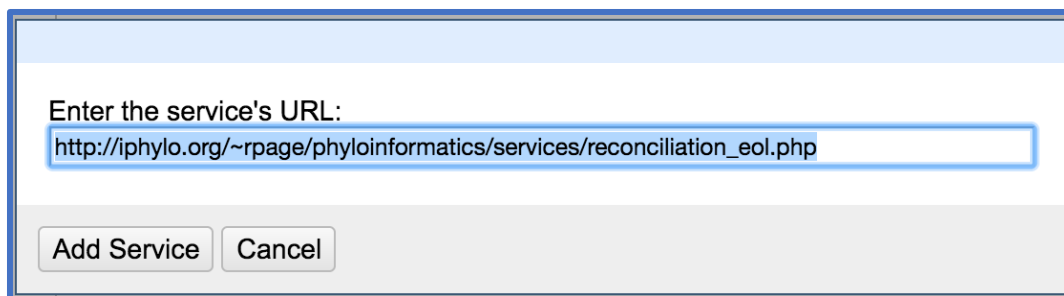


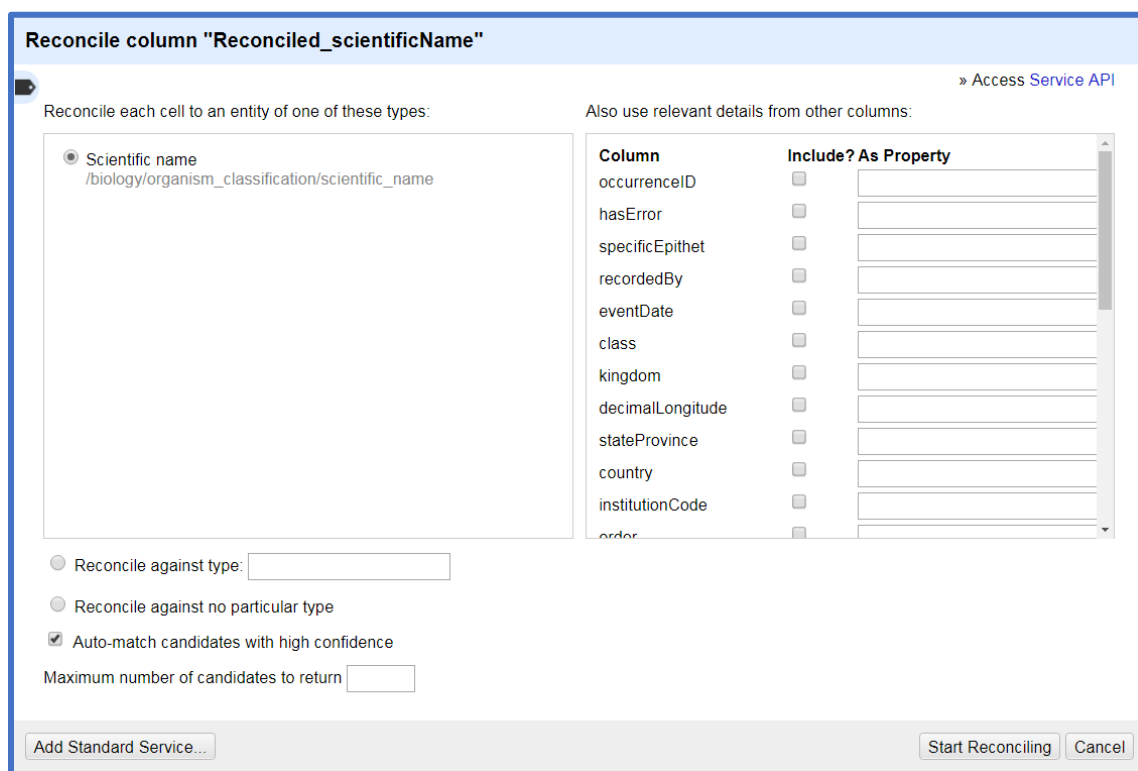
Figura 48.



Enter the service's URL:

Figura 49.

Cuando escoja el servicio, se abrirá una nueva ventana (Figura 50).



Reconcile column "Reconciled_scientificName"

» Access [Service API](#)

Reconcile each cell to an entity of one of these types:

- ☒ Scientific name
/biology/organism_classification/scientific_name

Also use relevant details from other columns:

Column	Include? As Property
occurrenceID	<input type="checkbox"/>
hasError	<input type="checkbox"/>
specificEpithet	<input type="checkbox"/>
recordedBy	<input type="checkbox"/>
eventDate	<input type="checkbox"/>
class	<input type="checkbox"/>
kingdom	<input type="checkbox"/>
decimalLongitude	<input type="checkbox"/>
stateProvince	<input type="checkbox"/>
country	<input type="checkbox"/>
institutionCode	<input type="checkbox"/>
order	<input type="checkbox"/>

☐ Reconcile against type:
☐ Reconcile against no particular type
☒ Auto-match candidates with high confidence
 Maximum number of candidates to return

Figura 50.

Sobre el panel de la derecha, puede seleccionar otros campos que podrían ser útiles para llevar a cabo la reconciliación. Para este ejemplo, no seleccionaremos ningún otro campo.

Note también que por defecto tiene seleccionado el campo "Auto-match candidates with high confidence". Esta opción seleccionará algunos valores automáticamente cuando los índices de coincidencia de los nombres con los encontrados en el servicio sean altos. Estas selecciones pueden ser modificadas luego, como se verá más adelante.

Haga click en "Start reconciling".

En el campo Reconciled_scientificName verá entonces resultados como los mostrados en la Figura 51.

▼ scientificName	▼ Reconciled_scientificName
Acacia platensis	Acacia platensis
Acacia adhaerens	Acacia adhaerens <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Acacia adhaerens Benth. (1) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Acacia adhaerens var. adhaerens (1) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic
Acacia spegazziniana	Acacia spegazziniana <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Acacia martiusiana (Steud.)Burkart (1) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Acacia martiusiana (Steud.)Burkart (1) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new topic
Acacia nitidifolia	Acacia nitidifolia Speg. Choose new match
	⋮

Figura 51.

Allí podrá ver que lo siguiente:

1. Barra con el porcentaje de registros cuyos nombres han sido reconciliados (barra verde bajo el nombre del campo)
2. Resolución de los nombres, donde puede ver los siguientes patrones:
 - El nombre NO ha sido encontrado en EOL. Ejemplo: *A. platensis*.
 - El nombre ha sido encontrado en EOL, pero con más de una variante, que pueden ser sólo en los autores o también indicar sinonimias. Ejemplos: *A. adhaerens* y *A. spegazziniana*, respectivamente.
 - El nombre ha sido encontrado en EOL con una única variante, a la cual se ha reconciliado el nombre provisto. Ejemplo: *A. nitidifolia*.

A partir de esta reconciliación, se pueden tomar varias vías de acción:

a. Aceptar los valores reconciliados para un nombre. Para aquellos nombres que tengan más de una opción de reconciliación, puede hacer click en los íconos con uno o dos ticks en el valor correcto. El primero implica que aceptará el valor únicamente para ese registro. El segundo implica que aceptará el valor para todos los registros que tengan contengan ese mismo nombre.

b. Rechazar los valores reconciliados para un nombre. Si no está de acuerdo con los valores reconciliados, haga click en los íconos con uno o dos tildes en la opción “Create new topic”. De esta forma, mantendrá el valor original (para un registro en particular o para todos los registros con ese nombre, respectivamente). En el caso de valores que han sido reconciliados automáticamente (ejemplo: *A. nitidifolia*), debe primero escoger “Choose new match”.

c. Aceptar o rechazar los valores reconciliados para todos los nombres. Para aceptar o rechazar todos los nombres reconciliados, puede seguir la siguiente ruta: haga click sobre la flecha azul del campo --> Reconcile --> Actions --> [algunas opciones descriptas a continuación].

Opciones:

--> Match each cell to its best candidate. Se reconciliarán todos los valores a la primer (y mejor) opción.

--> Create a new topic for each cell. Creará como valor el valor original, pero podrá ver aún las opciones obtenidas a partir de EOL en “Choose new match”.

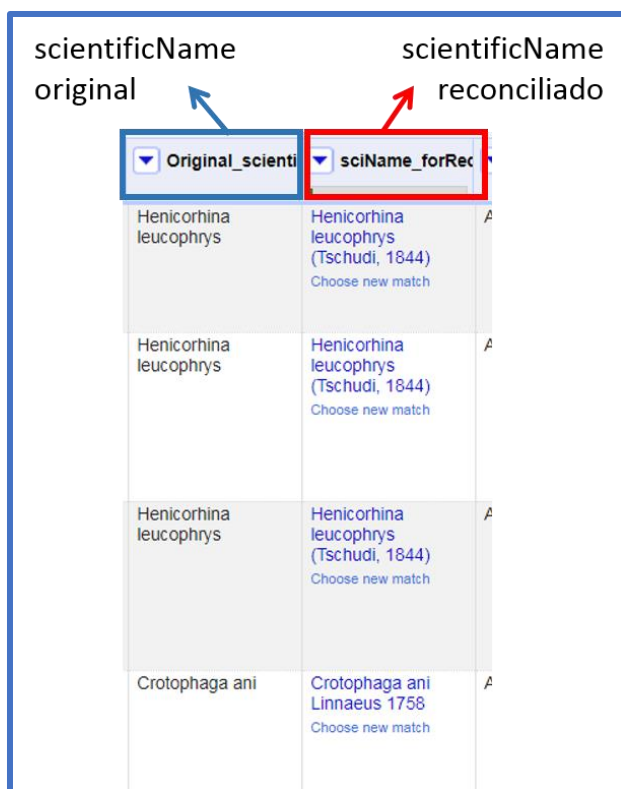
--> Discard reconciliation judgments. Esta opción le permite descartar todas las elecciones que haya hecho. Por ejemplo, si para algunos nombres escogió alguna de las opciones y luego se arrepiente, puede descartar todas sus elecciones con esta acción.

--> Clear reconciliation data. Descartará todos los resultados que obtuvo de EOL. Volverá a ver en el campo los valores originales, pero ya sin opciones para escoger (es como volver al paso anterior a la reconciliación).

NOTA IMPORTANTE: Cuando reconcilia un campo contra algún servicio los cambios sólo se incorporarán en el campo correspondiente. Por ello, es importante considerar qué otros campos podrían necesitar cambios similares. Por ejemplo, si reconcilia el campo *scientificName*, los cambios sólo se verán en dicho campo, pero los valores de campos como *genus*, *specificEpithet*, *scientificNameAuthorship*, etc., no serán afectados. Esto podría resultar en que un nombre en el campo *scientificName*, que ha sido corregido, no se corresponda con el género y el epíteto en los otros campos.

1. Cómo retener los nombres científicos incluyendo autor

Luego de la reconciliación con EOL, tendrá una columna con los valores reconciliados, como se muestra en la Figura 52.



scientificName original	scientificName reconciliado
Original_scienti	sciName_forRec
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match
Crotophaga ani	Crotophaga ani Linnaeus 1758 Choose new match

Figura 52.

Por default, si bien se ven los autores de los nombres, el valor de las celdas es sólo el nombre científico. Para retener los autores, puede armar una nueva columna en base a la columna reconciliada:

Edit column --> Add column based on this column...

Allí, nombre la nueva columna y en el cuadro coloque la siguiente expresión (Figura 53):

`cell.recon.match.name`

Esta expresión toma el nombre completo provisto por EOL, incluyendo autor.

Add column based on column sciName_forReconc

New column name: **scientificName_reconc_withAuthors**

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

Expression: **cell.recon.match.name** Language: Google Refine Expression Language (GREL) No syntax error.

Preview History Starred Help

row	value	cell.recon.match.name
146.	Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844)
247.	Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844)
260.	Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844)
353.	Crotophaga ani	Crotophaga ani Linnaeus 1758
414.	Colibri coruscans coruscans	Colibri coruscans coruscans (Gould, 1846)
448.	Colibri coruscans coruscans	Colibri coruscans coruscans (Gould, 1846)
481.	Colibri coruscans coruscans	Colibri coruscans coruscans (Gould, 1846)

OK Cancel

Figura 53.

Tendrá entonces un resultado como el que se muestra en la Figura 54:

scientificName original	scientificName reconciliado	scientificName reconciliado incluyendo autor
Original_scientificName	sciName_forRe	scientificName
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	Henicorhina leucophrys (Tschudi, 1844)
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	Henicorhina leucophrys (Tschudi, 1844)
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	Henicorhina leucophrys (Tschudi, 1844)
Crotophaga ani	Crotophaga ani Linnaeus 1758 Choose new match	Crotophaga ani Linnaeus 1758

Figura 54.

2. Cómo capturar las URL que corresponden a cada nombre en EOL

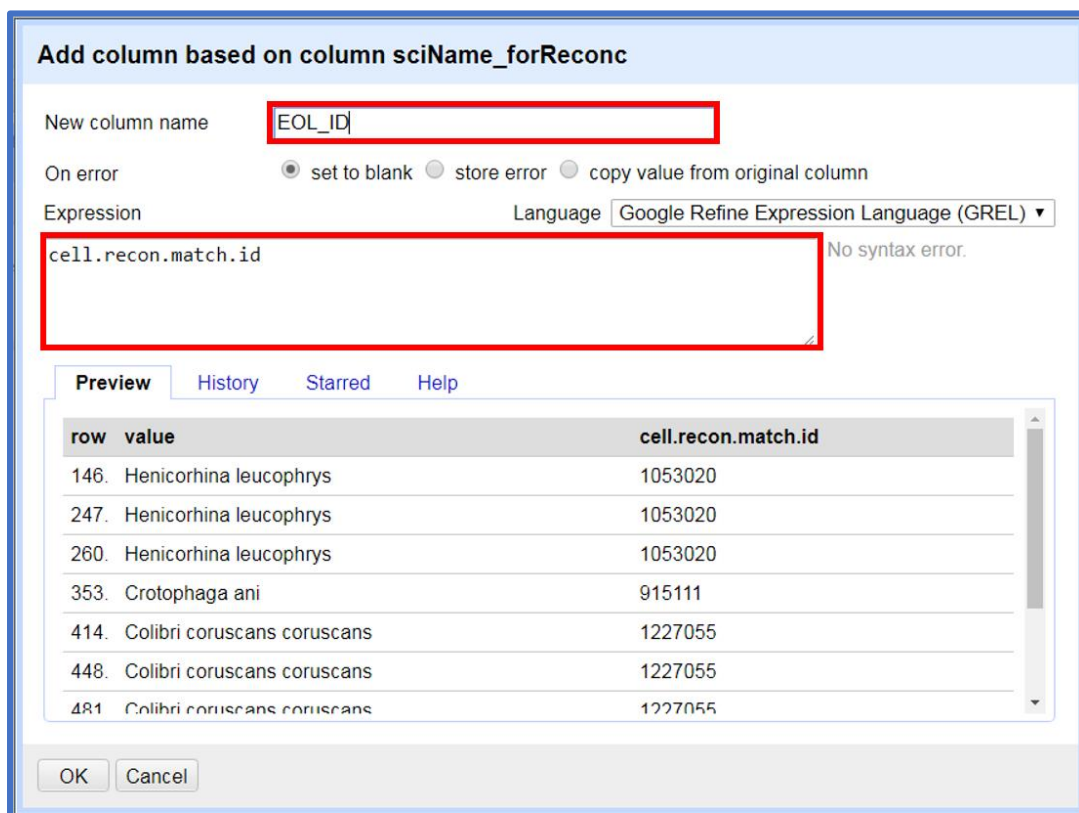
Como puede observar en el campo reconciliado, se puede acceder a las páginas correspondientes a cada nombre en EOL (haciendo click en los link). Si quisiera retener esas páginas de referencia, puede armar un nuevo campo que contenga las URL correspondientes.

Para ello, a partir del campo reconciliado primero arme una nueva columna que retenga los identificadores de cada nombre en EOL (los IDs) (Figura 55):

Edit column --> Add column based on this column...

Allí nombre la nueva columna y en el cuadro coloque la siguiente expresión:

cell.recon.match.id



Add column based on column sciName_forReconc

New column name: **EOL_ID**

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

Expression: **cell.recon.match.id** Language: Google Refine Expression Language (GREL) No syntax error.

Preview History Starred Help

row	value	cell.recon.match.id
146.	Henicorhina leucophrys	1053020
247.	Henicorhina leucophrys	1053020
260.	Henicorhina leucophrys	1053020
353.	Crotophaga ani	915111
414.	Colibri coruscans coruscans	1227055
448.	Colibri coruscans coruscans	1227055
481.	Colibri coruscans coruscans	1227055

OK Cancel

Figura 55.

Tendrá entonces un resultado como el que se muestra en la Figura 56:

scientificName original	scientificName reconciliado	ID del nombre en EOL
<input type="text" value="Original_scienti"/>	<input type="text" value="sciName_forRe"/>	<input type="text" value="EOL_ID"/>
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	1053020
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	1053020
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	1053020
Crotophaga ani	Crotophaga ani Linnaeus 1758 Choose new match	915111

Figura 56.

A continuación, arme una nueva columna en base a la columna de IDs:

Edit column --> Add column based on this column...

Nombre la nueva columna y en el cuadro coloque la siguiente expresión (Figura 57):

"http://eol.org/pages/" + value

Esta expresión arma una URL utilizando para cada celda el valor del ID correspondiente.

Add column based on column EOL_ID

New column name

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

Expression Language Google Refine Expression Language (GREL) ▼ No syntax error.

Preview History Starred Help

row	value	"http://eol.org/pages/" + value
146.	1053020	http://eol.org/pages/1053020
247.	1053020	http://eol.org/pages/1053020
260.	1053020	http://eol.org/pages/1053020
353.	915111	http://eol.org/pages/915111
414.	1227055	http://eol.org/pages/1227055
448.	1227055	http://eol.org/pages/1227055
481.	1227055	http://eol.org/pages/1227055

OK Cancel

Figura 57.

Tendrá entonces el resultado que se muestra en la Figura 58:

scientificName original	scientificName reconciliado	ID del nombre en EOL	URL para el nombre en EOL
Original_scientificName	sciName_forReconciliation	EOL_ID	EOL_URL
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	1053020	http://eol.org/pages/1053020
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	1053020	http://eol.org/pages/1053020
Henicorhina leucophrys	Henicorhina leucophrys (Tschudi, 1844) Choose new match	1053020	http://eol.org/pages/1053020
Crotophaga ani	Crotophaga ani Linnaeus 1758 Choose new match	915111	http://eol.org/pages/915111

Figura 58.