



OpenRefine para la limpieza de datos sobre biodiversidad

Paula F. Zermoglio, John R. Wieczorek

Versión 1.0, 2020-06-16 13:40:04 UTC

Tabla de Contenido

Colofón	1
Cita bibliográfica sugerida	1
Licencia	1
URI persistente	1
Control de documentos	1
Imagen de la portada	1
Prefacio	2
Objetivo	2
Cómo usar esta guía	2
1. Carga de datos y creación de un proyecto	3
2. Limpieza de datos	9
2.1. Manejo básico de columnas	9
2.2. Uso de Facetas	16
2.3. Uso de Filtros	26
2.4. Uso de Agrupamientos	32
2.5. Deshacer y rehacer cambios	34
2.6. Marcado de registros: banderas y estrellas	37
3. Guardado y exportación de datos y proyectos	41
3.1. Guardado de datos y proyectos	41
3.2. Exportación de datos y proyectos	41
4. Consultas a servicios externos	46
4.1. Consultas externas a través de URLs	46
Epílogo	65
Agradecimientos	65
Apéndice 1: instalación de OpenRefine	66
Requerimientos	66
Instalación en MS Windows	66
Instalación en Mac	66
Para saber más	66

Colofón

Cita bibliográfica sugerida

Zermoglio PF & Wieczorek JR (2020) Guía de Uso Básico de OpenRefine para la limpieza de datos sobre biodiversidad, Versión 3. Copenhagen: GBIF Secretariat. <https://doi.org/10.15468/doc-gzjg-af18>.

Licencia

El documento Guía de Uso Básico de OpenRefine para la limpieza de datos sobre biodiversidad se publica bajo una licencia [Atribución-CompartirIgual 4.0 Internacional](https://creativecommons.org/licenses/by-sa/4.0/deed.es) [<https://creativecommons.org/licenses/by-sa/4.0/deed.es>].

URI persistente

<https://doi.org/10.15468/doc-gzjg-af18>

Control de documentos

Versión 3, edición de revisión comunitaria, Marzo 2020.

Este documento se basa en una publicación anterior de la guía por los mismos autores.

Imagen de la portada

Una manada de guanaco (*Lama guanicoe*), Lago Argentino, Santa Cruz, Argentina. Foto 2016 Diego Carús via [iNaturalist Research-grade Observations](https://www.gbif.org/occurrence/2005372769) [<https://www.gbif.org/occurrence/2005372769>], licenciada bajo [CC BY-NC 4.0](http://creativecommons.org/licenses/by-nc/4.0/) [<http://creativecommons.org/licenses/by-nc/4.0/>].

Prefacio

Objetivo

La presente guía ha sido construida con fines únicamente pedagógicos. El objetivo de esta guía es mostrar cómo utilizar algunas de las funciones básicas de OpenRefine que pueden utilizarse para evaluar y mejorar la calidad de datos de biodiversidad.

Cómo usar esta guía

En esta guía se muestra cómo utilizar funciones básicas de OpenRefine para la evaluación y mejoramiento de la calidad de un conjunto de datos de biodiversidad. Los ejemplos de uso presentados en esta guía constituyen sólo algunas de las alternativas posibles para el tratamiento de datos en OpenRefine.

Para hacer un mejor uso de esta guía, se recomienda seguir los pasos utilizando el programa OpenRefine y el conjunto de datos modelo provisto junto con esta guía. Todos los ejemplos presentados corresponden a un conjunto de datos de biodiversidad que ha sido específicamente modificado por los autores.

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

Software

La versión de OpenRefine utilizada para la confección de esta guía es OpenRefine 3.1.

Para ver detalles sobre cómo descargar e instalar OpenRefine en su computadora, ver el [Apéndice 1](#).

Datos

El conjunto de datos modelo utilizado en esta guía puede obtenerse aquí: [EjercicioModelo_OpenRefine_Datos.csv](#) [..../data/EjercicioModelo_OpenRefine_Datos.zip]. Descargar el archivo, y no manipularlo en otros programas (e.g., MS Excel) antes de abrirlo en OpenRefine, dado que ello puede cambiar los formatos y/o codificación del archivo.

El conjunto de datos modelo fue derivado a partir del conjunto de datos:

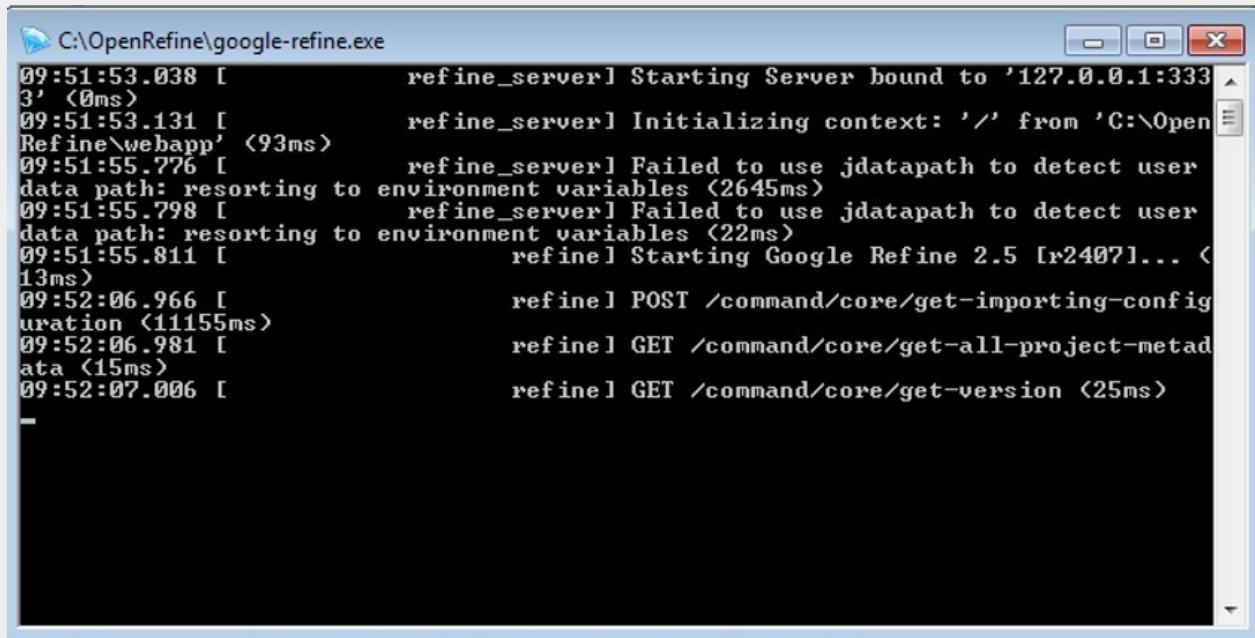
Williams J (2018). Colección de Herbario. Version 3.1. Facultad de Ciencias Naturales y Museo - U.N.L.P.. Occurrence dataset <https://doi.org/10.15468/i9bj5r> accessed via GBIF.org on 2019-04-18.

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:

Paso 1

Abra la aplicación OpenRefine. Si utiliza Windows, se abrirá una ventana de comandos que mostrará las acciones que 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'
3' <0ms>
09:51:53.131 [refine_server] Initializing context: '/' from 'C:\OpenRefine\google-refine.exe'
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 [r2407]... <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>



Figura 2

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

Además puede cambiar la configuración de idioma. Para ello, haga click en “Idioma” y en la siguiente pantalla (**Figura 3**) seleccione el idioma preferido. Aceite los cambios. En esta guía se utilizará el idioma Español.

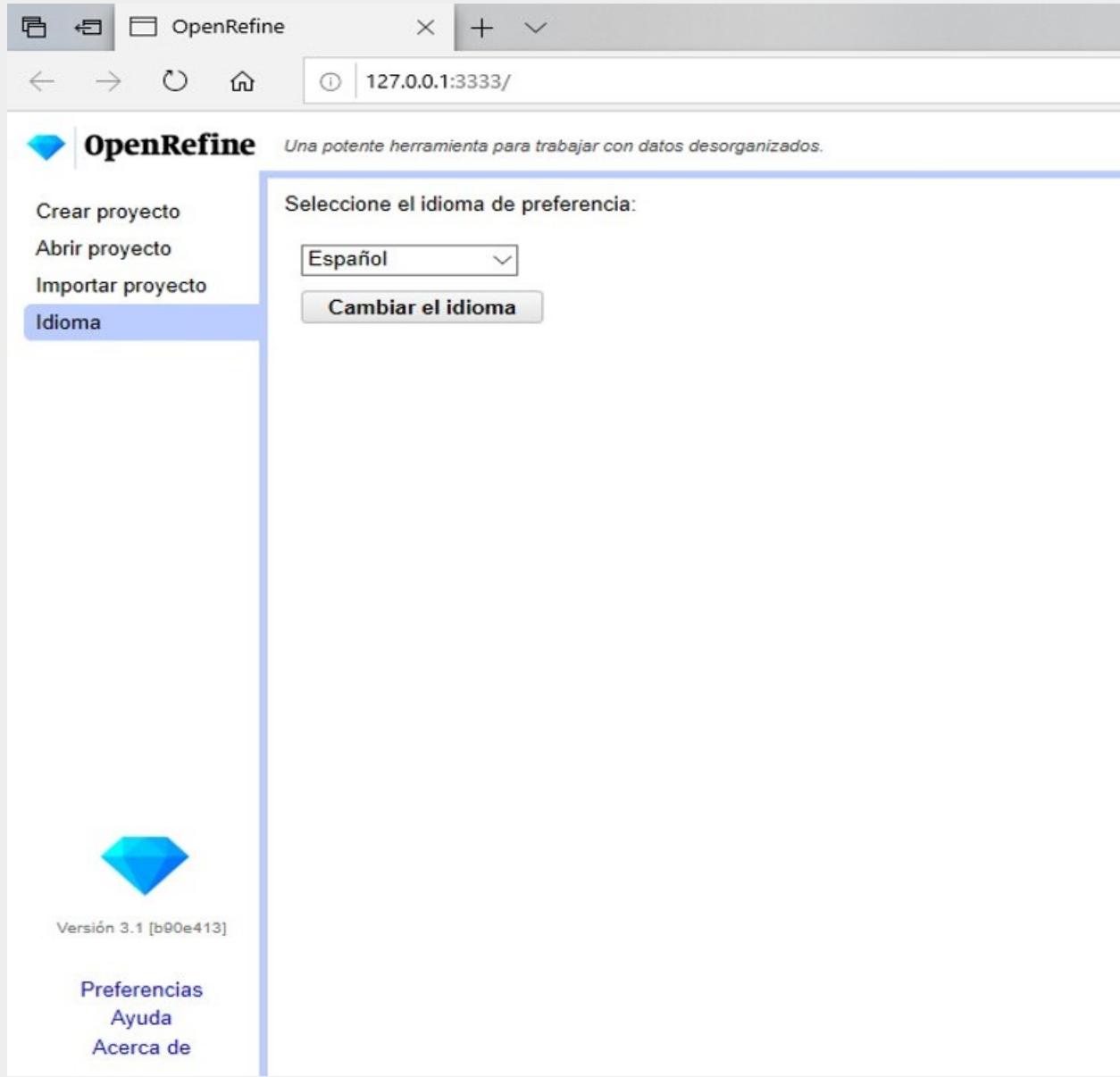


Figura 3

Paso 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 “Siguiente”.

Para seguir esta guía, cargue el archivo proporcionado, al que puede acceder a través del enlace provisto en la sección Comentarios Preliminares.



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).

Verá entonces una pantalla como la que se muestra en la [Figura 4](#). Allí puede ver una muestra de sus datos, dar nombre a su proyecto, 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 “Codificación de caracteres”. Asegúrese de escoger correctamente la codificación. Si está utilizando el conjunto de datos de prueba proporcionado, escoja UTF-8 ([Figura 5](#)).

OpenRefine presenta la opción de “Detectar y transformar texto en números, fechas, ...”. Si esta opción es seleccionada, el programa tratará de interpretar ciertos campos transformándolos a determinados formatos. Por ejemplo, si detecta campos de fecha, tratará de colocar los valores de las celdas de ese campo en formato de fecha estándar. Dada la naturaleza de los datos sobre biodiversidad con los que solemos trabajar, estas interpretaciones pueden ser incorrectas e introducir más errores. Asegúrese entonces de desmarque esta opción durante el paso de importación de datos.

OpenRefine Una potente herramienta para trabajar con datos desorganizados.

Crear proyecto

Abrir proyecto Importar proyecto Idioma

Nombre del proyecto: EjercicioModelo_OpenRefine_Data Tags: Crear proyecto »

	occurrenceID	specificEpithet	recordedBy	eventDate	class	kingdom	decimalLongitude	stateProvince	country	institutionCode	order	collect
1.	urn:catalog:fcnym.unlp.edu.ar:herb:005567			8/24/1967	Magnoliopsida	Plantae				fcnym.unlp.edu.ar		herb
2.	urn:catalog:fcnym.unlp.edu.ar:herb:009619		Arechavaleta	3/2/1937		Plantae		Montevideo	Uruguay	fcnym.unlp.edu.ar		herb
3.	urn:catalog:fcnym.unlp.edu.ar:herb:004997	argentinensis	Boffa, P.	2/20/1996	Magnoliopsida	Plantae		San Luis	Argentina	fcnym.unlp.edu.ar	Asterales	herb
4.	urn:catalog:fcnym.unlp.edu.ar:herb:002046	lasiocarpa	Lorentz, Paul(Pablo) GÃ¼nther	8/10/1981	Magnoliopsida	Plantae			Argentina	fcnym.unlp.edu.ar	Asterales	herb
5.	urn:catalog:fcnym.unlp.edu.ar:herb:002052	sprengeliana	Gardner, George	1/1/1987	Magnoliopsida	Plantae				fcnym.unlp.edu.ar	Asterales	herb
6.	urn:catalog:fcnym.unlp.edu.ar:herb:002048	doniana		12/27/1925	Magnoliopsida	Plantae				fcnym.unlp.edu.ar	Asterales	herb
7.	urn:catalog:fcnym.unlp.edu.ar:herb:002059	calcarasana	Frnck, Nicolas	2/8/1951	Magnoliopsida	Plantae				fcnym.unlp.edu.ar	Asterales	herb

Abrir archivo como: CSV / TSV / separator-based files Line-based text files Fixed-width field text files PC-Axis text files JSON files MARC files RDF/LD files RDF/N3 files RDF/Turtle files

Codificación de caracteres: Actualizar vista previa

Las columnas se encuentran separadas por: comas (CSV), tabulaciones (TSV), personalizado: , Ignorar caracteres especiales con \ Nombres de columna (separados por comas):

Ignorar primera(s) 0 linea(s) al inicio del archivo
 Seleccionar primera(s) 1 linea(s) para los nombres de las columnas
 Descartar primera(s) 0 fila(s) de datos
 Cargar al menos 0 fila(s) de datos
 Usar carácter " para encerrar celdas que contengan separadores de columnas
 Detectar y transformar texto en n鷊eros, fechas, ...
 Cargar filas en blanco
 Cargar celdas en blanco como nulas
 Cargar el origen del archivo (nombres, URLs) en cada fila

Versión 3.1 [b90e413]

Preferencias Ayuda Acerca de

Figura 4

Seleccionar codificación

Codificaciones populares Todas las Codificaciones

Codificación	Alias
ISO-8859-1	819, ISO8859-1, I1, ISO_8859-1:1987, ISO_8859-1, 8859_1, iso-ir-100, latin1, cp819, ISO8859_1, IBM819, ISO_8859_1, IBM-819, csISOLatin1
US-ASCII	ANSI_X3.4-1968, cp367, csASCII, iso-ir-6, ASCII, iso_646.irv:1983, ANSI_X3.4-1986, ascii7, default, ISO_646.irv:1991, ISO646-US, IBM367, 646, us
UTF-16	UTF_16, unicode, utf16, UnicodeBig
UTF-16BE	X-UTF-16BE, UTF_16BE, ISO-10646-UCS-2, UnicodeBigUnmarked
UTF-16LE	UnicodeLittleUnmarked, UTF_16LE, X-UTF-16LE
UTF-8	unicode-1-1-utf-8, UTF8

Cancelar

Figura 5

Paso 3

Cree el proyecto. Una vez que haya seleccionado las opciones de carga de datos, haga click en el botón “Crear Proyecto” arriba a la derecha.

Paso 4

¡Felicitaciones! Ya tiene un proyecto (lo verá como en la Figura 6).

The screenshot shows the OpenRefine interface with the following details:

- Title Bar:** EjercicioModelo_OpenR
- Address Bar:** 127.0.0.1:3333/project?project=1867141645905
- Toolbar:** Includes buttons for file operations (New, Open, Save, etc.), a search bar, and links to "Abrir...", "Exportar...", and "Ayuda".
- Header:** OpenRefine EjercicioModelo_OpenRefine_Datos csv | Enlace permanente
- Facet Panel:** Shows "24984 filas" and a "Mostrar como:" dropdown set to "filas registrosMostrar: 5 10 25 50 filas". It also includes "Facetas / Filtros" and "Deshacer / Rehacer 0 / 0".
- Help Panel:** Contains sections for "Usar facetas y filtros" (using facets and filters), "Problemas para comenzar?" (problems starting), and "Vea los videos de ayuda" (see help videos).
- Data Grid:** A table with 10 rows of data, each with a star icon and a blue question mark icon. The columns are labeled: Todo, occurrenceID, specificEpithet, recordedBy, eventDate, class, kingdom, decimalLongitude, stateProvince, and country. The first row shows: 1. urn:catalog:fcnym.unlp.edu.ar:herb:05567, Arechavaleta, 8/24/1967, Magnoliopsida, Plantae, Montevideo, Uruguay, fcny.

Figura 6



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

2.1. Manejo básico de 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.

2.1.1. Manejo básico de columnas

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

1. Renombrar un campo.

Hacer click en **la ▼ azul del campo > Editar columnas > Renombrar esta columna**

2. Eliminar un campo.

Hacer click en **la ▼ azul del campo > Editar columnas > Eliminar esta columna**

3. Mover un campo.

Hacer click en **la ▼ azul del campo > Editar columnas > Mover columna al principio**

... la ▼ azul del campo > Editar columnas > Mover columna al final

... la ▼ azul del campo > Editar columnas > Mover columna a la izquierda

... la ▼ azul del campo > Editar columnas > Mover columna a la derecha

Estas tres opciones pueden verse en la [Figura 7](#).

recordedBy	eventDate	class	kingdom	dec
Facetas	967	Magnoliopsida	Plantae	
	37		Plantae	
	996	Magnoliopsida	Plantae	
Editar columnas ►				
Dividir en varias columnas...				
Transponer ►				
Aregar columna basada en esta columna...				
Ordenar... ►				
Aregar columna accediendo a URLs...				
Ver ►				
Añadir columnas de valores conciliados...				
Cotejar ►				
Renombrar esta columna				
Eliminar esta columna				
Mover columna al principio				
Mover columna al final				
Mover columna a la izquierda				
Mover columna a la derecha				

Figura 7

4. Reordenar o eliminar varios campos a la vez.

Hacer click en la ▼ azul en el campo “Todo” > Editar columnas > Ordenar / Eliminar columnas... (Figura 8a).

Se abrirá entonces una ventana como la que se muestra en la Figura 8b. Allí puede ordenar los campos simplemente arrastrándolos arriba o abajo en la lista, y eliminarlos arrastrándolos hacia la parte derecha de la ventana. Una vez que termine de modificar el orden de los campos, haga click en “Aceptar”.

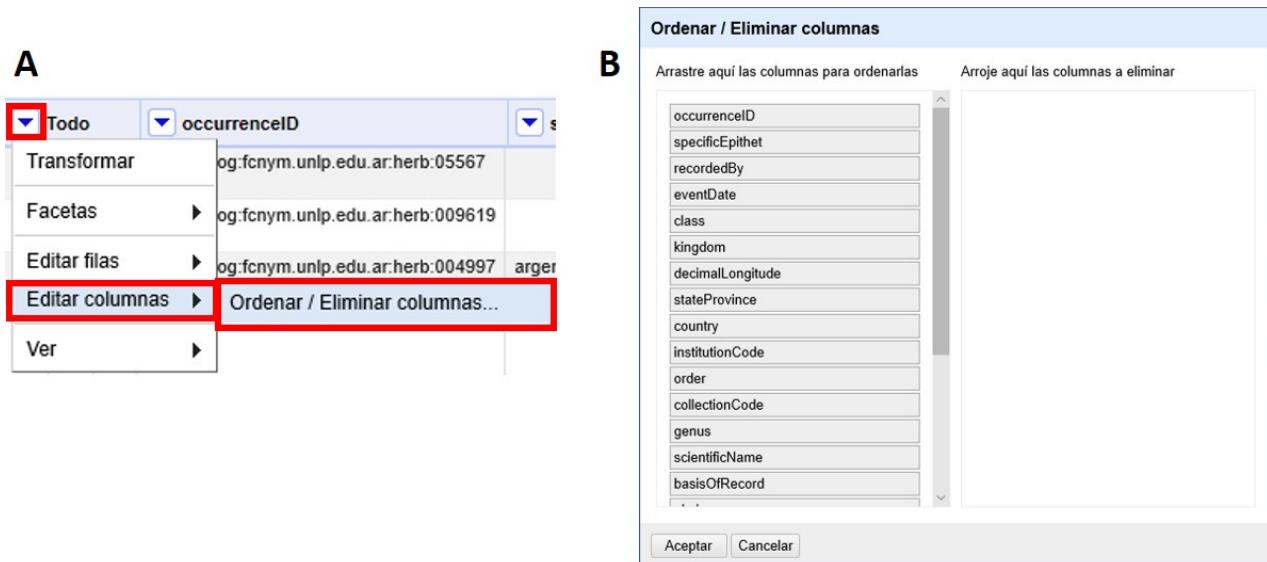


Figura 8

En OpenRefine se considera que cualquiera de los cuatro cambios descriptos anteriormente son cambios a los datos, y por ende se registran como tales en el historial de cambios (Ver más abajo sección **Deshacer y rehacer cambios**).

2.1.2. 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:

Editar columnas > Agregar columna basada en esta columna...

Se abrirá una ventana como la que se muestra en la **Figura 9**.

Arriba de todo, coloque el nombre del nuevo campo.

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 hayan sido utilizados para nombrar otros campos dentro del proyecto. 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 "Expresión" escriba: `null`. Ello quiere decir que se creará un campo con valores nulos. Luego oprima "Aceptar". Alternativamente, en vez de `null` puede colocar la expresión: `""`, y el nuevo campo tendrá valores en blanco.

Agregar columna basada en la columna occurrenceID

Nuevo nombre de la columna	CampoDePrueba																								
core-views/addasdasd	<input checked="" type="radio"/> cambiar a en blanco <input type="radio"/> guardar error <input type="radio"/> copiar valor de la columna original																								
Expresión	Lenguaje General Refine Expression Language (GREL) <input type="button" value="▼"/>																								
null																									
No hay error de sintaxis.																									
Vista previa Historial Con estrella Ayuda																									
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>row</th> <th>value</th> <th></th> </tr> </thead> <tbody> <tr><td>1.</td><td>urn:catalog:fcnym.unlp.edu.ar:herb:05567</td><td>null</td></tr> <tr><td>2.</td><td>urn:catalog:fcnym.unlp.edu.ar:herb:009619</td><td>null</td></tr> <tr><td>3.</td><td>urn:catalog:fcnym.unlp.edu.ar:herb:004997</td><td>null</td></tr> <tr><td>4.</td><td>urn:catalog:fcnym.unlp.edu.ar:herb:002046</td><td>null</td></tr> <tr><td>5.</td><td>urn:catalog:fcnym.unlp.edu.ar:herb:002052</td><td>null</td></tr> <tr><td>6.</td><td>urn:catalog:fcnym.unlp.edu.ar:herb:002048</td><td>null</td></tr> <tr><td>7.</td><td>urn:catalog:fcnym.unlp.edu.ar:herb:002050</td><td>null</td></tr> </tbody> </table>		row	value		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:002050	null
row	value																								
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:002050	null																							
<input type="button" value="Aceptar"/> <input type="button" value="Cancelar"/>																									

Figura 9

El nuevo campo, con el nombre que le haya dado, aparecerá a la derecha de aquel a partir del cual fue generado.

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, es posible 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.



2.1.3. Nuevas columnas a partir transformaciones simples de otras columnas

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

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 ▼ azul del campo **genus** > Editar columnas > Agregar columna basada en esta columna..."

Se abrirá una nueva ventana (Figura 10). 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 (" ").

Agregar columna basada en la columna genus

Nuevo nombre de la columna

core-views/addasdads cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje

`cells["genus"].value + " " + cells["specificEpithet"].value` No hay error de sintaxis.

Vista previa Historial Con estrella Ayuda

row	value
1.	null
2.	null
3.	null
4.	Filago
5.	Flotovia
6.	Flotovia
7.	Calinico...

cells["genus"].value + " " + c ...

Aceptar Cancelar

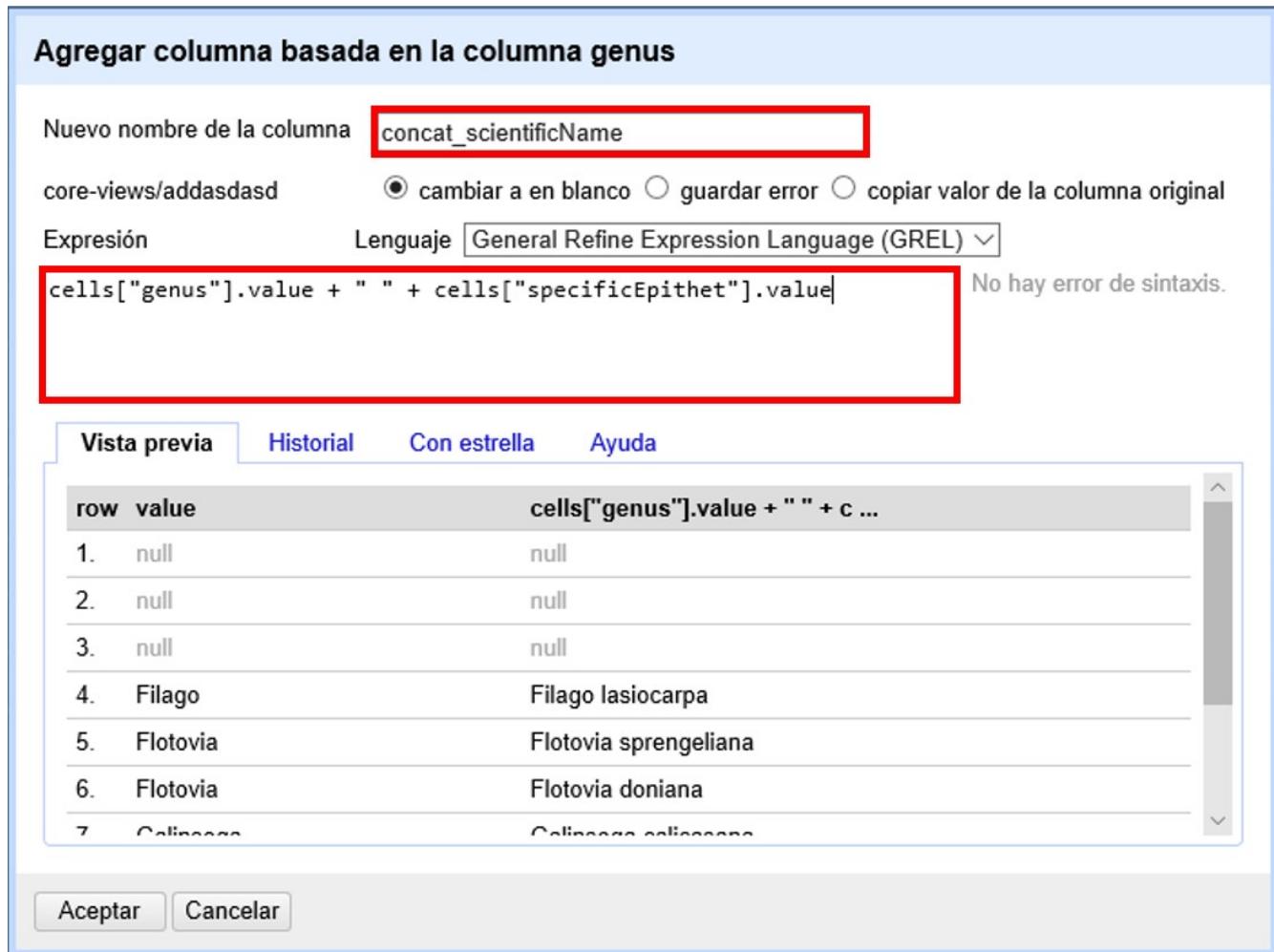


Figura 10

Note que esta expresión funciona cuando ambos campos, "genus" y "specificEpithet", tienen valores, es decir no son nulos. Si alguno de los dos campos tiene valores nulos, entonces no se lleva a cabo la concatenación. Por ejemplo, si hay un valor para genus pero specificEpithet está vacío, el campo de concatenación aparecerá vacío. Esto se debe a que no se puede operar sobre valores nulos.

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.

Los resultados esperados utilizando cada una de las dos fórmulas se resumen en la siguiente tabla:

Expresión	genus	specificEpithet	concatScientificName
1	Filago	lasiocarpa	Filago laiocarpa
	Filago	<i>null</i>	<i>null</i>
	<i>null</i>	lasiocarpa	<i>null</i>
2	Filago	lasiocarpa	Filago laiocarpa
	Filago	<i>null</i>	Filago
	<i>null</i>	lasiocarpa	lasiocarpa



Para evitar 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 4](#)).

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 ▼ azul del campo "`eventDate`" > Editar columnas > Dividir en varias columnas..."

Se abrirá una nueva ventana ([Figura 11](#)). 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 barras oblicuas (""/"), de modo que elegiremos esta barra como separador.



Desmarque la opción "Eliminar esta columna" a la derecha. Si la deja seleccionada, perderá el campo original y sólo tendrá los tres nuevos campos.

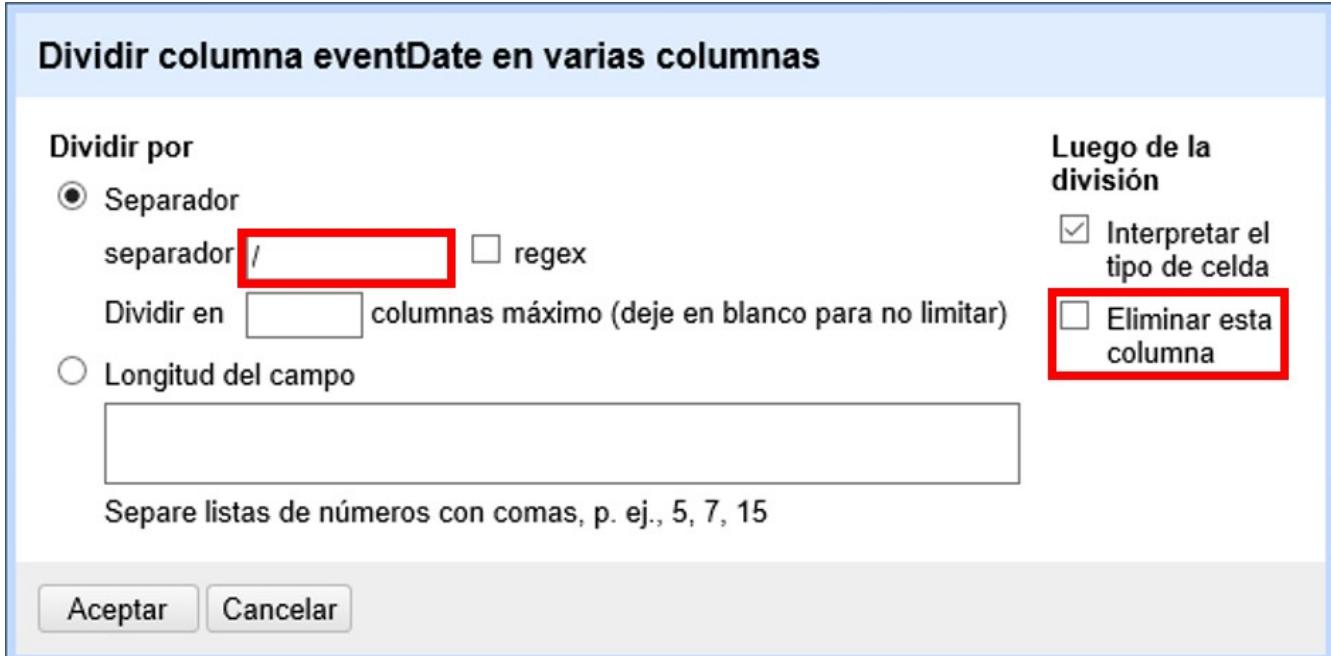


Figura 11

Una vez que oprima Aceptar, se crearán las nuevas columnas a la derecha del campo "eventDate". OpenRefine las nombrará automáticamente agregando números al final del nombre (en este caso: eventDate1, eventDate2 y eventDate3). Cambie los nombres de las columnas por los que corresponda (**la ▼ azul > Editar columnas > Renombrar esta columna**). En este caso, nómbrerlos "year", "month" y "day" según corresponda.

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" 0

2.2. Uso de Facetas

La función “Facetas” es una forma de visualización de los datos, que permite el tratamiento en bloque de grupos de registros. Las facetas se pueden aplicar a celdas que contengan cualquier tipo de texto, números o fechas.

2.2.1. Facetas de texto

Ubique la columna “kingdom” y haga click sobre la ▼ azul. Dentro de “Facetas”, escoja “Faceta de texto”, como se muestra a continuación ([Figura 12a](#)). Se abrirá entonces a la izquierda una ventana con la faceta ([Figura 12b](#)).

A**B**

<input checked="" type="checkbox"/> kingdom	cambiar
5 choices	Ordenar por: A-Z conteo
Plantae	24947
Plantae	15
Plantae	2
Plante	3
Plants	17
Facetas por conteo de opciones	

Figura 12

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

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 “editar”. Se abrirá entonces una pequeña ventana donde puede cambiar el valor (*Figura 13*). Para guardar el cambio haga click en “Aplicar”, 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”.

The screenshot shows the 'Facetas' (Facets) interface with a list of values under the 'kingdom' facet. The values are: Plantae 24947, Plantae 15, Plantae 2, Plante 3, Plants 17, and Facetas por conteo de opciones. A red box highlights the 'Plante' value. A modal dialog box is open over a table, with the input field containing 'Plante'. The dialog has 'Aplicar' (Apply) and 'Cancelar' (Cancel) buttons at the bottom.

	5567	8/24/1967	8	24	19
valeta	9619	3/2/1937	3	2	19
?	4997	2/20/1996	2	20	19
?	2046	8/10/1981	8	10	19

Figura 13

2.2.2. Facetas y espacios en blanco

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 14*). Sin embargo, estos valores sí son diferentes: tienen espacios adicionales al final del valor de texto.



Figura 14

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 ▼ azul y siga las siguientes opciones ([Figura 15](#)):

Editar celdas > Transformaciones comunes > Quitar espacios al inicio y final

Esta función permite eliminar espacios en blanco que puedan aparecer al principio y al final de cadenas de texto. Cuando termine este paso, los 24,984 registros deberían tener el valor "Plantae" en la columna "**kingdom**".

kingdom	phylum	class	order	family	scientificName	scientific
Facetas	phyta	Magnoliopsida				
Filtro de texto	hyta					
Editar celdas	Transformar...					
Editar columnas	Transformaciones comunes				Quitar espacios al inicio y final	
Transponer	Llenar hacia abajo				Contraer espacios consecutivos	
Ordenar...	Vaciar hacia abajo				Des-escapar entidades HTML	
Ver	Dividir celdas multi-valuadas...				A Tipo oración	
Cotejar	Unir celdas multi-valuadas...				A mayúsculas	
Plantae	Magnolio				A minúsculas	
Plantae	Magnolio				A número	
Plantae	Magnoliophyta	Magnoliopsida	Asterales		A fecha	
Plantae	Magnoliophyta	Magnoliopsida	Asterales		A texto	
					Establecer celdas en nulo	
					A la cadena vacía	

Figura 15

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 ▼ azul > Facetas > Faceta de texto). Luego, en la faceta, ordene los valores por número de registros asociados (seleccionando "conteo"). Verá entonces los valores que se encuentran en este campo como se muestra en la Figura 16.

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.

stateProvince		cambiar
478 choices Ordenar por: A-Z conteo		Agrupar
Buenos Aires	2820	↖
Jujuy	2217	
Buenos Aires	1804	↖
Salta	1265	
Córdoba	1151	
Misiones	992	
Corrientes	961	
Mendoza	674	
Neuquén	578	
Catamarca	515	
Río Negro	508	
Chubut	492	

Figura 16

Para corregir este error, sobre la columna "stateProvince", haga click sobre la ▼ azul y siga la siguiente ruta (Figura 17):

Editar celdas > Transformaciones Comunes > Contraer espacios consecutivos

Esta función le permite convertir múltiples espacios en blanco en un único espacio en blanco.

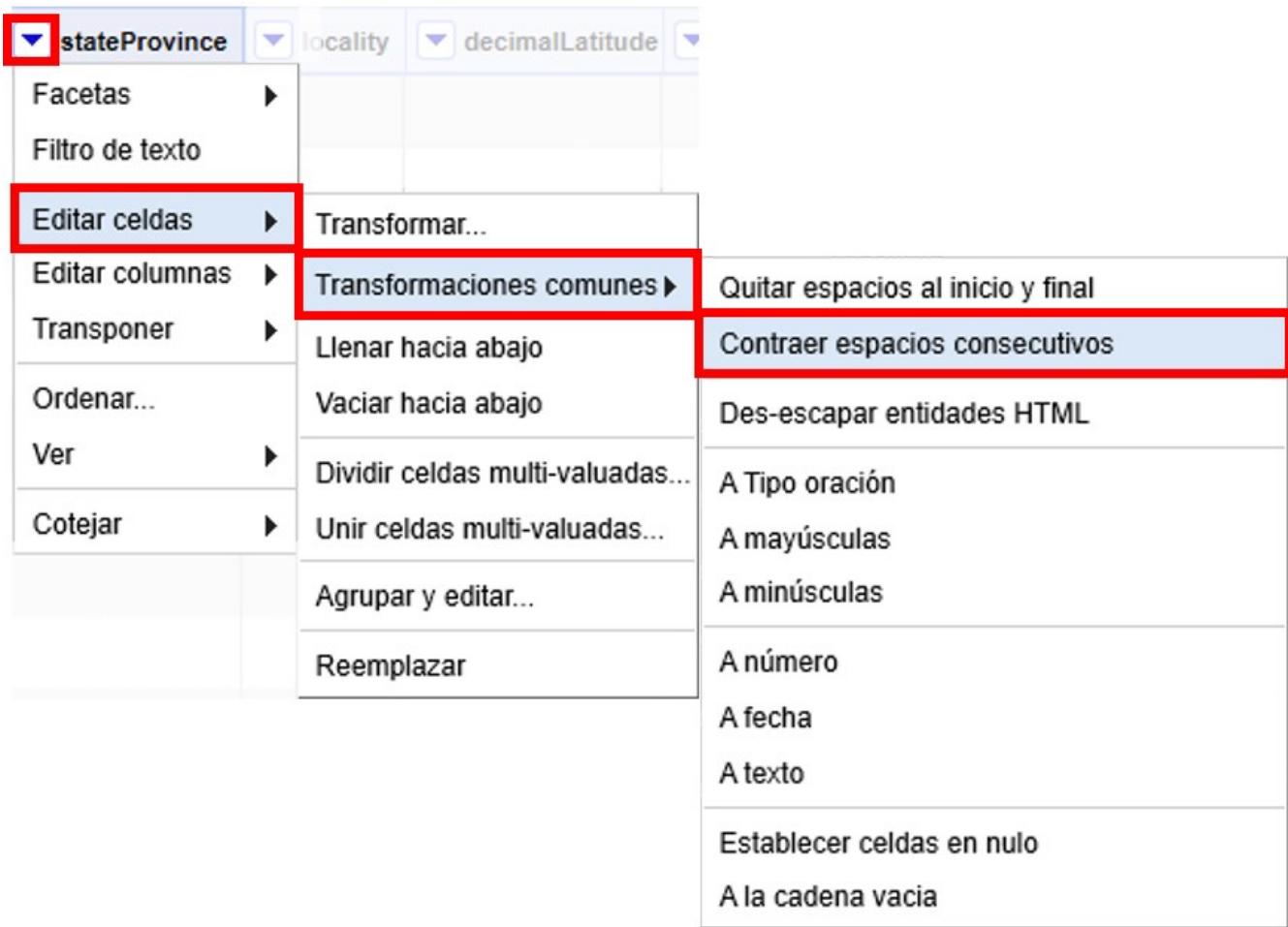


Figura 17

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.

2.2.3. Facetas numéricas

Las facetas también pueden aplicarse a campos numéricos, y en ese caso son muy útiles para, por ejemplo detectar valores fuera de rangos de interés.

A modo de ejemplo, armaremos una faceta numérica sobre el campo "`day`" que hemos creado más arriba. Para ello, hacer click en la ▼ azul del campo y seguir la ruta:

Facetas > Faceta numérica

Verá entonces una nueva ventana, la faceta, como se muestra en la Figura 18.

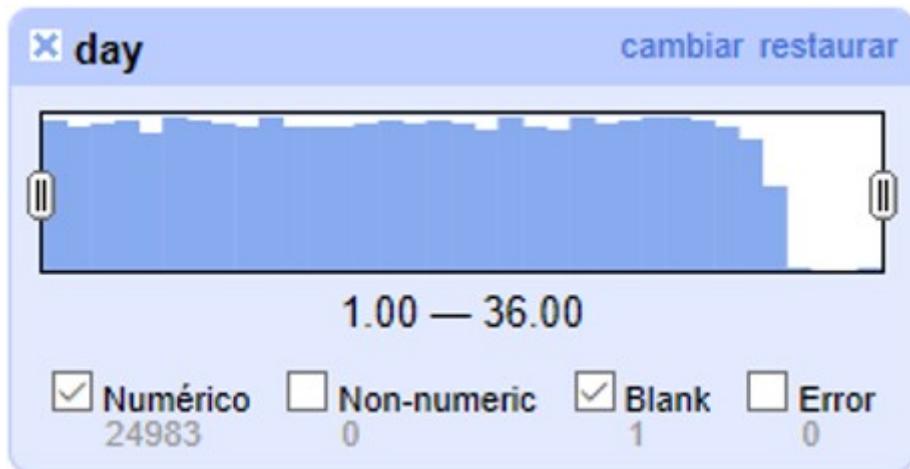


Figura 18

Allí se puede ver que el rango de días va desde 1 a 35 inclusive. Es decir, algunos números están fuera de rango, puesto que como máximo puede haber hasta día 31 en algunos meses.

Se pueden seleccionar los registros con los valores superiores desplazando el botón a la izquierda del rango hacia la derecha. Ello incluirá en la tabla los registros por encima del rango seleccionado y, si no desmarca la opción “Blank”, también los blancos, como se muestra en la Figura 19 (en el ejemplo, tres filas en total: un caso con día 32, un caso con día 35 y un caso con día vacío). Si hubiera valores en el campo que no son numéricos, también podría verlos utilizando esta faceta.

Facetas / Filtros		3 matching filas (24984 total)							
		Mostrar como: filas registros Mostrar: 5 10 25 50 filas							
		sectionCode	catalogNumber	recordedBy	recordNumber	eventDate	month	day	year
<input checked="" type="checkbox"/>	day	160	Spegazzini, Carlos Luigi (Carlos Luis)	160	2/35/1982	2	35	1982	
<input checked="" type="checkbox"/>		762	Schulz, August Gustavo	762	5/32/1979	5	32	1979	
<input checked="" type="checkbox"/>		3943	Mazzucchi, A. G.	3943	2/1/1978	2		1978	

Figura 19

Los tres errores encontrados deben ser consultados con la información original de los ejemplares en la colección, y los campos de fecha estrictamente deberían quedar vacíos para estos registros. Una opción es marcar estos registros para revisar más adelante, usando estrellas o banderas (ver sección sobre uso de estrellas y banderas).

2.2.4. Facetas y duplicados

Las facetas también permiten la detección y corrección de duplicados.



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 ▼ azul y luego siga la siguiente ruta:

Facetas > Facetas personalizadas > Faceta por duplicados

Verá entonces una ventana con la faceta, como se muestra en la Figura 20, donde "true" ("verdadero") refiere a los valores duplicados.

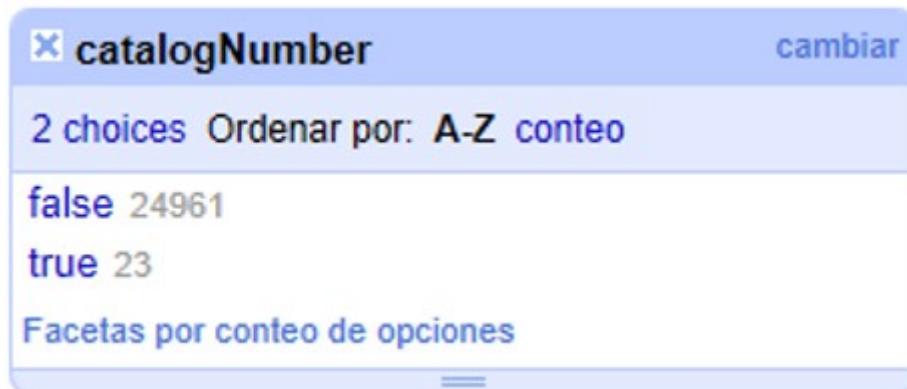


Figura 20

Si hace click en "true", la pantalla principal le mostrará los registros que tienen número de catálogo duplicado (Figura 21). 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 25 primeros) no tiene número de catálogo (el valor nulo es lo que está duplicado).
- etc.

Facetas / Filtros			23 matching filas (24984 total)		
			Mostrar como: filas registros Mostrar: 5 10 25 50 filas		
			<input type="checkbox"/> institutionCode	<input type="checkbox"/> collectionCode	<input type="checkbox"/> catalogNumber
<input checked="" type="checkbox"/> catalogNumber cambiar invertir restaurar			fcnym.unlp.edu.ar	herb	5567
2 choices Ordenar por: A-Z conteo			fcnym.unlp.edu.ar	herb	13305
false 24961			fcnym.unlp.edu.ar	herb	2246
true 23 exclude			fcnym.unlp.edu.ar	herb	5567
Facetas por conteo de opciones			fcnym.unlp.edu.ar	herb	4677
			fcnym.unlp.edu.ar	herb	4677
			fcnym.unlp.edu.ar	herb	4978
			fcnym.unlp.edu.ar	herb	4978
			fcnym.unlp.edu.ar	herb	1697

Figura 21

Corrija los números de catálogo. Para hacerlo, edite las celdas individualmente: sobre la celda haga click en el botón “editar”, modifique el valor y haga click en “Aplicar” (Figura 22).



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.

23 matching filas (24984 total)			
Mostrar como: filas registros Mostrar: 5 10 25 50 filas			
	institutionCode	collectionCode	catalogNumber
Tipo de Dato: texto		5567	edit
5567		13305	
Aplicar	Aplicar a todas las celdas iguales	Cancelar	
Aceptar	Ctrl-Enter	Cancelar	
	fcnym.unlp.edu.ar	herb	5567
	fcnym.unlp.edu.ar	herb	4677
	fcnym.unlp.edu.ar	herb	4677
	fcnym.unlp.edu.ar	herb	4978
	fcnym.unlp.edu.ar	herb	4978
	fcnym.unlp.edu.ar	herb	1697

Figura 22

2.2.5. 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.

Por ejemplo, si tiene configurado el valor a 2000 y trata de armar una faceta de texto en el campo "[specificEpithet](#)", verá que a la derecha la faceta no muestra los valores esperados sino un mensaje que dice que hay demasiados valores para mostrar (Figura 23a).

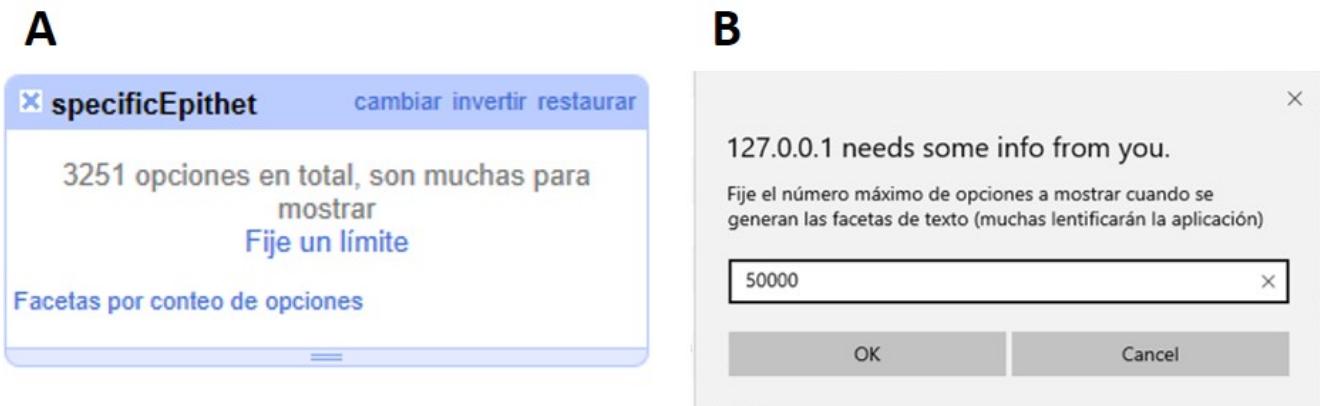


Figura 23

Haciendo click en "Fije un límite", se abrirá otra ventana donde puede cambiar el límite al valor preferido (**Figura 23b**).

Una vez que haya cambiado el valor límite, y si este valor es lo suficientemente grande, podrá ver todos los valores en la faceta del campo de interés (en el ejemplo anterior, el campo "**specificEpithet**").

Alternativamente, para modificar en cualquier momento el límite en el número de valores que se pueden desplegar por faceta, puede ir a la siguiente dirección en su navegador web:

<http://127.0.0.1:3333/preferences>

El navegador mostrará una ventana como ciertas opciones (**Figura 24a**). Allí, establezca el límite preferido para las facetas editando la clave "ui.browsing.listFacet.limit". Para ello haga click en "core-index/edit", y en la ventana que se abre, coloque el nuevo valor límite y oprima "OK" (**Figura 24b**).

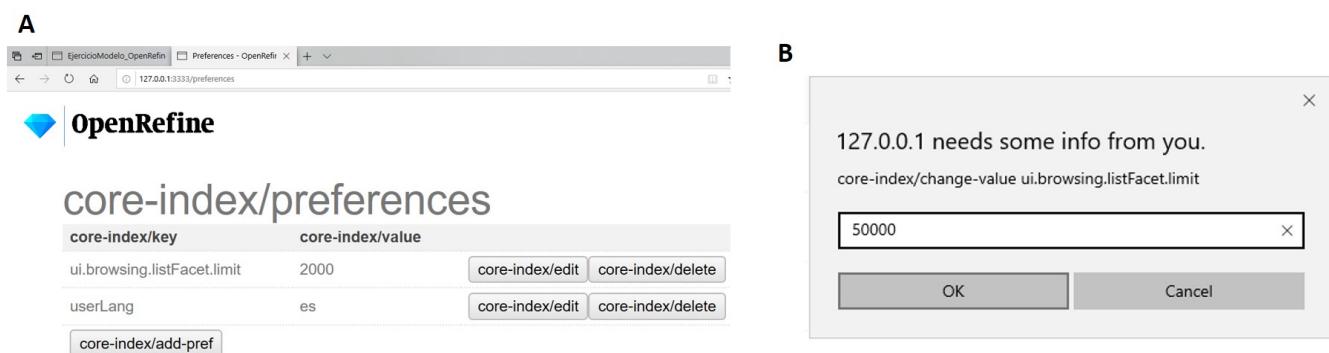


Figura 24

2.3. Uso de Filtros

2.3.1. 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 ▼ azul > **Facetas** > **Faceta de texto**). Luego vaya nuevamente a la ▼ azul y cree un filtro de texto ("Filtro de texto").

Sobre el menú de la izquierda se abrirá una ventana como la que se muestra en la Figura 25.

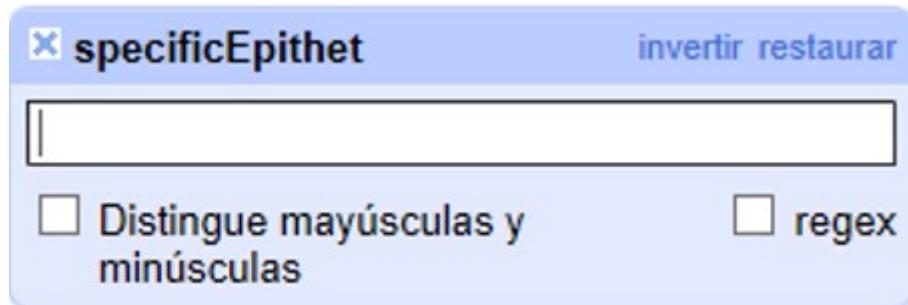


Figura 25

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 26).

Facetas / Filtros		102 matching filas (24984 total)				
		Mostrar como: filas registros Mostrar: 5 10 25 50 filas				
		scientificNameA	genus	concat_scientific	specificEpithet	infraspecificEpit
		Muhlenbergia	Muhlenbergia sp.	sp.		
		Muhlenbergia	Muhlenbergia sp.	sp.		
		Sporobolus	Sporobolus sp.	sp.		
		Phleum	Phleum sp.	sp.		
		Phleum	Phleum sp.	sp.		
		Phleum	Phleum sp.	sp.		
		Croton	Croton sp.	sp.		
		Muhlenbergia	Muhlenbergia sp.	sp.		

Figura 26

Note que verá dos valores, uno en letra minúscula y otro en letra mayúscula. Si sólo desea ver los valores escritos con minúscula, en el filtro debe seleccionar "Distingue mayúsculas y minúsculas", o puede seleccionar "sp." directamente sobre la faceta de "**specificEpithet**".

Corrija los valores "sp." y "SP." utilizando la función "editar" 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 27.

Facetas / Filtros
91 matching filas (24984 total)

Deshacer / Rehacer 13 / 13
Mostrar como: filas registros Mostrar: 5 10 25 50 filas

Actualizar
Restablecer todos
Remover todos

scientificName cambiar

7 choices Ordenar por: A-Z conteo Agrupar

- [Aegiphila sp.](#) 2
- [Croton sp.](#) 1
- [Ctenitis sp.](#) 1
- [Muhlenbergia sp.](#) 18
- [Nassella sp.](#) 1
- [Phleum sp.](#) 22
- [Sporobolus sp.](#) 46

[Facetas por conteo de opciones](#)

scientificName invertir restaurar

×

Distingue mayúsculas y minúsculas regex

<input type="button" value="▼ class"/>	<input type="button" value="▼ order"/>	<input type="button" value="▼ family"/>	<input type="button" value="▼ scientificName"/>
.iliopsida	Poales	Poaceae	Muhlenbergia sp.
.iliopsida	Poales	Poaceae	Muhlenbergia sp.
.iliopsida	Poales	Poaceae	Sporobolus sp.
.iliopsida	Poales	Poaceae	Phleum sp.
.iliopsida	Poales	Poaceae	Phleum sp.
.iliopsida	Poales	Poaceae	Phleum sp.
Magnoliopsida	Euphorbiales	Euphorbiaceae	Croton sp.
.iliopsida	Poales	Poaceae	Muhlenbergia sp.
.iliopsida	Poales	Poaceae	Sporobolus sp.
.iliopsida	Poales	Poaceae	Phleum sp.

Figura 27

Debe 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 ▼ azul en "scientificName" > Editar celdas > Transformar..." (Figura 28).

scientificName	scientificNameA	genus
Facetas		Muhlenbergia
Filtro de texto		Muhlenbergia

Menu options:

- Editar celdas ► Transformar...
- Editar columnas ► Transformaciones comunes ►
- Transponer ► Llenar hacia abajo
- Ordenar... ► Vaciar hacia abajo
- Ver ► Dividir celdas multi-valuadas...
- Cotejar ► Unir celdas multi-valuadas...
- Agrupar y editar...
- Reemplazar

Figura 28

Se abrirá entonces una ventana como la mostrada en la Figura 29. 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 29](#) puede observar cómo se vería el resultado del cambio en la pestaña "Vista previa".

Transformación personalizada en `scientificName`

Expresión Lenguaje General Refine Expression Language (GREL) ▾

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

No hay error de sintaxis.

Vista previa Historial Con estrella Ayuda

row	value	value.replace(" sp.", "")
1607.	Muhlenbergia sp.	Muhlenbergia
1608.	Muhlenbergia sp.	Muhlenbergia
2422.	Sporobolus sp.	Sporobolus
2555.	Phleum sp.	Phleum
2556.	Phleum sp.	Phleum
2557.	Phleum sp.	Phleum
2558.	Ostrea	Ostrea

En error mantener original cambiar a en blanco guardar error

Re-transformar hasta 10 veces hasta que no haya cambios

Aceptar Cancelar

Figura 29

Oprima “Aceptar” para ejecutar la transformación, y verá que en la faceta que ha sido filtrada ya no hay registros que contengan “sp.” como parte del valor en el campo “`scientificName`”.

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

2.3.2. Filtros con expresiones regulares

Los filtros se pueden utilizar también incluyendo expresiones regulares, que permite buscar ciertos patrones en los valores de los campos. Por ejemplo, se pueden buscar palabras que comiencen con ciertas letras, o que comiencen con mayúscula o minúscula, etc.

A modo de ejemplo, buscaremos valores en el campo “`genus`” que comiencen con minúscula. Para ello, abra una faceta y un filtro de texto para el campo “`genus`”. En el filtro coloque la siguiente expresión en el cuadro de texto: `^[a-z]`, y seleccione las opciones “Distingue mayúsculas y minúsculas” y “regex” (Figura 30a). Con dicha expresión se pueden buscar los valores en los que la primera letra es minúscula.

A

genus invertir restaurar

Distingue mayúsculas y minúsculas regex

B

genus cambiar

2 choices Ordenar por: A-Z conteo Agrupar

gentianella 24
stratiotes 1

Facetas por conteo de opciones

Figura 30

Siguiendo estos pasos, debería poder ver dos valores (Figura 30b). Corrija estos 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/>. 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 carácter 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 carácter 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 carácter (*) 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 carácter (*), 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 [<https://github.com/OpenRefine/OpenRefine/wiki>] en GitHub.

2.4. Uso de Agrupamientos

2.4.1. Agrupamientos simples

Los agrupamientos permiten, como su nombre lo indica, agrupar valores de acuerdo a diferentes criterios. Por ejemplo, pueden agruparse valores de acuerdo al grado de similitud 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 "Agrupar". Se abrirá entonces una ventana como la mostrada en la [Figura 31](#).

Allí verá que algunos valores que son similares han sido agrupados por un algoritmo. El método y la función utilizados se muestran y se pueden modificar arriba de la lista de valores.

La ventana también muestra el tamaño del clúster ("Número de valores" agrupados), cuántos registros hay por cluster ("Número de filas") y por valor (entre paréntesis junto a los valores en "Valores en la agrupación").

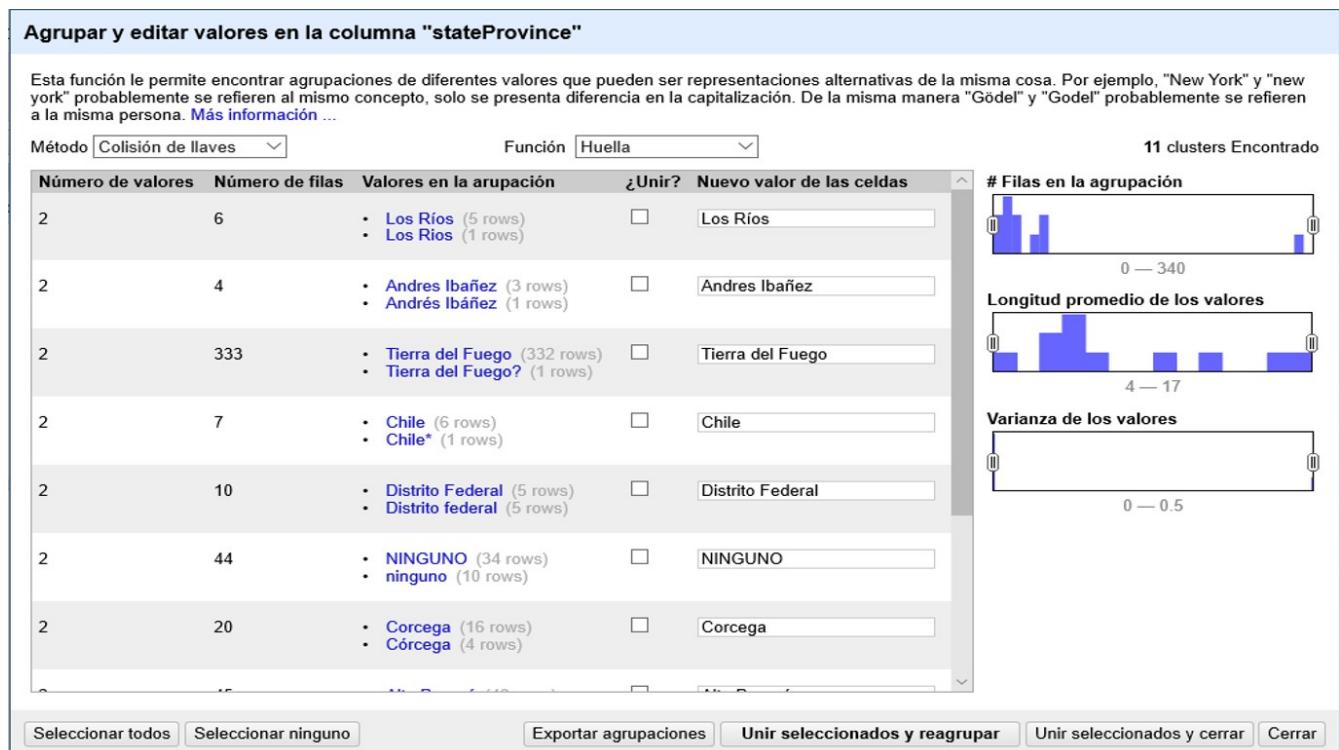


Figura 31

Además, para cada cluster verá una opción para fusionar los valores ("¿Unir?") y el nuevo valor que se asignará a todos los registros del cluster ("Nuevo valor de las celdas").

OpenRefine asigna de forma predeterminada como nuevo valor aquel que presenta mayor número de registros asociados. Esto no es necesariamente correcto. Por ejemplo, en el caso "Corcega" y "Córcega" el valor correcto lleva tilde. Puede modificar el nuevo valor al que unificará haciendo click en el valor deseado si está listado o, en caso de ser diferente, editando directamente el campo "Nuevo valor de las celdas". Recuerde que todos los valores dentro de un agrupamiento dado se

unificarán al valor escogido. Por ejemplo, en el caso en que los valores agrupados son “NINGUNO” y “ninguno”, podría agrupar a un nuevo valor vacío (pues “ninguno” no es un valor válido para una provincia).

Explore los valores agrupados por el algoritmo y corrija los que considere apropiados, seleccionando el valor correcto y marcando la casilla “¿Unir?”.

Haga click sobre “Unir seleccionados y reagrupar”.

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 se evalúa el campo epíteto específico, pueden tenerse dos palabras iguales con diferente terminación -um, -us), no necesariamente lo sean (por ejemplo, si se evalúa también el campo género podría encontrarse que esos epítetos se aplican a géneros distintos, y que ambos son válidos). Por ello, si tiene dudas, consulte los registros completos. Y si aún tiene dudas, consulte en la colección. Otro ejemplo en que debe tenerse extremo cuidado es cuando se agrupan valores que difieren en el orden de las palabras. Un ejemplo típico se da en el campo de colectores. Aún cuando los agrupamientos pueden sugerir que “Colector A y Colector B” es lo mismo que “Colector B y Colector A”, ello puede no ser cierto, y el orden de los colectores puede tener en sí un valor particular. Nuevamente, antes de unificar, es fundamental consultar con la colección.

Una vez resueltos los agrupamientos, si ha decidido no agrupar algunas de las opciones, las verá nuevamente en el re-agrupamiento; en caso contrario, el programa le indicará que no se han encontrado agrupaciones con el método seleccionado. Puede cambiar el método y la función que se utiliza para agrupar escogiendo entre las opciones del menú, como se muestra en la [Figura 32](#).

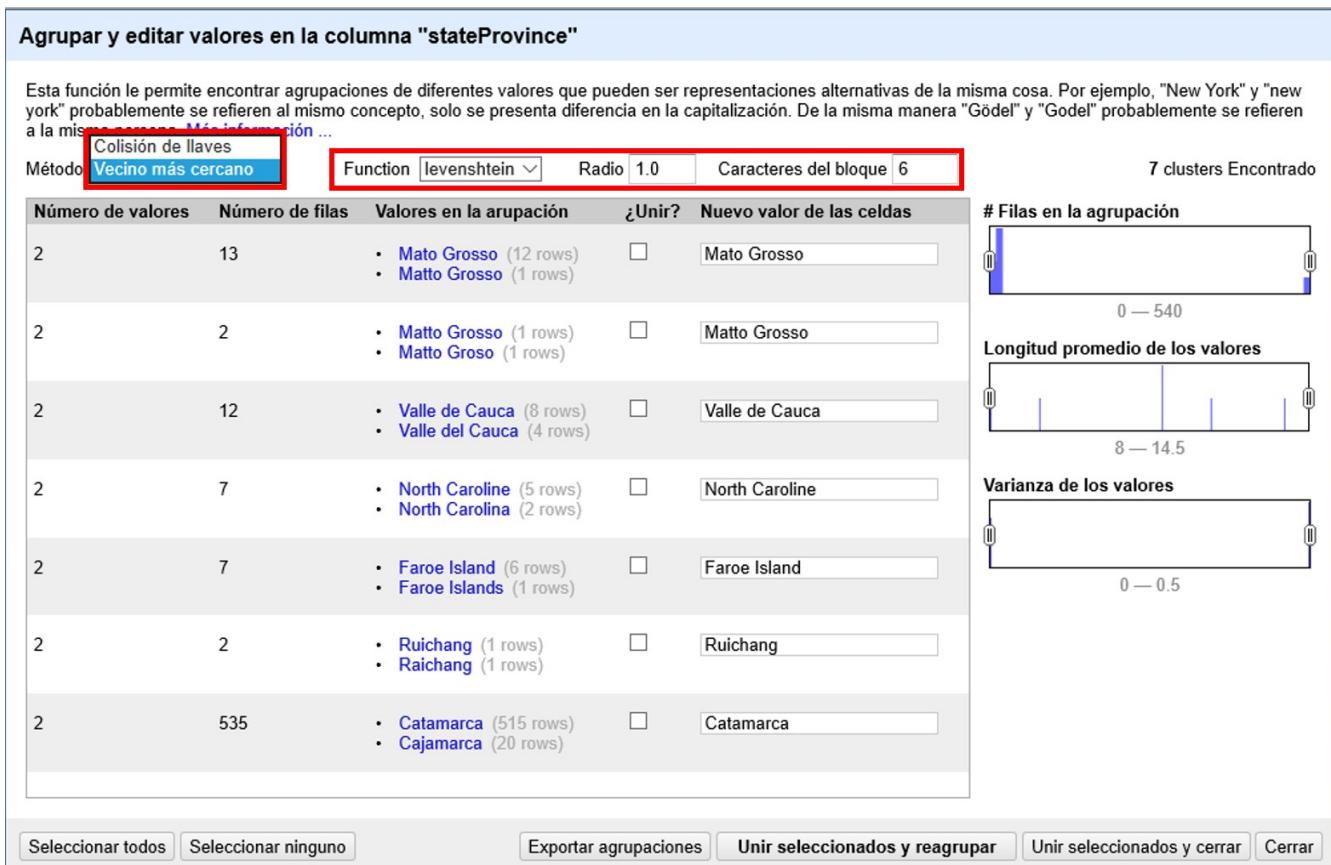


Figura 32

Pruebe agrupamientos con distintos métodos para limpiar los datos.

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

2.5. 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 arriba a la izquierda, abra la pestaña “Deshacer/Rehacer”, que está asociada a un número que indica número de cambios acumulados hasta ahora. Verá entonces una lista de pasos realizados, como se muestra en la Figura 33a.

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.

2.5.1. 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 pasos a partir del paso 5, haga click en el paso 5, y los todos los posteriores se revertirán automáticamente (Figura 33b).

A Deshacer / Rehacer 11 / 18

Extraer... Aplicar...

Filtrar:

0. Create project
1. Reorder columns
2. Create new column CampoDePrueba based on column occurrenceID by filling 0 rows with grel:null
3. Create new column concat_scientificName based on column genus by filling 22174 rows with grel:cells["genus"].value + " " + cells["specificEpithet"].value
4. Split 24984 cell(s) in column eventDate into several columns by separator
5. Rename column eventDate 1 to month
6. Rename column eventDate 2 to day
7. Rename column eventDate 3 to year
8. Mass edit 3 cells in column kingdom
9. Mass edit 17 cells in column kingdom
10. Text transform on 17 cells in column kingdom: value.trim()
11. Text transform on 1809 cells in column stateProvince: value.replace(/\s+/,'')

B Deshacer / Rehacer 5 / 18

Extraer... Aplicar...

Filtrar:

0. Create project
1. Reorder columns
2. Create new column CampoDePrueba based on column occurrenceID by filling 0 rows with grel:null
3. Create new column concat_scientificName based on column genus by filling 22174 rows with grel:cells["genus"].value + " " + cells["specificEpithet"].value
4. Split 24984 cell(s) in column eventDate into several columns by separator
5. **Rename column eventDate 1 to month**
6. Rename column eventDate 2 to day
7. Rename column eventDate 3 to year
8. Mass edit 3 cells in column kingdom
9. Mass edit 17 cells in column kingdom
10. Text transform on 17 cells in column kingdom: value.trim()
11. Text transform on 1809 cells in column stateProvince: value.replace(/\s+/,'')

Figura 33

Para rehacer un paso luego de haberlo deshecho, simplemente haga click en ese paso, teniendo en cuenta que entonces se llevarán a cabo todos los pasos intermedios también.



El hacer y deshacer en OpenRefine trabaja sobre “estados”. Eso quiere decir que se 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 se vuelve a un estado anterior y luego se realiza una nueva modificación a partir de ese estado, entonces perderá los pasos originales y no podrá recuperarlos. En el ejemplo de la Figura 33, si se vuelve al paso 5 y luego realiza sobre los datos alguna otra operación, no podrá volver a los pasos 6 a 11 previos.

2.5.2. Guardar pasos para rehacer luego

Es importante entonces que guarde sus pasos, especialmente para aquellos procesos más complejos. Para ello, en la pestaña “Deshacer/Rehacer”, haga click en el botón “Extraer...”. Se abrirá una nueva ventana, como se muestra en la Figura 34, donde pueden seleccionar los pasos que desea guardar. Los pasos están dados en formato JSON en el panel de la derecha.

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.

Puede marcar y desmarcar pasos en el panel de la izquierda para seleccionar los pasos de interés. 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).

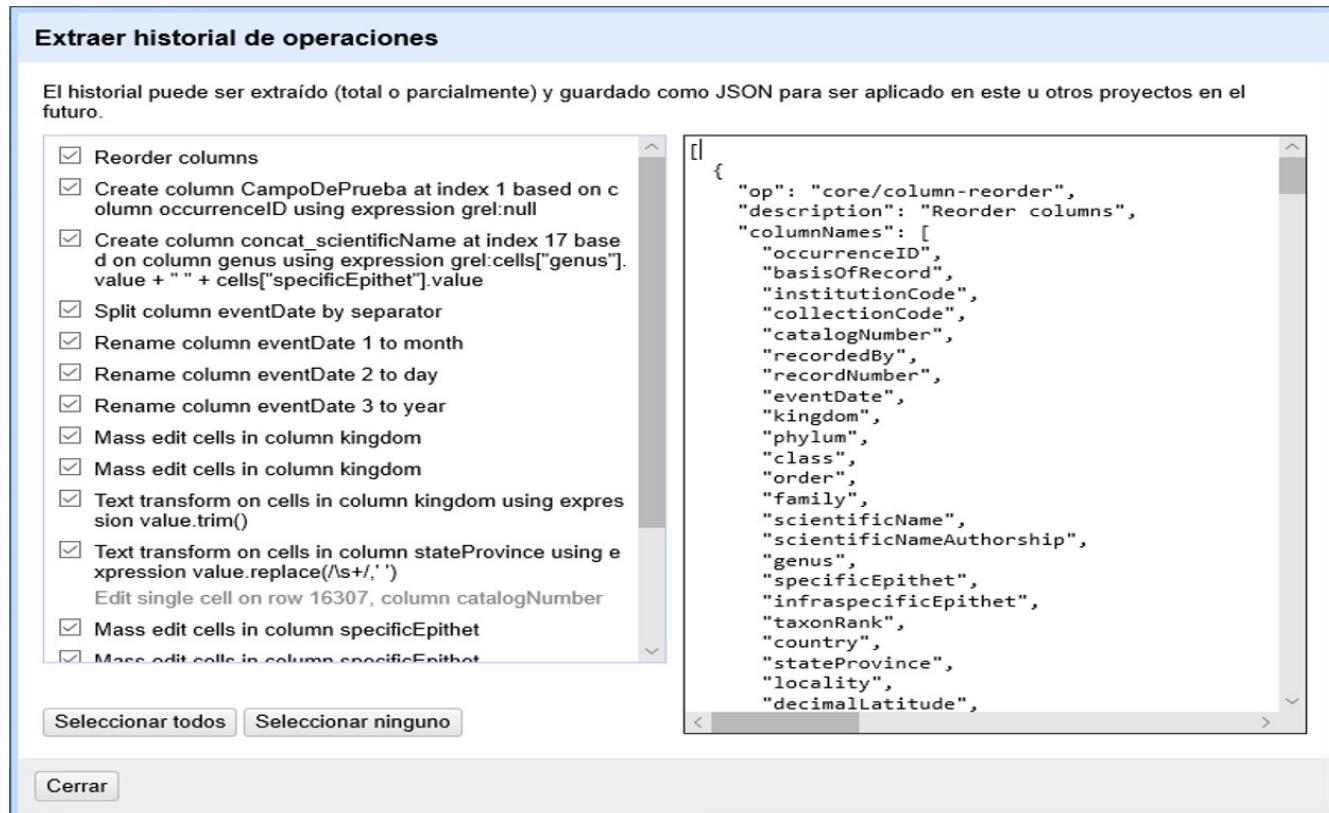


Figura 34



Los cambios hechos a celdas particulares no tienen la opción de guardar expresiones. En el ejemplo anterior, Figura 34, 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””).

2.5.3. Rehacer pasos guardados

Si desea rehacer pasos que tenga guardados (en formato JSON), dentro de la pestaña “Deshacer/Rehacer” haga click en el botón “Aplicar...”. Se abrirá entonces una ventana como la que se muestra en la Figura 35, 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 “Ejecutar Operaciones”.

Aplicar historial de operaciones

Pegue un historial de operaciones extraido en JSON para que sea ejecutado:

```
[  
  {  
    "op": "core/column-split",  
    "description": "Split column eventDate by separator",  
    "engineConfig": {  
      "facets": [],  
      "mode": "row-based"  
    },  
    "columnName": "eventDate",  
    "guessCellType": true,  
    "removeOriginalColumn": false,  
    "mode": "separator",  
    "separator": "/",  
    "regex": false,  
    "maxColumns": 0  
  }  
]
```

Ejecutar Operaciones

Cancelar

Figura 35

De este modo, puede rehacer pasos particulares o toda una rutina de trabajo, sobre el mismo conjunto de datos, o sobre otros conjuntos de datos (siempre y cuando las columnas sean las mismas).

2.6. 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, aunque el marcado se registra como un cambio en el historial de cambios del proyecto, 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.



2.6.1. Marcado con banderas y estrellas

Las banderas y estrellas se encuentran dentro del campo “Todo”. 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 ▼ azul del campo → Facetas → Faceta de texto).

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 ▼ azul del campo "Todo" y siga la siguiente ruta ([Figura 36](#)):

Editar filas > Marcar filas con bandera



Figura 36

Una vez que lo haya hecho, verá que todos los registros seleccionados están marcados ahora con una bandera.

Para desmarcar todos esos registros, puede hacer click en la ▼ azul del campo "Todo" y seguir la ruta:

Editar filas > Desmarcar filas con bandera

Para marcar y desmarcar registros con estrellas, siga el mismo procedimiento con "estrellas" en lugar de "banderas".

2.6.2. Conservación de banderas y estrellas en la exportación

Si desea marcar los registros de modo que al exportar se conserven 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 ▼ azul → Editar columnas → Agregar columna basada en esta columna...

Se abrirá una ventana como la mostrada en la [Figura 37](#). Asigne un nombre al campo. Por ejemplo, si sus banderas significan que ha detectado errores en los registros, puede llamarlo “tieneError”.

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 oprimir “Aceptar” su campo se habrá creado. Verifique los valores que toma asignando a algunos registros una bandera.

row	value
324.	urn:catalog:fcnym.unlp.edu.ar:herb:001065
3452.	urn:catalog:fcnym.unlp.edu.ar:herb:000983
3453.	urn:catalog:fcnym.unlp.edu.ar:herb:000979
3799.	urn:catalog:fcnym.unlp.edu.ar:herb:000987
5130.	urn:catalog:fcnym.unlp.edu.ar:herb:000980
5131.	urn:catalog:fcnym.unlp.edu.ar:herb:000981
E122	urn:catalog:fcnym.unlp.edu.ar:herb:000001

Figura 37

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

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

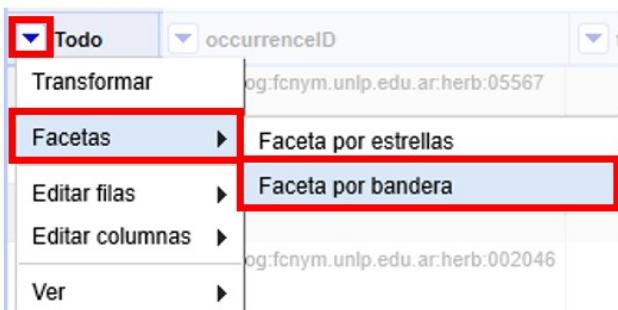
Para ver los pasos de exportación de datos, vea la sección de [Exportación de proyectos](#).

2.6.3. 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 ▼ azul sobre el campo “Todo” > Facetas > Faceta por bandera** ([Figura 38a](#)).
3. En esta nueva faceta, a la izquierda, seleccione la opción “true” haciendo click sobre ella. Ello le mostrará los registros a los que se ha asignado una bandera.
4. Haga click nuevamente sobre **la ▼ azul del campo “Todo” > Editar filas > Eliminar todas las filas que encajen** ([Figura 38b](#)).

A



B



Figura 38

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 vayan a verse reflejados automáticamente en su base de datos original. Para ello, deberá exportar los datos desde OpenRefine e importarlos nuevamente a su base de datos.

3.1. 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 encontrará 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

MacOS:

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

Linux:

- ~/.local/share/openrefine/

3.2. 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 “Exportar” ([Figura 39](#)).

Note que la primera opción, “Exportar proyecto”, permite exportar el proyecto completo, mientras que otras opciones (e.g., delimitado por..., Excel, etc.) permiten exportar los datos.

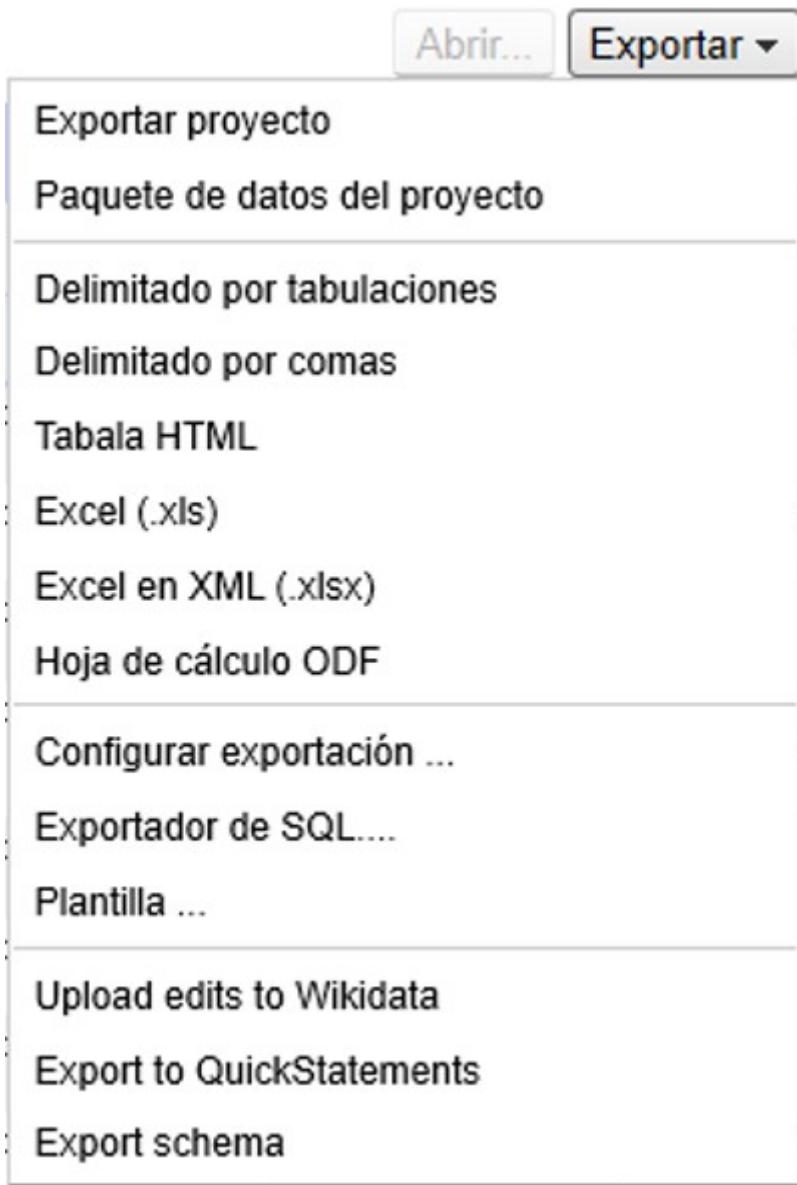


Figura 39

La exportación de proyectos es útil cuando uno quiere abrir el mismo proyecto en OpenRefine en otra computadora.

Haciendo click en “Exportar proyecto” se abrirá una ventana en la que puede escoger si exportar como archivo local o si exportar a Google Drive ([Figura 40](#)).

Por favor, elija el destino para la exportación del proyecto

- Exportar a local
- Exportar a Google Drive

Exportar

Cancelar

Figura 40

Escoja la opción deseada y haga click en “Exportar”. El archivo exportado tendrá una extensión .tar.gz, 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.



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 el menú “Exportar” escoja “Configurar exportación...”. Se abrirá una ventana como la mostrada en la [Figura 41](#), en la cual puede escoger una serie de opciones.

Exportador Tabular Personalizado

Contenido Descarga Cargar Código

Seleccione y ordene las columnas a exportar

occurrenceID
 tieneError
 CampoDePrueba
 basisOfRecord
 institutionCode
 collectionCode
 catalogNumber
 recordedBy
 recordNumber

Opciones para occurrenceID

Para celdas cotejadas, descargar

Nombre cotejado El contenido de las celdas
 ID cotejado Descargar nada para celdas sin correspondencia
 Enlace a la página cotejada ISO 8601, p. ej., 2011-08-24T18:36:10+08:00
 Formato corto local Formato mediano local
 Formato largo local Formato local completo
 Personalizado Usar zona horaria local Omitir hora/fecha

Ayuda

Seleccionar todos De-seleccionar todos

Incluir encabezados de columnas Incluir filas en blanco (p. ej. todas las celdas nulas) Ignorar facetas y filtros filters y exportar todas las filas

[Cancelar](#)

Figura 41

En la pestaña “Contenido” puede elegir qué campos exportar y modificar ciertos parámetros para cada campo individualmente.

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 “Descarga”, como se ve en la Figura 42.

Exportador Tabular Personalizado

Contenido Descarga Cargar Código

Formatos de texto

Delimitado por tabulaciones (TSV)
 Delimitado por comas (CSV)
 Otro delimitador \t

Separador de línea \n

Codificación de caracteres UTF-8

Otros formatos

Excel (.xls)
 Excel en XML (.xlsx)
 Tabla HTML

[Vista previa](#) [Descargar](#)

[Cancelar](#)

Figura 42

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

También, para ver una vista previa de los datos que descargará, puede hacer click en "Vista previa", 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. Para el caso particular de datos sobre biodiversidad, 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. En esta guía sólo se incluyen los métodos referidos a las consultas a través de URLs. Para ver explicaciones referidas al uso de algunos servicios de reconciliación consultar versiones anteriores de este documento; tener en cuenta que esos servicios no han sido actualizados en concordancia con las actualizaciones de OpenRefine, y muchos no funcionan a partir de OpenRefine 2.8.



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



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.

4.1. 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.

4.1.1. 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>].

Para acortar el tiempo de consulta, cree una faceta para el campo "**genus**" (click en **la ▼ azul > Facetas > Faceta de texto**) 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).

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 haciendo click en la ▼ azul del campo y siguiendo la siguiente ruta (**Figura 43**):

Editar columnas > Agregar columna accediendo a URLs...

The screenshot shows a spreadsheet interface with several columns: scientificName, scientificNameA, genus, and concat_scien. The 'scientificName' column header is highlighted with a red box. A context menu is open over this column, with the 'Editar columnas' option also highlighted in red. The submenu under 'Editar columnas' contains the following items: 'Dividir en varias columnas...', 'Aregar columna basada en esta columna...', 'Aregar columna accediendo a URLs...', 'Añadir columnas de valores conciliados...', 'Renombrar esta columna', 'Eliminar esta columna', 'Mover columna al principio', 'Mover columna al final', 'Mover columna a la izquierda', and 'Mover columna a la derecha'. The 'Aregar columna accediendo a URLs...' option is currently selected.

scientificName	scientificNameA	genus	concat_scien
Facetas	Iter	Cinna	Cinna lateralis
Filtro de texto		Cinna	Cinna arundinacea
Editar celdas		Cinna	Cinna_arundinacea
Editar columnas	Dividir en varias columnas...		
Transponer	Aregar columna basada en esta columna...		
Ordenar...	Aregar columna accediendo a URLs...		
Ver	Añadir columnas de valores conciliados...		
Cotejar	Renombrar esta columna		
Cinna arundinacea	Eliminar esta columna		
Cinna arundinacea	Mover columna al principio		
Cinna latifolia	Mover columna al final		
	Mover columna a la izquierda		
	Mover columna a la derecha		

Figura 43

Se abrirá una ventana como la mostrada en la Figura 44. 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". Es importante que el nombre del campo que utiliza en la expresión sea idéntico al nombre del campo del cual tomará los valores originales, o de otro modo el llamado será infructuoso.

Agregar columna accediendo a URIs basada en la columna scientificName

Nuevo nombre de la columna	GNR_Json_sciName	Tiempo de retraso	5000 milisegundos																							
En error	<input checked="" type="radio"/> cambiar a en blanco <input type="radio"/> guardar error	<input checked="" type="checkbox"/> core-views/cache-responses																								
Cabeceras HTTP que se utilizarán cuando se obtengan URLs: Mostrar																										
Ingrese las URLs a acceder:																										
Expresión	Lenguaje	General Refine Expression Language (GREL) <input type="button" value="▼"/>																								
<pre>"http://resolver.globalnames.org/name_resolvers.json? names="+escape(cells["scientificName"].value,"url")</pre>			No hay error de sintaxis.																							
<table border="1"> <tr> <td>Vista previa</td> <td>Historial</td> <td>Con estrella</td> <td>Ayuda</td> </tr> <tr> <td colspan="4"> <table border="1"> <thead> <tr> <th>row</th> <th>value</th> <th>...</th> </tr> </thead> <tbody> <tr> <td>534.</td> <td>Cinna lateralis</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+lateralis</td> </tr> <tr> <td>535.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> <tr> <td>536.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> <tr> <td>537.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> </tbody> </table> </td> </tr> </table>				Vista previa	Historial	Con estrella	Ayuda	<table border="1"> <thead> <tr> <th>row</th> <th>value</th> <th>...</th> </tr> </thead> <tbody> <tr> <td>534.</td> <td>Cinna lateralis</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+lateralis</td> </tr> <tr> <td>535.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> <tr> <td>536.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> <tr> <td>537.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> </tbody> </table>				row	value	...	534.	Cinna lateralis	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+lateralis	535.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea	536.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea	537.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea
Vista previa	Historial	Con estrella	Ayuda																							
<table border="1"> <thead> <tr> <th>row</th> <th>value</th> <th>...</th> </tr> </thead> <tbody> <tr> <td>534.</td> <td>Cinna lateralis</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+lateralis</td> </tr> <tr> <td>535.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> <tr> <td>536.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> <tr> <td>537.</td> <td>Cinna arundinacea</td> <td>http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea</td> </tr> </tbody> </table>				row	value	...	534.	Cinna lateralis	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+lateralis	535.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea	536.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea	537.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea								
row	value	...																								
534.	Cinna lateralis	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+lateralis																								
535.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea																								
536.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea																								
537.	Cinna arundinacea	http://resolver.globalnames.org/name_resolvers.json? names=Cinna+arundinacea																								
<input type="button" value="Aceptar"/> <input type="button" value="Cancelar"/>																										

Figura 44

Note que en esta ventana, arriba a la derecha, tiene una opción para modificar el "Tiempo de retraso". Este valor indica el tiempo que transcurre entre llamados o consultas sucesivas que se hacen al servicio en cuestión. Por defecto, el valor es de 5000 milisegundos. Puede reducir este tiempo para acelerar el proceso de comparación. Tenga en cuenta, sin embargo, que muchos servicios bloquean los llamados si éstos ocurren muy cercanos en el tiempo, pues consideran que puede tratarse de un ataque. El máximo número de consultas que pueden realizar por unidad de tiempo depende de cada servicio en particular.



La expresión utilizada es muy general, y devolverá los valores de todos los parámetros que GNR provee respecto de un nombre científico. Puede consultar dichos parámetros en <http://resolver.globalnames.org/api>. Si no quiere obtener en el resultado todos los valores, puede modificar la expresión especificando valores para todos o algunos de los parámetros. Por ejemplo: GNR resuelve los nombres consultando diferentes fuentes, a las que asigna identificadores (data_source_id); si sólo quiere obtener los resultados provenientes de la fuente Catalogue of Life (que en GNR tiene id=1), puede utilizar la siguiente expresión:

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

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 (haga click en **la ▼ azul > Editar columnas > Agregar columna basada en esta columna**).

Dé un nombre al nuevo campo (por ejemplo, GNR_sciName_options) y en el cuadro de texto, coloque la siguiente expresión (**Figura 45**):

```
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 **|**.

Agregar columna basada en la columna GNR_Json_sciName

Nuevo nombre de la columna **GNR_sciName_options**

core-views/addasdadsd cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje General Refine Expression Language (GREL)

```
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")), "",
```

No hay error de sintaxis.

Vista previa Historial Con estrella Ayuda

row	value	forEach(value.parseJson().get(...
534.	{"id":"beu78ok4w86m","url":"http://resolver.global	uBio NameBank; Cinna lateralis; Catalogue of
	[],"data":[{"supplied_name_string":"Cinna	Life; Cinna lateralis Walter; Andropogon
	lateralis","is_known_name":true,"results":	virginicus L. ITIS; Cinna lateralis Walter;
	[{"data_source_id":169,"data_source_title":"uBio	Andropogon virginicus L. GBIF Backbone
	NameBank","gni_uuid":"41981160-be42-55d3-	Taxonomy; Cinna lateralis Walter; Andropogon
	8292-6605441a7e28","name_string":"Cinna	virginicus L. EOL; Cinna lateralis Walter;
	lateralis","canonical_form":"Cinna	Tropicos - Missouri Botanical Garden; Cinna
	lateralis","classification_path":["Cinna	lateralis Walter; The International Plant
	lateralis","classification_path_ranks":"kingdom","	Names Index; Cinna lateralis Walter; uBio
	namebankID=10751399","imported_at":"2013-	NameBank; Cinna lateralis Walter; uBio
	05-	NameBank; Cinna lateralis Walter. 1788;

Aceptar Cancelar

Figura 45

Una vez que haya creado el campo, verá que contiene, aún 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;
```



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"
```

y luego arme un nuevo campo extrayendo los resultados de interés, usando la 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(" | ")
```

A partir de los resultados obtenidos, puede extraer los nombres separando la nueva columna en columnas distintas utilizando separadores apropiados (ver sección de separación de columnas).

4.1.2. 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 simultáneamente.

Llevaremos a cabo la georreferenciación a partir del campo "`locality`". Para ello, cree un nuevo campo a partir de éste siguiendo la ruta: click en la ▼ azul > **Editar columnas** > **Agregar columna**

accediendo a URLs....

Se abrirá una nueva ventana (Figura 46). 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.

Agregar columna accediendo a URLs basada en la columna locality

Nuevo nombre de la columna **GeoLocate_Json_georref** Tiempo de retraso **5000 milisegundos**

En error cambiar a en blanco guardar error core-views/cache-responses

Cabeceras HTTP que se utilizarán cuando se obtengan URLs: [Mostrar](#)

Ingrese las URLs a acceder:

Expresión Lenguaje **General Refine Expression Language (GREL)**

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

No hay error de sintaxis.

Vista previa Historial Con estrella Ayuda

row	value
3452.	Ledesma
3453.	Ledesma
3799.	null
5130.	Yuto

Aceptar Cancelar

row	value
3452.	"http://www.museum.tulane.edu/webservices/geolocatesvcv2/glcwrap.aspx?Country=Argentina&fmt=json&Locality=Ledesma"
3453.	"http://www.museum.tulane.edu/webservices/geolocatesvcv2/glcwrap.aspx?Country=Argentina&fmt=json&Locality=Ledesma"
3799.	"http://www.museum.tulane.edu/webservices/geolocatesvcv2/glcwrap.aspx?Country=Argentina&fmt=json&Locality=null"
5130.	"http://www.museum.tulane.edu/webservices/geolocatesvcv2/glcwrap.aspx?Country=Argentina&fmt=json&Locality=Yuto"

Figura 46

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 tres 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, ①  
  "executionTimems" : 171.6003  
}
```

① Ningún resultado encontrado.

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 ] ①  
        },  
        "properties": {  
          "parsePattern": "YUTO", ②  
          "precision": "High",  
          "score": 79,  
          "uncertaintyRadiusMeters": 3036, ③  
          "uncertaintyPolygon": "Unavailable", ④  
          "displacedDistanceMiles": 0, ⑤  
          "displacedHeadingDegrees": 0,  
          "debug":  
            ":GazPartMatch=False|:inAdm=True|:Adm=JUJUY|:NPExtent=5040|:NP=YUTO|:KFID=|YUTO" ⑥  
        }  
      }  
    ],  
    "crs": { "type": "EPSG", "properties": { "code": 4326 } }  
  }  
}
```

① Las coordenadas: "coordinates": [-64.471941, -23.643418]

② La localidad original que consultó: "parsePattern" : "YUTO"

③ El radio de incertezza en metros: "uncertaintyRadiusMeters" : 3036

④ El polígono de incertezza asociado: "uncertaintyPolygon" : "Unavailable", en este caso no disponible.

- ⑤ 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).
- ⑥ 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, ①
  "executionTimems": 187.2004,
  "resultSet": {
    "type": "FeatureCollection",
    "features": [
      {
        "type": "Feature",
        "geometry": {
          "type": "Point",
          "coordinates": [ -64.158097, -26.21252 ] ②
        },
        "properties": {
          "parsePattern": "TARTAGAL", ③
          "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" ④
        }
      },
      {
        "type": "Feature",
        "geometry": {
          "type": "Point",
          "coordinates": [ -59.846115, -28.671732 ] ②
        },
        "properties": {
          "parsePattern": "TARTAGAL", ③
          "precision": "High",
          "score": 83,
          "uncertaintyRadiusMeters": 3036,
          "uncertaintyPolygon": "Unavailable",
          "displacedDistanceMiles": 0,
          "displacedHeadingDegrees": 0,
          "debug": ":GazPartMatch=False|:inAdm=True|:Adm=SANTA
CLARA|:NPExtent=500|:NP=TARTAGAL|:KFID=|TARTAGAL" ④
        }
      }
    ]
  }
}
```

```

    "FE|:NPExtent=5040|:NP=TARTAGAL|:KFID=|TARTAGAL" ④
    }
  },
  {
    "type": "Feature",
    "geometry": {
      "type": "Point",
      "coordinates": [ -63.801314, -22.516365 ] ②
    },
    "properties": {
      "parsePattern": "TARTAGAL", ③
      "precision": "High",
      "score": 83,
      "uncertaintyRadiusMeters": 3036,
      "uncertaintyPolygon": "Unavailable",
      "displacedDistanceMiles": 0,
      "displacedHeadingDegrees": 0,
      "debug":
":GazPartMatch=False|:inAdm=True|:Adm=SALTA|:NPExtent=5040|:NP=TARTAGAL|:KFID=|TARTAGA
L" ④
    }
  }
],
"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.

Visualizando JSON

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 47](#)). Esto puede ser muy útil a la hora de armar expresiones para desglosar el contenido de los campos en nuevos campos sin perder información.



```
{
  "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": ":GasPartMatch=False|:inAdm=True|:Adm=SANTIAGO DEL ESTERO|NPExtent=500|:NP=TARTAGAL|:KFID=TARTAGAL"
        },
        "crs": {
          "type": "EPSG",
          "properties": {
            "code": 4326
          }
        }
      }
    ]
  }
}
```

```

String parse

object {
  "engineVersion": string "GLC:5.21|U:1.01374|eng:1.0",
  "numResults": number 3,
  "executionTimems": number 187.2004,
  "resultSet": object {
    "type": string "FeatureCollection",
    "features": array {
      object {
        "type": string "Feature",
        "geometry": object {
          "type": string "Point",
          "coordinates": array {
            number -64.158097,
            number -26.21252
          }
        },
        "properties": object {
          "parsePattern": string "TARTAGAL",
          "precision": string "High",
          "score": number 83,
          "uncertaintyRadiusMeters": number 301,
          "uncertaintyPolygon": string "Unavailable",
          "displacedDistanceMiles": number 0,
          "displacedHeadingDegrees": number 0,
          "debug": string ":GasPartMatch=False|:inAdm=True|:Adm=SANTIAGO DEL ESTERO|NPExtent=500|:NP=TARTAGAL|:KFID=TARTAGAL"
        }
      }
    }
  },
  "crs": object {
    "type": string "EPSG",
    "properties": object {
      "code": number 4326
    }
  }
}

```

Figura 47



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:

1. 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.
2. 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/geolocatesvcv2/glcwrap.aspx?country=Argentin
a&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 ▼ azul del campo **"GeoLocate_Json_georref"** > Editar columnas > Agregar columna basada en esta columna".

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, isNonBlank(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 47](#)).

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
```



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 ▼ azul del campo > Editar columnas > Dividir en varias columnas**.

Escoja como separador **|**. Desmarque la opción "Eliminar esta columna" 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 ▼ azul del campo **"GeoLocate_Json_georref"** > Editar columnas > Agregar columna basada en esta columna".

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, isNonBlank(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 ;.

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 crear 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).

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/#location>, y consultar: <https://github.com/tdwg/dwc-qa/wiki/Georeferences>.

4.1.3. Limpieza de fechas utilizando Canadensys Date Parsing

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:

- **AAA**: 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".



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.

Tabla 1. Ejemplos

Fecha original	Fecha estandarizada
12 Feb 1809	1809-02-12
12/02/1809	1809-02-12
Jun 1906	1906-06
1971	1971

Fecha original	Fecha estandarizada
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

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) [<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 48](#)):

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 48

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, primero seleccione algunas fechas mediante una faceta, para reducir el tiempo de consulta. Luego, sobre la columna "`eventDate`" haga click en **la ▼ azul > Editar columna > Agregar columna accediendo a URLs...** ([Figura 49](#)). En la

ventana que aparece, nombre la nueva columna (por ejemplo "Canadensys_eventDate") y pegue en el cuadro de texto la siguiente expresión:

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

Esta expresión le indica a la herramienta que evalúe los valores del campo "eventDate" y que envíe los resultados en formato JSON.

Agregar columna accediendo a URLs basada en la columna eventDate

Nuevo nombre de la columna Tiempo de retraso milisegundos

En error cambiar a en blanco guardar error core-views/cache-responses

Cabeceras HTTP que se utilizarán cuando se obtengan URLs: [Mostrar](#)

Ingrese las URLs a acceder:

Expresión Lenguaje General Refine Expression Language (GREL)

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

No hay error de sintaxis.

Vista previa	Historial	Con estrella	Ayuda	
row	value	'http://data.canadensys.net/to ...'		
3531.	9/19/2013	http://data.canadensys.net/tools/dates.json?data=9%2F19%2F2013		
7150.	6/4/1969	http://data.canadensys.net/tools/dates.json?data=6%2F4%2F1969		
7905.	9/2/1910	http://data.canadensys.net/tools/dates.json?data=9%2F2%2F1910		
12375.	4/29/1994	http://data.canadensys.net/tools/dates.json?data=4%2F29%2F1994		
15264.	9/16/1967	http://data.canadensys.net/tools/dates.json?data=9%2F16%2F1967		
19767.	10/11/1903	http://data.canadensys.net/tools/dates.json?data=10%2F11%2F1903		
22442.	8/16/1907	http://data.canadensys.net/tools/dates.json?data=8%2F16%2F1907		

Aceptar Cancelar

Figura 49



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 50).

eventDate	Canadensys_eventDate
9/19/2013	{"data":{"results":[{"originalValue":"9/19/2013","year":2013,"month":9,"day":19,"iso8601":"2013-09-19","partial":false}]}}
6/4/1969	{"data":{"results":[{"originalValue":"6/4/1969","error":"The date [6-4-1969] could not be precisely determined.","partial":true}]}}

Figura 50

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, extraeremos 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: **Editar columnas > Agregar columna basada en esta columna** (Figura 51).

Para extraer sólo la fecha en formato ISO, en la ventana nombre la nueva columna (por ejemplo, "ISO_eventDate") y en el cuadro de texto pegue la siguiente expresión:

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

Agregar columna basada en la columna Canadensys_eventDate

Nuevo nombre de la columna ISO_eventDate

core-views/addasdads cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje General Refine Expression Language (GREL)

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

No hay error de sintaxis.

Vista previa Historial Con estrella Ayuda

row	value	forEach(value.parseJson().get(...
3531.	{"data":{"results":[{"originalValue":"9/19/2013","year":2013,"month":9,"day":19,"iso8601":"2013-09-19","partial":false}]}}	2013-09-19
7150.	{"data":{"results":[{"originalValue":"6/4/1969","error":"The date [6-4-1969] could not be precisely determined.","partial":true}]}}	null
7905.	{"data":{"results":[{"originalValue":"9/2/1910","error":"The date [9-2-1910] could not be precisely determined."}]}}	null

Aceptar Cancelar

Figura 51

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 52).

▼ eventDate	▼ Canadensys_eventDate	▼ ISO_eventDate
9/19/2013	{"data":{"results":[{"originalValue":"9/19/2013","year":2013,"month":9,"day":19,"iso8601":"2013-09-19","partial":false}]}}	2013-09-19
6/4/1969	{"data":{"results":[{"originalValue":"6/4/1969","error":"The date [6-4-1969] could not be precisely determined.","partial":true}]}}	

Figura 52

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.

Epílogo

Agradecimientos

Los autores agradecen especialmente a Anabela Plos (GBIF Argentina, Museo Argentino de Ciencias Naturales “Bernardino Rivadavia”) y David Bloom (VertNet) por sus comentarios sobre versiones anteriores de esta guía. Esta guía ha sido actualizada en el marco de la iniciativa de actualización de documentación del Secretariado del Global Biodiversity Information Facility (GBIF). Este documento se ha actualizado en parte a partir de su uso en distintos cursos de entrenamiento organizados por diversas instituciones, entre las que se cuentan GBIF Argentina - Museo Ciencias Naturales “Bernardino Rivadavia”, y GBIF España - Real Jardín Botánico de Madrid, incorporando los comentarios de los respectivos alumnos. Se extienden los agradecimientos a otras personas que también contribuyeron de una u otra manera en esta versión del documento: Katia Cezón, Kyle Copas, Melanie Raymond, members of the GBIF Documentation Editorial Panel, OTHEr .

Apéndice 1: instalación de OpenRefine



Estas instrucciones han sido adaptadas y traducidas al castellano a partir de un instructivo preparado por el GBIF-BID Programme.

Requerimientos

1. **Java JRE** [<https://www.oracle.com/technetwork/java/javase/downloads/index.html>] instalado.
2. **Google Chrome** [<https://www.google.com/chrome/browser/desktop/>] o **Mozilla Firefox** [<https://www.mozilla.org/en-US/firefox/new/>] instalados, evitar usar Internet Explorer.

Para instalar OpenRefine 3.1 en su computadora, siga los siguientes pasos:

Instalación en MS Windows

1. Descargue el **kit de Windows** [<https://github.com/OpenRefine/OpenRefine/releases/download/3.1/openrefine-win-3.1.zip>] aquí.
2. Descomprima, y copie la carpeta en su computadora. Abra la carpeta y haga doble click en openrefine.exe. Si encuentra algún problema en este punto, haga doble click sobre refine.bat.
3. Aparecerá una ventana de comando (que no debe cerrar) e inmediatamente después su navegador web mostrará una nueva ventana con la aplicación.

Instalación en Mac

1. Descargue el **kit de Mac** [<http://openrefine.org/download.html>], ésta es la última versión estable.
2. Abra y arrastre el ícono en la carpeta Applications.
3. Haga doble clic en él y su navegador web mostrará una nueva ventana con la aplicación.



Si tiene problemas para instalar Open Refine en Mac puede deberse a que sólo trabaja con Java 6 y 7. También puede probar instalando la última versión (no estable) disponible en la **web de OpenRefine** [<http://openrefine.org/download.html>].



Existe también una **versión para Linux** [<https://github.com/OpenRefine/OpenRefine/releases/download/3.1/openrefine-linux-3.1.tar.gz>].

Para saber más

- **Instrucciones adicionales sobre cómo instalar Open Refine** [<https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions>]
- Si además quiere hacer algunas pruebas de uso, puede consultar las funciones básicas en la sección **Introduction to OpenRefine** [<http://openrefine.org/index.html>]