

OpenRefine - Guía básica

Limpieza de datos sobre biodiversidad

Junio- 2019

Versión - 2.1



OpenRefine - Guía básica
Limpieza de datos sobre biodiversidad
2019

URI:

Cómo citar: SiB Colombia (2019). OpenRefine - Guía básica, Limpieza de datos sobre biodiversidad. Sistema de Información sobre Biodiversidad de Colombia, Bogotá D.C., Colombia, 22 pp. Disponible en:

© **Copyright Sistema de Información sobre Biodiversidad de Colombia – SiB Colombia, 2017**

Contenidos: Equipo Coordinador del SiB Colombia.

Diseño y diagramación: Equipo Coordinador del SiB Colombia

Control del documento:

Versión	Descripción	Fecha publicación	Autor(es)
1.0	Creación del documento	2014-11-18	Néstor Beltrán
2.0	Modificación sección de uso avanzado	2016-08-23	Leonardo Buitrago
2.1	Ajustes generales del documento	2019-05-13	Leonardo Buitrago Camila Plata Ricardo Ortiz

Este material circula bajo una licencia Creative Commons CC BY-SA 4.0



Puedes remezclar, modificar y crear a partir de esta obra, incluso con fines comerciales, siempre y cuando des los créditos correspondientes y licencies las nuevas creaciones bajo las mismas condiciones. Para ver una copia de esta licencia visita:

https://creativecommons.org/licenses/by-sa/4.0/deed.es_ES

Acerca del SiB Colombia

El SiB Colombia es la red nacional de datos abiertos sobre biodiversidad. Esta iniciativa de país nace con el Decreto 1603 de 1994 como parte del proceso de creación del Sistema Nacional Ambiental (Sina), establecido en la Ley 99 de 1993, y es el nodo oficial del país en la Infraestructura Mundial de Información en Biodiversidad (GBIF). Su principal propósito es brindar acceso abierto a información sobre la diversidad biológica del país para la construcción de una sociedad sostenible. Además, facilita la publicación en línea de datos e información sobre biodiversidad, y promueve su uso por parte de una amplia variedad de audiencias, apoyando de forma oportuna y eficiente la gestión integral de la biodiversidad.

El SiB Colombia es una realidad gracias a la participación de cientos de organizaciones y personas que comparten datos e información bajo los principios de libre acceso, transparencia, cooperación, reconocimiento y responsabilidad compartida.

Lo coordina el Instituto Humboldt y es liderado por un Comité Directivo (CD-SiB), conformado por el Ministerio de Ambiente y Desarrollo Sostenible, los 5 institutos de investigación del SINA (Ideam, Invemar, IIAP, Sinchi e Instituto Humboldt), la Universidad Nacional de Colombia y Parques Nacionales Naturales de Colombia. El CD-SiB se apoya en un Comité Técnico (CT-SiB), grupos de trabajo para temas específicos y un Equipo Coordinador (EC-SiB) que cumple las funciones de secretaría técnica, acogiendo e implementando las recomendaciones del CD-SiB.

El SiB Colombia promueve la participación activa del gobierno, la academia, el sector productivo y la sociedad civil para lograr la consolidación de información confiable y oportuna que apoye la toma de decisiones a nivel nacional e internacional. Es además, el nodo oficial del país en la infraestructura mundial de información en biodiversidad -GBIF-.

La implementación del SiB Colombia, a partir del 2000, constituyó el primer resultado del nuevo enfoque de gestión de datos e información en el ámbito nacional y se encuentra articulado con el Sistema de Información Ambiental de Colombia (SIAC) como el subsistema de información que soporta el componente de biodiversidad.



OpenRefine - Guía básica

Limpieza de datos sobre biodiversidad

ÍNDICE

CONVENCIONES DE LOS EJERCICIOS

I. INSTALACIÓN

II. FUNCIONES BÁSICAS

CARGA DE DATOS

'FACETING'

FILTROS

CONJUNTOS

III. VALIDACIÓN TAXONÓMICA


USO DEL API DE GBIF

IV. EXPORTACIÓN

V. FUNCIONES ADICIONALES

VI. ENLACES ADICIONALES

CONVENCIONES DE LOS EJERCICIOS

Fórmulas o información a usar en la herramienta (copiar y pegar).	cell.recon.match.id
Comandos y rutas en OpenRefine.	Edit column
Nombres de las columnas.	nombreRecon
Enlaces a sitios informativos.	www.sibcolombia.net
Menú Columna	

I. INSTALACIÓN

OpenRefine (anteriormente Google Refine) es una herramienta que dispone de un conjunto de características para trabajar con datos tabulares que mejoran la calidad general de un conjunto de datos. Se trata de una aplicación que se ejecuta fuera de su propia computadora como un pequeño servidor web, al que se accede desde un navegador web. Debe pensar en OpenRefine como una aplicación web personal y de acceso privado.

Para instalar el software en su computadora, siga los siguientes pasos:

Instalación en Windows

1. Descargue el Open Refine versión 2.8 para Windows [aquí](#)¹. Ésta es la última versión estable.
2. Descomprima el archivo descargado y copie la carpeta resultante en el disco local (C:/)
3. Abra la carpeta y haga doble clic en **openrefine.exe**. Si encuentra algún problema en este punto, haga doble clic 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 Open Refine versión 2.8 para Mac [aquí](#). Ésta es la última versión estable.
2. Abra y arrastre el icono en la carpeta Aplicación.
3. Haga doble clic en él y su navegador web mostrará una nueva ventana con la aplicación.

Notas

- 1: 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.
- 2: Existe también una versión para Linux.
3. Si al dar doble clic no abre el OpenRefine en el navegador, puede escribir la siguiente dirección en el buscador (preferiblemente Chrome): **http://127.0.0.1:3333/**

¹ <https://github.com/OpenRefine/OpenRefine/releases/download/2.8/openrefine-win-2.8.zip>

Requerimientos

1. [Java JRE](#) instalado.
2. Google Chrome o Mozilla Firefox instalados, evitar usar Internet Explorer.

Para saber más

- Instrucciones adicionales sobre la [instalación](#)² de OpenRefine:
- Si además quiere hacer algunas pruebas de uso, puede consultar las funciones básicas en la sección [Introduction to OpenRefine](#)³.

II. FUNCIONES BÁSICAS

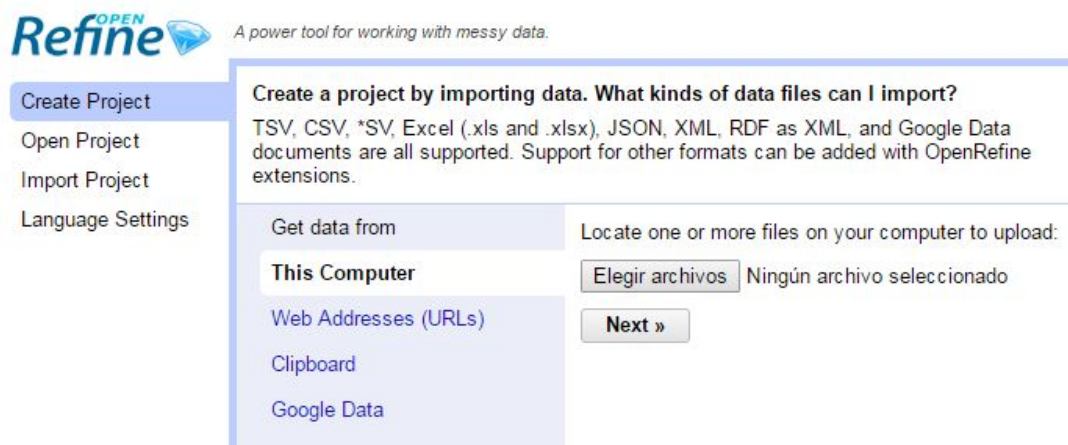
CARGA DE DATOS

ANTES DE EMPEZAR

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

CREAR UN PROYECTO

1. Tenga presente el lugar donde almacenó el archivo **Datos_Estructurados.xlsx**
2. Abra *OpenRefine* y diríjase a la pestaña **Create Project**. Para cargar el archivo siga la ruta **Get data from > This Computer**, y haga clic en **Choose Files (Elegir archivos)**:



3. Seleccione el archivo **Datos_Estructurado.xlsx** y haga clic en **Next**.
4. Un panel de selección aparecerá, este le permite especificar el tipo de datos que se cargan y configurar la manera en la que los datos son leídos.

² <https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions>

³ <http://openrefine.org/index.html>

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



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

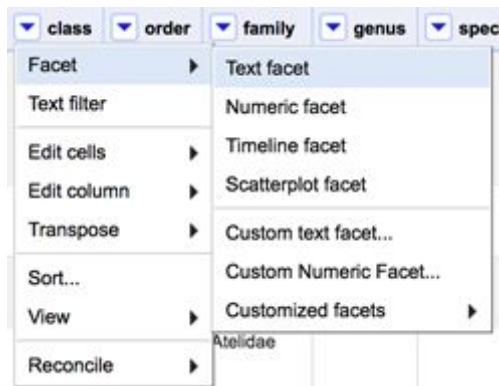
'FACETING'

ANTES DE EMPEZAR

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

EJERCICIO 1. 'Faceting' y correcciones masivas

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



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



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

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



Verá que todos los valores serán corregidos de manera automática y las celdas se transformarán de forma masiva.


- Realice el mismo proceso con la columna **basisOfRecord** y **sex** para que se ajusten al vocabulario controlado de este elemento (ver Guía rápida de DwC).



- Al finalizar este ejercicio diríjase en el menú lateral y seleccione la opción **Remove All**. Así removerá todos los Facets y Filtros que tenga en uso



EJERCICIO 2. 'Faceting' y espacios en blanco

- Diríjase a la columna **individualCount**, haga clic en el Menú Columna  y realice un **Text Facet**

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




Aunque a simple vista los datos se encuentran sin errores, al realizar este procedimiento vemos que el programa ha detectado espacios extra y por eso nos muestra cuatro opciones diferentes para el valor '1'.

3. Corrija las inconsistencias desde el Menú Columna de `individualCount`, siguiendo la ruta `Edit Cells > Common transforms > Trim leading and trailing whitespace`, verá un mensaje de notificación:

**Text transform on # cells in column individualCount:
value.trim() Undo**

4. Si observa la ventana del Facet de `individualCount`, notará que ahora solo existe una opción y que los espacios fueron eliminados.
5. Al finalizar este ejercicio diríjase en el menú lateral y seleccione la opción `Remove All`. Así removerá todos los Facets y Filtros que tenga en uso

EJERCICIO 3. 'Faceting' y duplicados

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



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

2. Haga clic en `true` y verá los registros. De esta manera se pueden detectar los duplicados para un análisis posterior. En este caso corrija el registro de Feb 2001 por 46-2300MI2008AV0248 tanto en `catalogNumber` como en `occurrenceID`; como se puede observar en la imagen a continuación.

ID del registro b	Código de la ins	Código de la col	Número de catál	Base del registr	Registrado por	Fecha del event
UCN:MH-CZO:46-2300MI2008AV0242	UCN	MH-CZO	46-2300MI2008AV0242	En Colección	Cardona F; López C	2001/03/19
UCN:MH-CZO:46-2300MI2008AV0242					Cardona F; López C	Feb 2001

Data type: **text**

UCN:MH-CZO:46-2300MI2008AV0242


Apply Apply to All Identical Cells Cancel

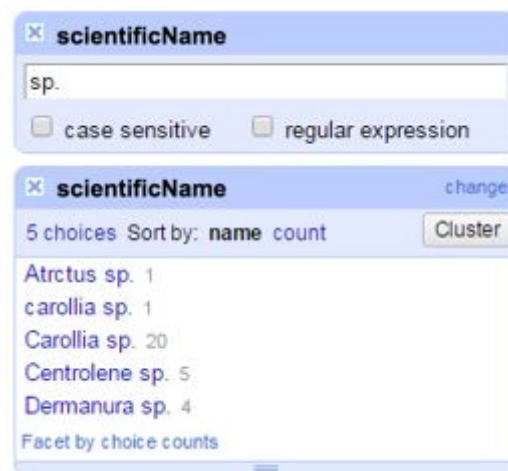
Enter Ctrl-Enter Esc

- Al finalizar este ejercicio diríjase en el menú lateral y seleccione la opción **Remove All**. Así removerá todos los Facets y Filtros que tenga en uso

FILTROS

EJERCICIO 4. Filtro básico y reemplazo de valores


- Diríjase a la columna **scientificName**, haga clic en el Menú Columna  y luego en **Text filter**, aparecerá la ventana del Filtro.
- Escriba en el campo de texto **sp.** y realice un **Text Facet** en **scientificName** para visualizar los registros con este valor:

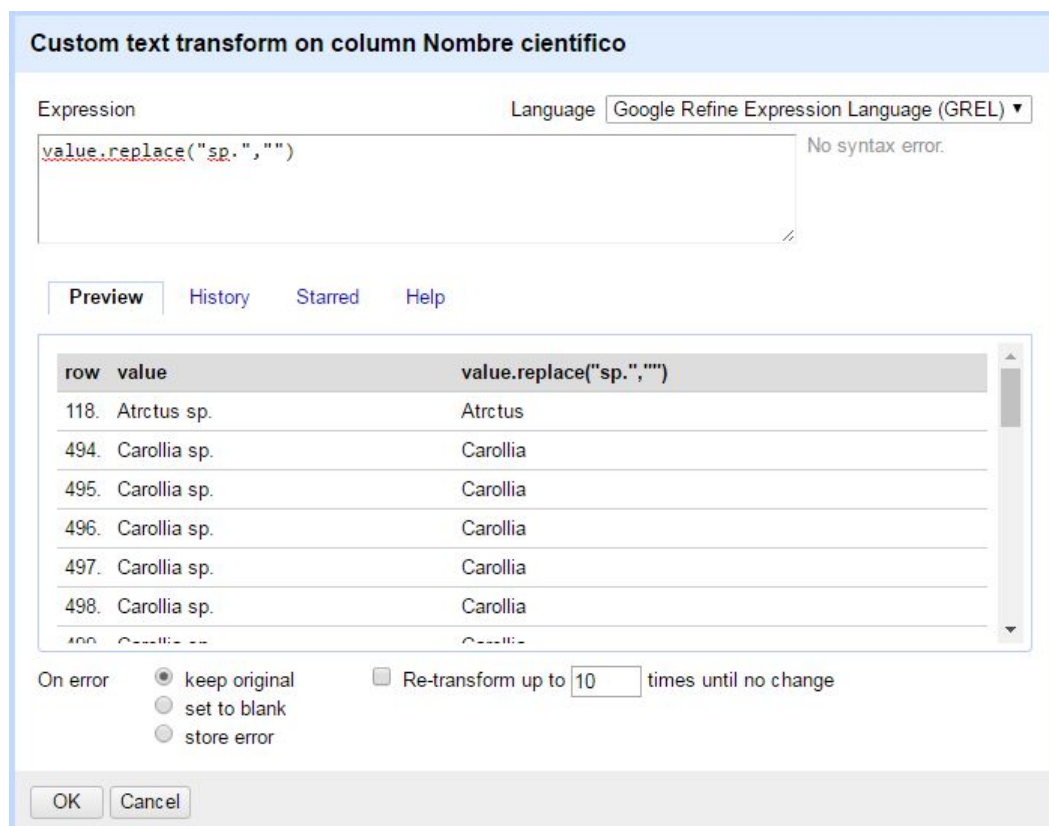


Este tipo de clasificaciones no determinadas no deben documentarse en el elemento **scientificName**, para ello se emplea **verbatimTaxonRank**.

- Realice un **Text Facet** en **verbatimTaxonRank** y edite masivamente reemplazando las celdas vacías (**blank**) con **sp.**, haga clic en **Apply**.



4. Diríjase nuevamente al Menú Columna  de `scientificName` y siga la ruta **Edit cells > Transform...**, luego ingrese la fórmula `value.replace(" sp.", "")` tal y como se muestra a continuación:



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


Empleando este comando `value.replace` podemos sustituir cualquier valor de una columna poniendo dentro de un paréntesis inicialmente el valor a buscar (ej. " sp."), entre comillas ["] y luego separado por una coma [,] el valor de reemplazo (en este caso ninguno por lo cual se ponen unas comillas vacías ["]).

6. Vaya ahora a las columnas `recordedBy` y `identifiedBy` y, empleando la misma función del punto 4, reemplace en cada una el carácter de separación entre los nombres ["; "] por

el que acepta el estándar Darwin Core actualmente para este elemento [" | "] (ver Guía rápida de DwC).

- Al finalizar este ejercicio diríjase en el menú lateral y seleccione la opción **Remove All**. Así removerá todos los Facets y Filtros que tenga en uso

EJERCICIO 5. Filtro avanzado

- Diríjase a la columna **family** realice un **Text Facet**.
- Haga clic en el Menú Columna  y luego en **Text filter**. Aparecerá la ventana del Filtro.
- Marque la casilla **regular expression**. Escriba en el campo de texto la expresión regular `.*(?:?!ae).$` esta expresión nos permite excluir todas las palabras de la columna que no terminan en "ae".




Podrá observar como los registros que no corresponden a familias han sido filtrados, usted puede editarlos tal cual como en el punto 3 de Ejercicio 1. En este caso particular reemplace **Bolitoglossa** (que corresponde a un Género) por **Plethodontidae** (la Familia a la que pertenece el nombre científico).

Para conocer más de las expresiones regulares haga clic [aquí](#).

- Al finalizar este ejercicio diríjase en el menú lateral y seleccione la opción **Remove All**. Así removerá todos los Facets y Filtros que tenga en uso.

EJERCICIO 6. Filtro avanzado II

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

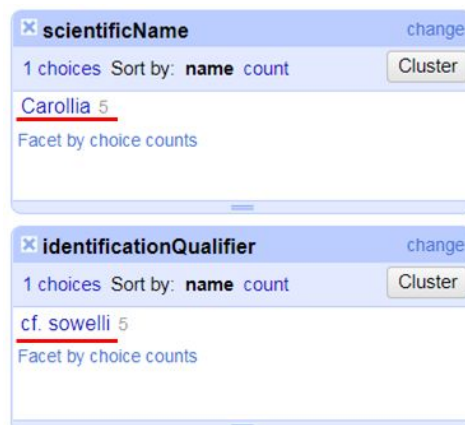


Podrá observar los registros que cumplen este criterio. El elemento `identificationQualifier` está diseñado para almacenar este tipo de información y por su parte el elemento `scientificName` debe encontrarse sin calificadores.

- Para hacer el ajuste realice un **Text Facet** en el elemento `identificationQualifier` para editar masivamente, de manera que en los **blank** se documente "`cf. sowelli`" y se borre en el `scientificName`.




- Finalmente estos registros deben quedar documentados con el género `Carollia` en `scientificName` y en `identificationQualifier` el valor `cf. sowelli`.



- Al finalizar este ejercicio diríjase en el menú lateral y seleccione la opción **Remove All**. Así removerá todos los Facets y Filtros que tenga en uso.

CONJUNTOS

EJERCICIO 7. Conjunto básico

- Diríjase a la columna **recordedBy**, haga clic en el Menú Columna  y luego en **Text facet**, aparecerá la ventana del Facet con 254 diferentes entradas de datos (**choices**):



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

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	4	<ul style="list-style-type: none"> David H; Arango A; Bedoya J (2 rows) David H; Arango A; Bedoya J (1 rows) David H; Arango A; Bedoya J (1 rows) 	<input type="checkbox"/>	David H; Arango A; Bedoya J
3	132	<ul style="list-style-type: none"> Vargas I (130 rows) Vargas I (1 rows) Vargas I (1 rows) 	<input type="checkbox"/>	Vargas I
2	3	<ul style="list-style-type: none"> Rodríguez Wilson; Giraldo F; Marín M (2 rows) Wilson Rodríguez; Giraldo F; Marín M (1 rows) 	<input type="checkbox"/>	Rodríguez Wilson; Giraldo F; M
2	12	<ul style="list-style-type: none"> Giraldo D (11 rows) D Giraldo (1 rows) 	<input type="checkbox"/>	Giraldo D

- Podrá ver la siguiente información:

Cluster size: La cantidad versiones que el algoritmo muestra como similares.

Row count: El número de registros por cluster.

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

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

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

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

Cluster & Edit column "recordedBy"

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

Method: key collision | Keying Function: ngram-fingerprint | Ngram Size: 1 | 7 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	4	<ul style="list-style-type: none"> David H; Arango A; Bedoya J (2 rows) David H; Arango A; Bedoya J (1 rows) David H; Arango A; Bedoya J (1 rows) 	<input type="checkbox"/>	David H; Arango A; Bedoya J
3	132	<ul style="list-style-type: none"> Vargas I (130 rows) Vargas I (1 rows) Vargas I (1 rows) 	<input type="checkbox"/>	Vargas I
2	3	<ul style="list-style-type: none"> Rodríguez Wilson; Giraldo F; Marín M (2 rows) Wilson Rodríguez; Giraldo F; Marín M (1 rows) 	<input type="checkbox"/>	Rodríguez Wilson; Giraldo F; M
2	15	<ul style="list-style-type: none"> López J; Idárraga A; Correa D; Sánchez L (13 rows) López J; Idárraga A; Correa D; Sánchez Lorena (2 rows) 	<input type="checkbox"/>	López J; Idárraga A; Correa D;
2	12	<ul style="list-style-type: none"> Giraldo D (11 rows) D Giraldo (1 rows) 	<input type="checkbox"/>	Giraldo D
2	4	<ul style="list-style-type: none"> Molina L; Agudelo A; González A; Sierra S; Zapata F (3 rows) Agudelo M; González A; Sierra S; Zapata F (1 rows) 	<input type="checkbox"/>	Molina L; Agudelo A; González

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

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

- Para el primer cluster asigne un valor nuevo, para esto vaya al cuadro de texto de **New cell value** y escriba **David H | Arango A | Bedoya J** (dejando espacios sencillos). Luego haga check en el cuadro de **Merge?:** para ese cluster.
- Para el segundo clúster haga clic en **Vargas I** (la primera opción: sin espacios adicionales), automáticamente el valor en **New cell value** cambiará y la casilla **Merge?** se chequeará.
- Con los restantes evalúe si se deben o no agrupar dependiendo de las opciones disponibles y escoja en tal caso si selecciona o no la casilla.
- Una vez escoja las entradas que desee fusionar vaya a **Merge Selected & close** para agrupar los valores y volver a la ventana principal. El resultado del proceso debería verse así:



Observe que la cantidad de entradas de datos disminuyó y que la primera entrada de nombres ha cambiado, es decir la información se simplificó y organizó correctamente gracias a este proceso.

- Al finalizar este ejercicio diríjase en el menú lateral y seleccione la opción **Remove All**. Así removerá todos los Facets y Filtros que tenga en uso.

III. VALIDACIÓN TAXONÓMICA

USO DEL API DE GBIF

ANTES DE EMPEZAR

En este ejercicio se utilizará el API de la *Global Biodiversity Information Facility* (GBIF) a través de OpenRefine para verificar la validez taxonómica de una lista de nombres determinada.


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

<http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c>

Entre los proveedores de clasificaciones se encuentran:

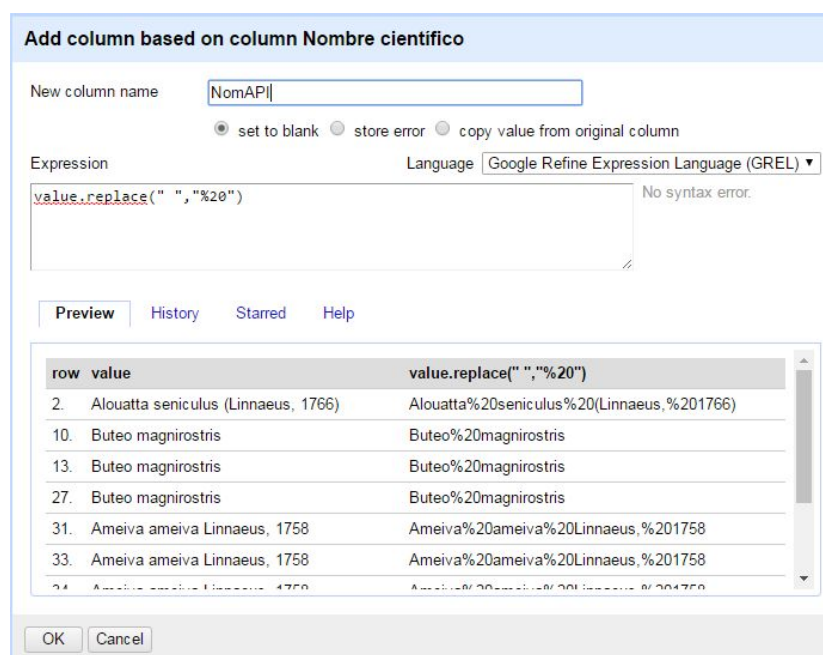
CoL	http://www.catalogueoflife.org/
Index Fungorum	http://www.indexfungorum.org/
ITIS	http://www.itis.gov/
IUCN	http://www.iucn.org/
International Plant Names Index	http://www.ipni.org/
The Paleobiology Database	http://www.paleodb.org/
Integrated Taxonomic Information System (ITIS)	http://www.itis.gov/
World Register of Marine Species (WoRMS)	http://www.marinespecies.org/


EJERCICIO 8. Validación taxonómica

1. Elimine los facets o filtros que tenga activos.
2. Para tener una aproximación inicial al funcionamiento del API diríjase a la columna `recordedBy`, haga clic en el Menú Columna  y realice un **Text Facet**. Luego haga clic en **count** y seleccione al investigador(es) con mayor número de registros asociados (Vargas I).



3. Vaya a la columna `scientificName`. Es importante que estos nombres no contengan calificadores de como "cf.", "sp." o "spp.", de ser este el caso elimínelos como se mostró en el ejercicio 4 y 6 y deje solamente como valor el nombre científico sin autoría.
4. Para realizar la validación es necesario que los espacios en blanco en cada nombre científico sean reemplazados por un valor que reconozca el API ("%20"). Para ello vaya a **Edit column > Add column based on this column** e introduzca la expresión (tal y como aparece) `value.replace(" ", "%20");` nombre la columna `NomAPI`.



5. Cree una columna llamada `validTax` a partir de la columna `NomAPI`. Para esto haga clic en Menú Columna  y luego siga la ruta **Edit column > Add column by fetching URLs...** e introduzca la expresión (tal y como aparece):

"[http://api.gbif.org/v1/species/match?strict=true&name="+value](http://api.gbif.org/v1/species/match?strict=true&name=)

En el campo **Throttle delay** escriba 250, haga clic en **OK** y espere a que finalice el proceso; el tiempo de consulta depende de la cantidad de información y de la velocidad de la red (para este caso no tardará más de un par de minutos).

Add column by fetching URLs based on column NomAPI

New column name: Throttle delay: milliseconds

On error: set to blank store error

Formulate the URLs to fetch:

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

No syntax error.

Preview History Starred Help

row	value	"http://api.gbif.org/v1/species/match?strict=true&name="+value
2.	Alouatta%20seniculus%20(Linnaeus,%201766)	http://api.gbif.org/v1/species/match?strict=true&name=Alouatta%20seniculus%20(Linnaeus,%201766)
10.	Buteo%20magnirostris	http://api.gbif.org/v1/species/match?strict=true&name=Buteo%20magnirostris
13.	Buteo%20magnirostris	http://api.gbif.org/v1/species/match?strict=true&name=Buteo%20magnirostris
27.	Buteo%20magnirostris	http://api.gbif.org/v1/species/match?strict=true&name=Buteo%20magnirostris

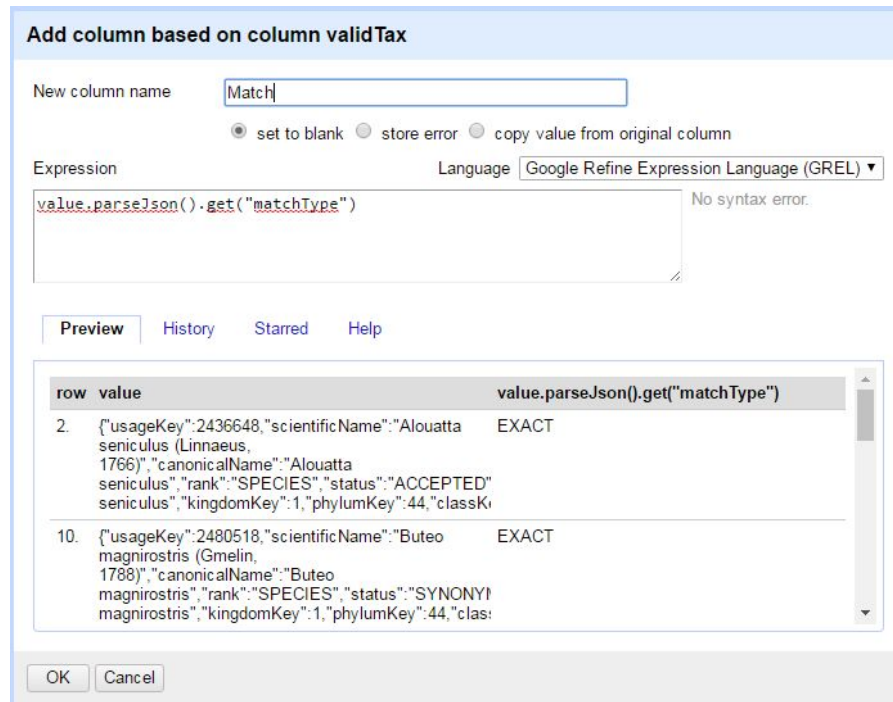
OK Cancel

Podrá observar que en cada celda de la columna **validTax** aparecen expresiones a partir del llamado al API de GBIF para cada nombre científico consultado.

132 matching rows (1000 total)		
Show as: rows records Show: 5 10 25 50 rows		
Nombre científico	NomAPI	validTax
Alouatta seniculus (Linnaeus, 1766)	Alouatta%20seniculus%20(Linnaeus,%201766)	{ "usageKey": "2436648", "scientificName": "Alouatta seniculus", "rank": "SPECIES", "status": "ACCEPTED", "kingdomKey": "1", "phylumKey": "44", "classKey": "1", "orderKey": "1", "familyKey": "1", "genusKey": "1" }
Buteo magnirostris	Buteo%20magnirostris	{ "usageKey": "2480518", "scientificName": "Buteo magnirostris", "rank": "SPECIES", "status": "SYNONYM", "kingdomKey": "1", "phylumKey": "4", "classKey": "1", "orderKey": "1", "familyKey": "1", "genusKey": "1" }
Buteo magnirostris	Buteo%20magnirostris	{ "usageKey": "2480518", "scientificName": "Buteo magnirostris", "rank": "SPECIES", "status": "SYNONYM", "kingdomKey": "1", "phylumKey": "4", "classKey": "1", "orderKey": "1", "familyKey": "1", "genusKey": "1" }
Buteo magnirostris	Buteo%20magnirostris	{ "usageKey": "2480518", "scientificName": "Buteo magnirostris", "rank": "SPECIES", "status": "SYNONYM", "kingdomKey": "1", "phylumKey": "4", "classKey": "1", "orderKey": "1", "familyKey": "1", "genusKey": "1" }
Ameiva ameiva (Linnaeus, 1758)	Ameiva%20ameiva%20(Linnaeus,%201758)	{ "usageKey": "2472164", "scientificName": "Ameiva ameiva", "rank": "SPECIES", "status": "ACCEPTED", "kingdomKey": "1", "phylumKey": "44", "classKey": "1", "orderKey": "1", "familyKey": "1", "genusKey": "1" }

6. Para observar claramente con cuál hubo o no coincidencia respecto al nombre científico agregue una columna basada en `validTax` (**Edit column > Add column based on this column...**), nómbrala `Match` y emplee la expresión:

```
value.parseJson().get("matchType")
```



7. Realice un **Text Facet** a la columna `Match` y seleccione **FUZZY**, que denota los nombres científicos con los que no hubo coincidencia exacta.



No encuentra coincidencia total para este caso con *Dermanura cinereus* ni *Dermanura glaucus*.

8. Para limpiar este error GBIF también le retorna a través del API una posible opción de nombres científicos válidos de acuerdo a los que no reconoció totalmente. Para revisar esto agregue una columna basada en `validTax` (**Edit column > Add column based on this column...**), nómbrala `validName` y emplee la expresión:
`value.parseJson().get("species")`, luego haga clic en **OK**.

Add column based on column validTax

New column name:

set to blank
 store error
 copy value from original column

Expression: Language: Google Refine Expression Language (GREL) ▾

No syntax error.

Preview History Starred Help

row	value	value.parseJson().get("species")
854.	{ "usageKey":8234645,"scientificName":"Dermanura cinerea Gervais, 1856","canonicalName":"Dermanura cinerea","rank":"SPECIES","status":"ACCEPTED","coi cinerea","kingdomKey":1,"phylumKey":44,"classKey":3	Dermanura cinerea
855.	{ "usageKey":8234645,"scientificName":"Dermanura cinerea Gervais, 1856","canonicalName":"Dermanura cinerea","rank":"SPECIES","status":"ACCEPTED","coi cinerea","kingdomKey":1,"phylumKey":44,"classKey":3	Dermanura cinerea
861.	{ "usageKey":4832192,"scientificName":"Dermanura cinerea Gervais, 1856","canonicalName":"Dermanura cinerea","rank":"SPECIES","status":"ACCEPTED","coi cinerea","kingdomKey":1,"phylumKey":44,"classKey":3	Dermanura glauca

OK Cancel

Haciendo un **Text Facet** en la columna **validName** verá que GBIF reconoce que la especie que seguramente desea documentar es *Dermanura cinerea* y *Dermanura glauca*, respectivamente.

Nombre científico change

2 choices Sort by: **name** count Cluster

Dermanura cinereus 2

Dermanura glaucus 1

validName change

2 choices Sort by: **name** count Cluster

Dermanura cinerea 2

Dermanura glauca 1

- Corrija y edite entonces las inconsistencias en la columna **Nombre científico** de acuerdo al **validName**.

Nombre científico change

2 choices Sort by: **name** count Cluster

Dermanura cinerea 2


Dermanura glaucus 1 edit include

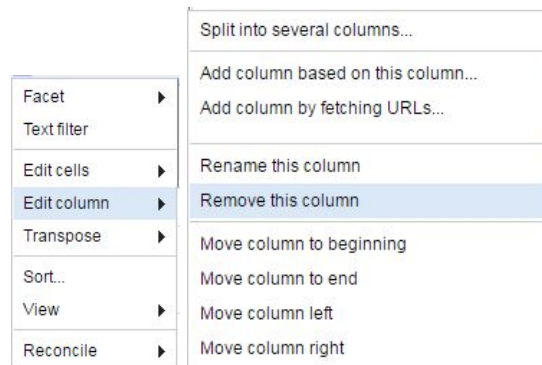
Dermanura glauca|

Apply Cancel

Enter Esc

- Habiendo realizado el proceso de verificación y limpieza de nombres científicos elimine las columnas adicionales que se crearon para este fin (**NomAPI**, **validTax**, **Match** y

validName) Para ello haga clic en el Menú Columna  de cada una y siga la ruta **Edit column** > **Remove this column**.

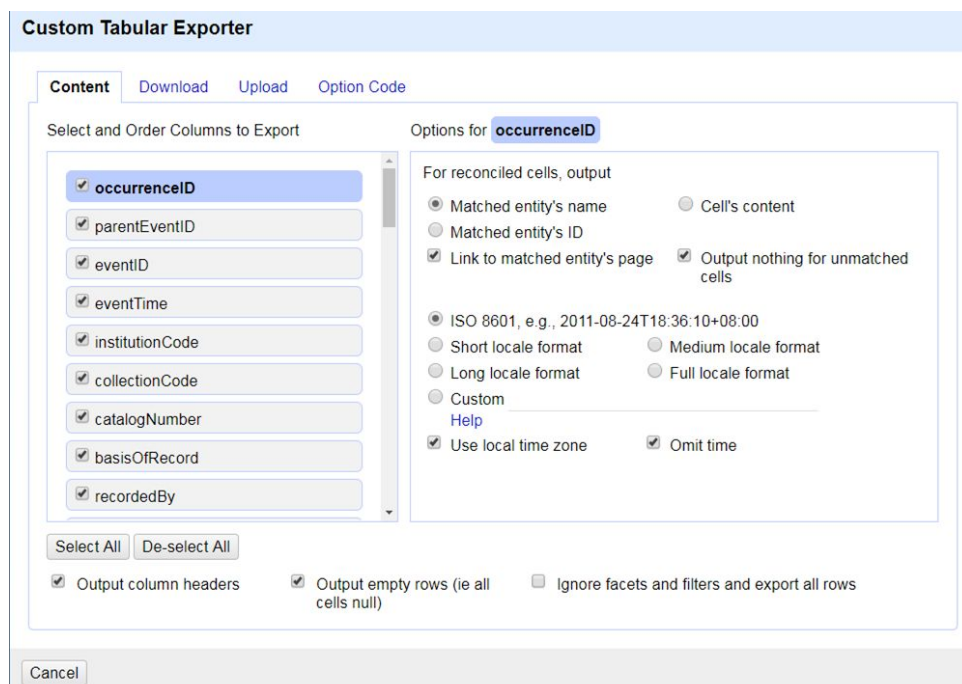


IV. EXPORTACIÓN

EJERCICIO 9. Exportar el archivo

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

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



3. En la pestaña **Content** puede seleccionar las columnas que quiere exportar, si selecciona **Ignore facets and filters and export all rows** todos los facets y filtros serán ignorados, esto sucede cuando solo queremos hacer visualizaciones de los datos.

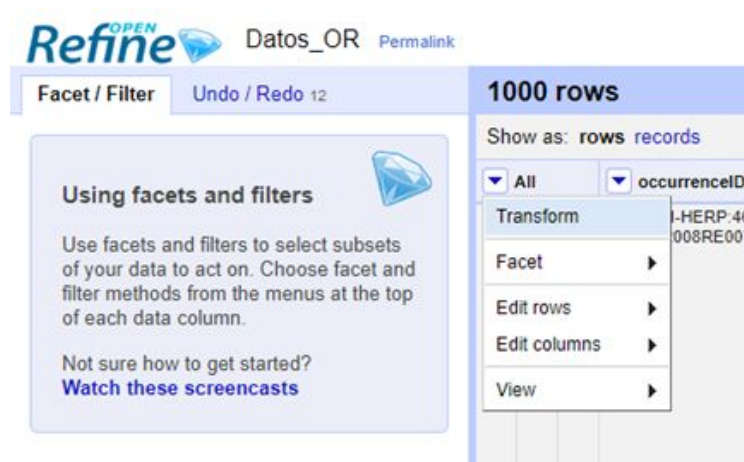
- Vaya a la pestaña **Download** y seleccione el separador de caracteres que desee, en este caso use Other formats y elija la opción *Excel(.xls)*. Haga clic en **Download** y guarde su archivo.

*Para próximas ocasiones, también puede exportar el proyecto siguiendo la ruta **Export** > **Export project**. De esta manera puede descargar el proyecto para trabajarlo en OpenRefine desde otro equipo.

V. FUNCIONES ADICIONALES

Puede hacer esta operación de manera masiva con la opción **Transform** en la columna **All**, que es la primera columna de OpenRefine. Para hacerlo sigas estos pasos.

- Diríjase a la columna **All**, haga clic en el Menú Columna  y de clic en Transform.



- En la ventana emergente inserte el siguiente texto: `" value.replace(/s+/, ' ').trim()"`. Para finalizar de clic en **OK**.

Expression

```
value.replace(/\s+/, ' ').value.trim()
```

Preview History Starred Help

```
row value value.replace(/\s+/, ' ').value ...
```

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

OK Cancel

Al ejecutarlo verá que hay cambios sobre el elemento **recordedBy** en 3 celdas, y en el elemento **verbatimElevation** en 1000 celdas.

VI. ENLACES ADICIONALES

Rutinas de validación del SiB Colombia en Open Refine:
<https://github.com/SIB-Colombia/data-quality-open-refine>

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

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

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

Documentación para usuarios del programa:
<https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users>

Lista de recursos disponibles para OpenRefine:
<https://github.com/OpenRefine/OpenRefine/wiki/External-Resources>