Uso de Open Refine

Biodiversity Data Mobilization - Data Cleaning - OpenRefine Exercise (Español)

CONTENIDOS

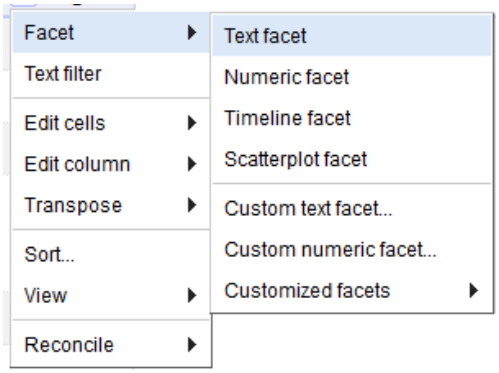
CONVENCIONES

Fórmulas (para copiar y pegar) `Texto en azul`

Ejemplo: ... pega la expresión `^[a-z]`

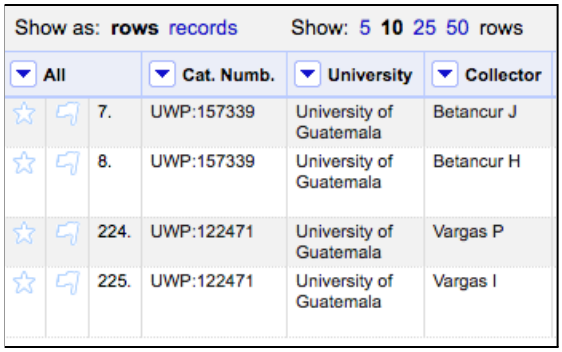
Comandos en Refine Texto en rojo

Ejemplo: ...y seguir la ruta de la faceta de Texto



Nombres de las columnas Texto en verde

Ejemplo: ...ve a la columna Cat. Numb



Enlaces a sitios informativos www.gbif.org

Menú columna

2. USO BÁSICO

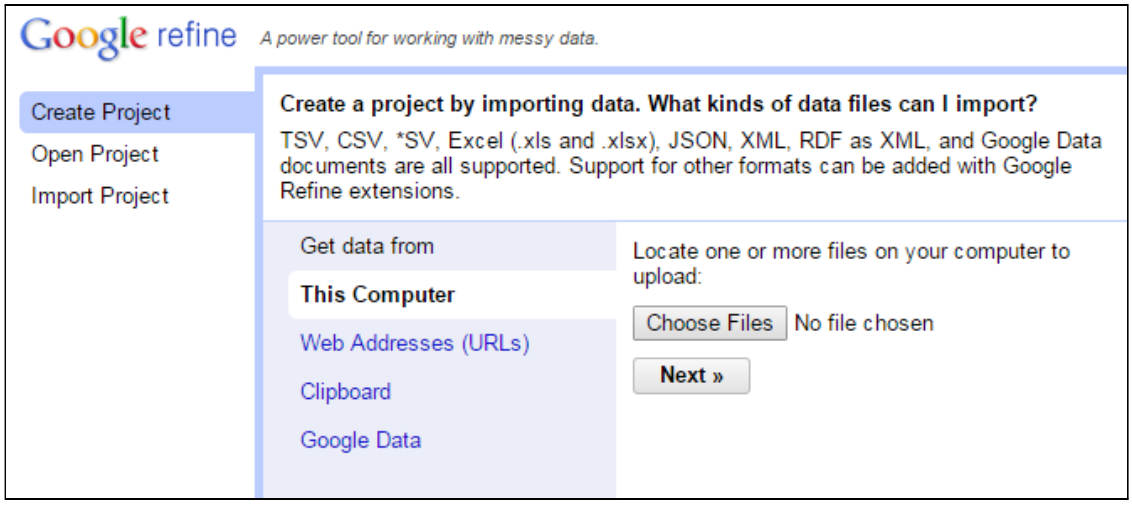
2.1. CARGA DE DATOS Y PROYECTOS

2.1.1. Antes de empezar

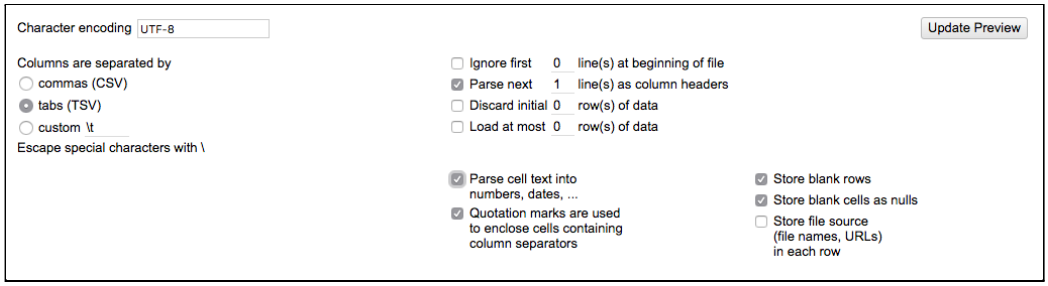
La carga de datos se puede hacer desde varias fuentes de datos: TSV, CSV, SV, Excel (.xls y .xlsx), JSON, XML, RDF y datos XML como Google Docs. La carga de datos implica dos etapas: la primera es la carga del archivo y la segunda es la creación del proyecto.

2.1.2. Ejercicio 1. Crear un proyecto

1. Cargue el archivo de datos base desde el enlace indicado en la plataforma e-learning.
2. Abra *OpenRefine* (GoogleRefine), seleccione Create Project, y siga la ruta Get data from > This Computer, después seleccione Choose Files. Seleccione el archivo.
3. Click en Next.



4. Aparecerá un menú de opciones de análisis. Asegúrese de dejar las opciones como se muestra en la imagen:



5. En la esquina superior derecha verá un cuadro de texto en el que puede cambiar el nombre del proyecto, haga clic en el botón Create Project ¡Y estará listo para trabajar!

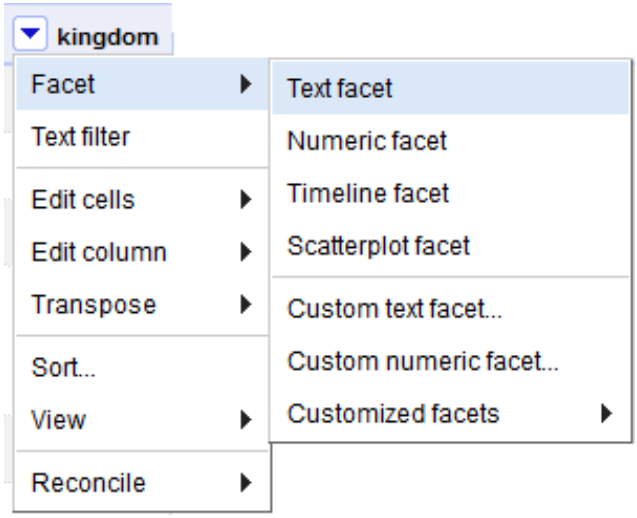
2.2. FACETAS

2.2.1. Antes de empezar

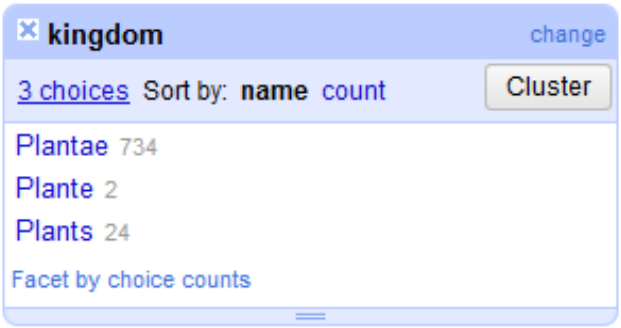
Faceting es una función que nos permitirá obtener un panorama general de los datos y filtrar sólo el subconjunto de filas que queremos cambiar o ver en bloque. Facilita el uso y análisis de datos y se puede hacer con células que contienen cualquier tipo de texto, números y fechas.

2.2.2. Ejercicio 2. Facetas y correcciones masivas

1. Diríjase a la columna kingdom, haga clic en la columna menú  y siga la ruta que se muestra en la imagen para hacer una Faceta de texto:

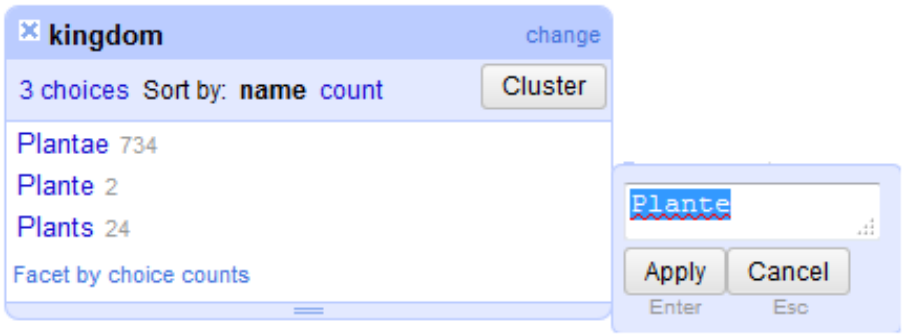


2. A la izquierda aparecerá una ventana con el nombre de la columna, que es la faceta:



Selecciona count para ordenar por cuenta, luego haga clic en name para ordenar alfabéticamente.

Corregir los errores de ortografía. Coloque el cursor sobre el texto en la ventana y haga clic en edit, luego corrija el error en el cuadro de texto, y para guardar haga clic en apply.



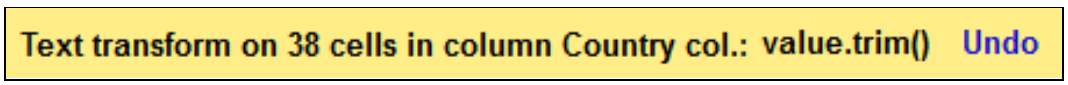
Todos los valores se corregirán automáticamente.

2.2.3. Ejercicio 3. Facetas y espacios en blanco I

1. Vaya a Country col. Y haga clic en el menú columna  y ejecute Text Facet.

En una vista rápida, el país parece estar escrito correctamente, pero la faceta muestra tres valores diferentes debido a los espacios adicionales al final del texto.

2. Corrija el error en la columna Country col., siguiendo la ruta Edit Cells > Common transforms > Trim leading and trailing whitespace. Verá un mensaje de notificación:



3. Ahora compruebe la ventana de la faceta; Sólo quedará un valor.

2.2.4. Ejercicio 4. Facetas y espacios en blanco II

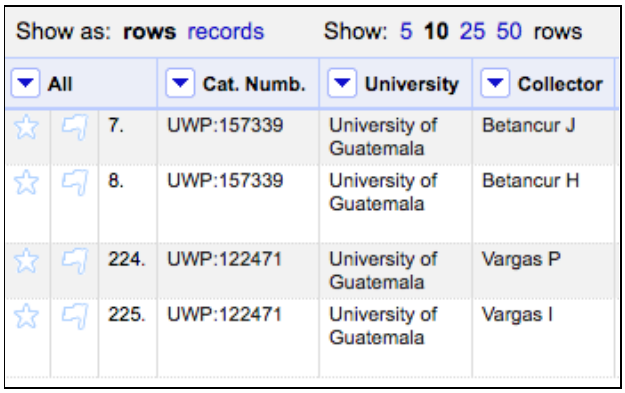
1. Vaya a la columna Full name y haga clic en  luego vaya a Text facet. A continuació seleccione count. La faceta mostrará lo siguiente:

   Como se ha visto anteriormente, *Guzmania lingulata* es el primer elemento de la lista con 25 especímenes, pero también está presente en el cuarto lugar con 20 especímenes

2. Corrija el error de la columna Full name, Edit Cells > Common transforms > Collapse consecutive whitespaces.
3. Una vez que los espacios en blanco se eliminan, *Guzmania lingulata* sólo debe aparecer en la lista con 45 registros.

2.2.5. Ejercicio 5. Facetas y duplicados

1. Diríjase a la columna Cat. Numb, haga clic en y siga la ruta Facet > Customized facets > Duplicates facet. La faceta mostrará 4 duplicados.
2. Haga clic en true, y verá los valores que se muestran en la ventana principal:



All	Cat. Numb.	University	Collector
7.	UWP:157339	University of Guatemala	Betancur J
8.	UWP:157339	University of Guatemala	Betancur H
224.	UWP:122471	University of Guatemala	Vargas P
225.	UWP:122471	University of Guatemala	Vargas I

Después de una comprobación con las etiquetas de los especímenes, corregir los valores haciendo clic en editar directamente en la celda con los números de catálogo correctos:

```
UWP:122470    Vargas P
UWP:122471    Vargas I
UWP:157351    Betancur H
UWP:157339    Betancur J
```

2.3. FILTROS

2.3.1. Ejercicio 6. Filtro básico

1. Diríjase otra vez a Full name y ejecute Text facet para visualizar los valores, después vaya otra vez a ▾ y haga clic en Text filter, Realice los siguientes filtros y corríjalos como se muestra a continuación:

Filter	How to fix	Correct value
	Edit directly in the cell	Cyperus
	Edit directly in the cell, check case sensitive	Cyperus
	1. Go to ▾ on Full name, then click Edit cells > Transform... 2. In the text box paste the formula `value.replace(" spp.", "")` 3. Click OK	Aechmea Chusquea Eleocharis Greigia Navia Neurolepis Rhynchospora Tillandsia Xyris

2.3.2. Ejercicio 7. Filtro avanzado I

1. Diríjase a la columna genus, haga clic en el Menú Columna y luego realice un filtro de texto Text filter.
2. Marque las casillas regular expression y case sensitive, después pegue la expresión `^[a-z]`

   Esta expresión regular filtra las cadenas en las que la primera letra es minúscula.

3. Corrija, ya que el género debe comenzar con mayúsculas.

Nota: Si desea obtener más información sobre las expresiones regulares, haga clic aquí.

2.3.3. Ejercicio 8. Filtro avanzado II

1. Diríjase a la columna Full name y realice un filtro de texto Text filter.
2. Marque las casillas regular expression y case sensitive, y a continuación pegue la expresión `^[A-Z].*\s[A-Z]`

   Esta expresión regular filtra las cadenas que comienzan con una letra mayúscula seguida de cualquier carácter, luego un espacio y luego una letra mayúscula.

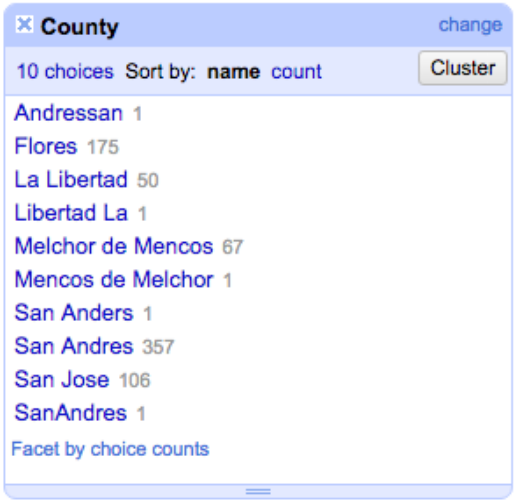
3. Corrija, ya que la segunda palabra del nombre debe estar en minúscula.

Nota: Si desea obtener más información sobre las expresiones regulares, haga clic aquí.

2.4. AGRUPACIONES

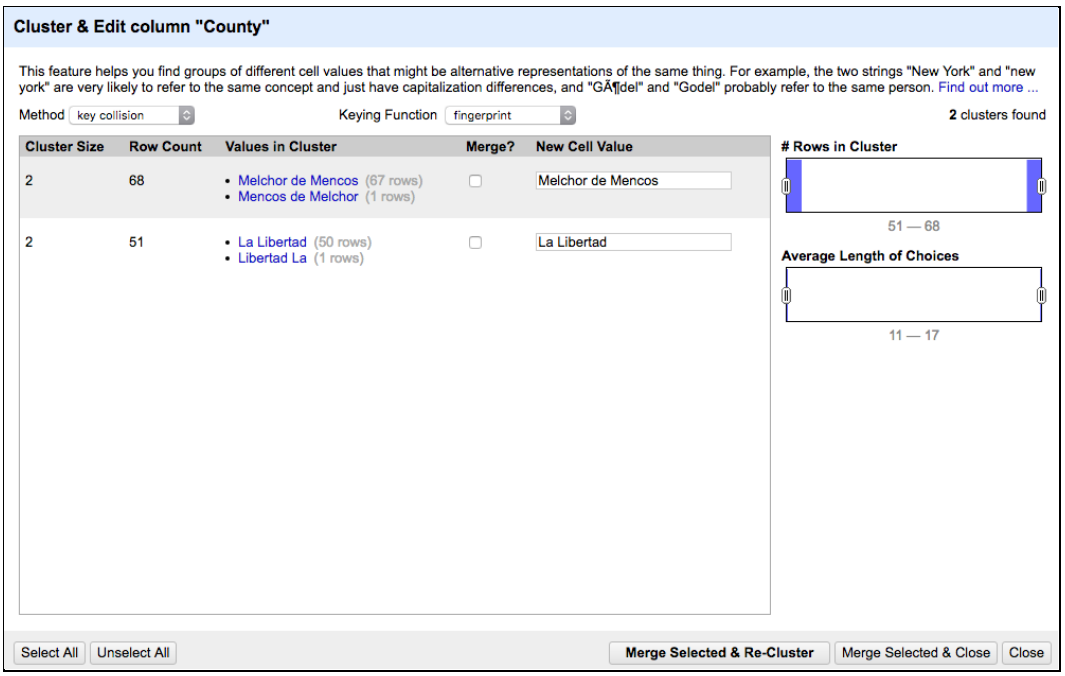
2.4.1. Ejercicio 9. Agrupaciones básicas

1. Diríjase a County, a continuación en el menú columna haga clic en Text facet.



Tenga en cuenta que los condados correctos son: Flores, La Libertad, Melchor de Mencos, San Andrés y San José.

2. En la parte superior derecha de la ventana de facetas, haga clic en Cluster, una ventana nueva aparecerá:



3. Ahora puede ver información sobre los clústeres:

- Cluster size: el número de versiones diferentes que el algoritmo de agrupación cree que son iguales.
- Row count: El número de registros con cualquiera de los valores del clúster.
- Values in cluster: los valores reales que el algoritmo cree que son los mismos. También aparece el número de registros con cada valor en particular, y la posibilidad de examinar el contenido del clúster en una pestaña diferente.
- Merge?: comprueba si los valores deben fusionarse en un solo valor estándar.
- New cell value: el valor que se aplicará a cada registro del clúster. De forma predeterminada, es el valor con la mayoría de los registros. También puede hacer clic en cualquier valor para aplicarlo al New cell value.

Nota: Si quiere saber más acerca de las agrupaciones haga clic aquí.

4. Haga clic en Select All y después en Merge Selected & close, verá un mensaje de notificación:

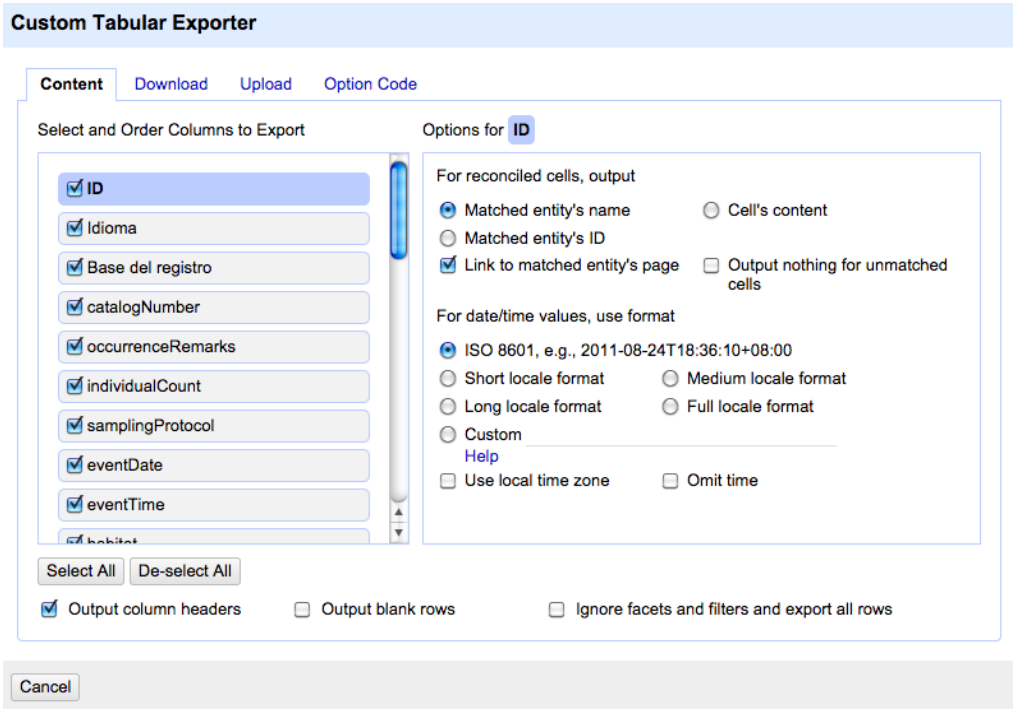
5. Para arreglar los condados restantes vaya de nuevo a Cluster la ventana de la faceta de County.
6. En la ventana de la agrupación, diríjase a Keying Function, seleccione ngram-fingerprint, y establezca 1 como valor en Ngram Size. Pulse la tecla intro.
7. Haga clic en Select All y luego en Merge Selected & close, verá un mensaje de notificación:

8. Sus condados ahora están corregidos y su ventana debe verse como la imagen de abajo:

2.5. EXPORTACIÓN

Existen varias opciones para exportar los datos limpios, pero la opción siguiente es útil en la mayoría de los casos.

1. En la esquina superior derecha, haga clic en Export y seleccione Custom tabular exporter...
2. **Verá la siguiente ventana de exportación:**



3. En la pestaña content puede seleccionar las columnas que desea exportar, si selecciona Ignore facets and filters and export all rows, se ignorarán todas las facetas y filtrados, esto es útil si se olvida de borrarlos antes de exportar.
4. **Vaya a la pestaña Download y seleccione el separador que prefiera. No modifique las otras opciones a menos que lo necesite.**

También puede exportar todo el proyecto para abrirlo en OpenRefine en otro equipo siguiendo la ruta Export > Export project. En este caso, no está descargando un archivo de datos para abrir en una hoja de cálculo o procesador de texto, sino un archivo GZIP que sólo será accesible a través de OpenRefine.

3. USO BÁSICO DE LAS API

3.1. Antes de empezar

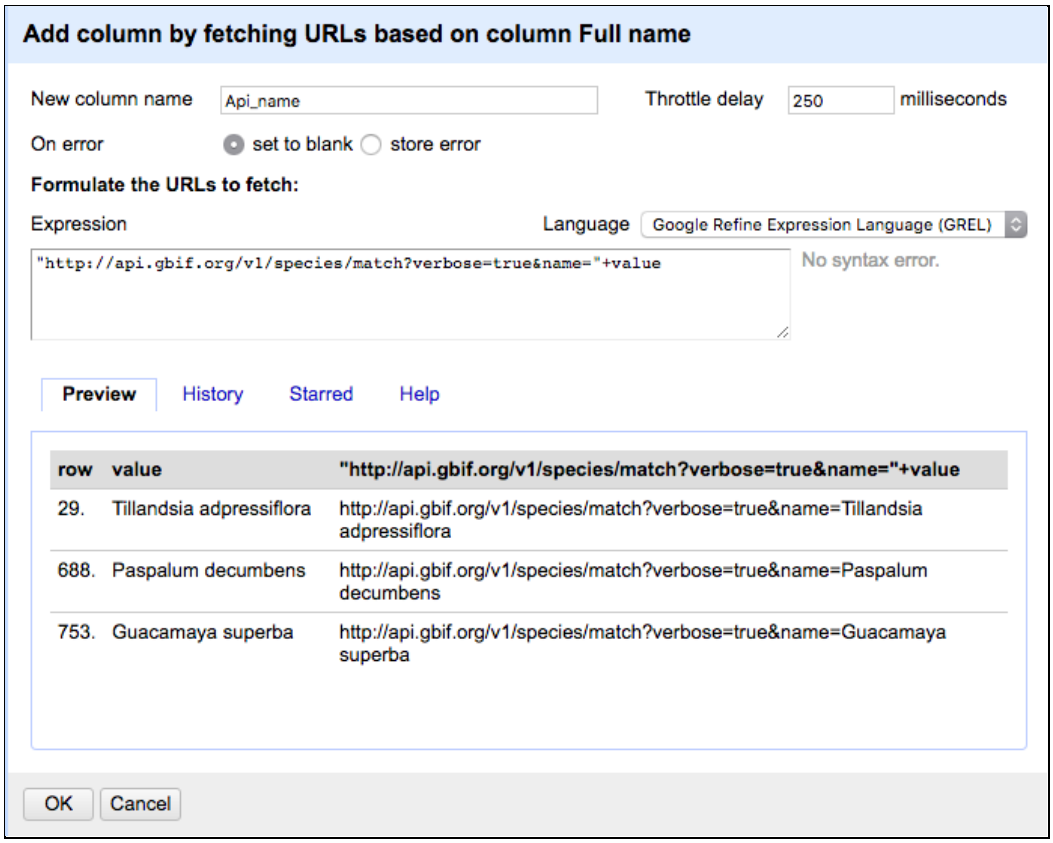
La reconciliación busca la correspondencia entre la información de una de sus columnas y una base de datos externa. Esto es particularmente útil cuando se trata de validación de nombres, ya que demuestra que el nombre que tiene en su base de datos existe en algún otro lugar. Este es un servicio muy útil, pero puede llevar mucho tiempo. En este caso vamos a pasar por el proceso con sólo tres registros utilizando la API de GBIF. Se requiere conexión a Internet.

3.2. Ejercicio 1. Taxonomía superior

1. Diríjase a Collector y haga una faceta de texto desde Text facet. Seleccione el colector Elsa P

2. Bajo Full name, haga clic en la columna menú y siga la ruta Edit column > Add column by fetching URLs..., renombre la nueva columna como Api_name
3. Cambien el Throttle Delay to `250` y pegue la expresión:

```
"http://api.gbif.org/v1/species/match?verbose=true&name="+escape(value,'url')
```

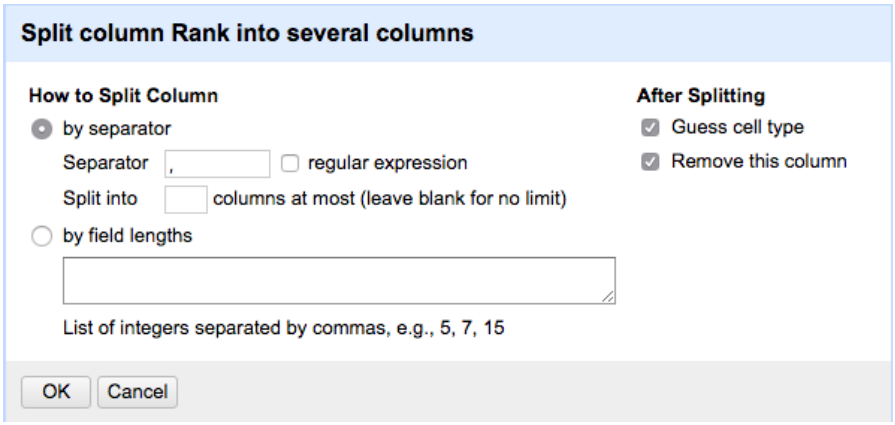


4. Haga clic en ok and wait, esto podría llevar algún tiempo dependiendo de su conexión a Internet y el número de taxones.
5. Diríjase a Api_name, haga clic en el menú de columnas y luego siga la ruta Edit column > Add column based on this column.... Renombre la nueva columna como Rank y pegue la expresión:

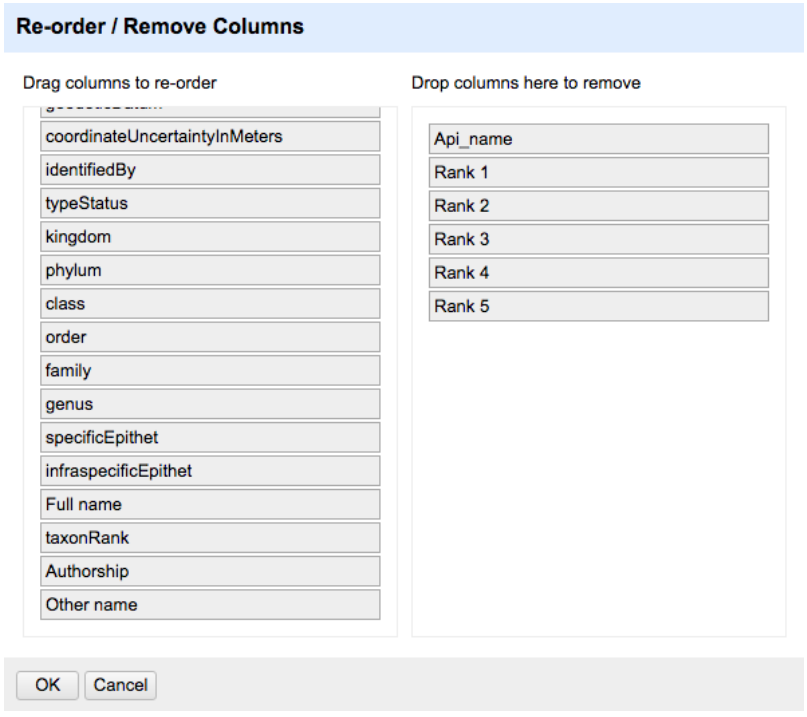
```
value.parseJson().get("kingdom")+
", "+value.parseJson().get("phylum")+
", "+value.parseJson().get("class")+
", "+value.parseJson().get("order")+
", "+value.parseJson().get("family")
```

   Verás los campos Kingdom, Phylum, Class, Order y family para cada taxon,.

6. Bajo Rank siga la ruta Edit column > Split into several columns..., deje los ajustes como se muestra:



7. Ahora ya sabe cómo obtener las categorías taxonómicas de un taxón dado si está disponible en el API de GBIF. Los nombres de columna se pueden editar desde Edit column > Rename this column.
8. Para el propósito del taller, las columnas creadas en este ejercicio (Higher taxonomy) deben ser eliminadas. En All, que es la primera columna, vaya a Edit columns > Re-order / remove columns...
9. Suelte las columnas como se muestra y haga clic en OK:



4. LINKS Y REFERENCIAS ÚTILES

- Tutorial de validación de nombres: https://docs.google.com/document/d/1tkDRXlYhmassYAk5T4v5oac5prF0jAiSMr_JEGTvhRo/edit
- Tutorial de Taxonomía Superior: https://docs.google.com/document/d/1XZ_pM9gIdQzHzl8wfUCVea-52yub5T_3tc-snBgPRa0/edit
- Documentación https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- Lista de recursos para OpenRefine: https://github.com/OpenRefine/OpenRefine/wiki/External-Resource

---

Exercise concept and content developed by Néstor Beltrán.
Updated: 03 July 2019. Sophie Pamerlon
Updated: 11 July 2019. Laura Russell, Sophie Pamerlon.