

Uso de Open Refine

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



CONTENIDOS

CONTENIDOS

CONVENCIONES

- 2. USO BÁSICO
 - 2.1. CARGA DE DATOS Y PROYECTOS

2.1.1. Antes de empezar

2.1.2. Ejercicio 1. Crear un proyecto

- 2.2. FACETAS
 - 2.2.1. Antes de empezar
 - 2.2.2. Ejercicio 2. Facetas y correcciones masivas
 - 2.2.3. Ejercicio 3. Facetas y espacios en blanco I

2.2.4. Ejercicio 4. Facetas y espacios en blanco II

2.2.5. Ejercicio 5. Facetas y duplicados

2.3. FILTROS

2.3.1. Ejercicio 6. Filtro básico

2.3.2. Ejercicio 7. Filtro avanzado I

2.3.3. Ejercicio 8. Filtro avanzado II

2.4. AGRUPACIONES

2.4.1. Ejercicio 9. Agrupaciones básicas

2.5. EXPORTACIÓN

- 3. USO BÁSICO DE LAS API
 - 3.1. Antes de empezar

3.2. Ejercicio 1. Taxonomía superior

4. LINKS Y REFERENCIAS ÚTILES



CONVENCIONES

Fórmulas (para copiar yTexto en azul pegar)

Ejemplo: ... pega la expresión ^[a-z]

× genus	
^[a-z]	
case sensitive	regular expression

Comandos en Refine

Texto en rojo

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

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

Nombres de las columnas

Texto en verde

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

Show as: rows records		Show: 5 10 25 50 rows			
			Cat. Numb.	University	Collector
☆		7.	UWP:157339	University of Guatemala	Betancur J
ಬ	5	8.	UWP:157339	University of Guatemala	Betancur H
		224.	UWP:122471	University of Guatemala	Vargas P
☆	9	225.	UWP:122471	University of Guatemala	Vargas I

Enlaces a sitios informativos

Menú columna



www.gbif.org



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.

Google refine	A power tool for working with messy data.	
Create Project Open Project Import Project	Create a project by importing da TSV, CSV, *SV, Excel (.xls and .) documents are all supported. Supp Refine extensions.	ta. What kinds of data files can I import? dsx), JSON, XML, RDF as XML, and Google Data bort for other formats can be added with Google
	Get data from	Locate one or more files on your computer to
	This Computer	
	Web Addresses (URLs)	Choose Files No file chosen
	Clipboard	Next »
	Google Data	

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

Character encoding UTF-8		Update Preview
Columns are separated by commas (CSV) tabs (TSV) custom tt Escape special characters with \	Ignore first 0 line(s) at beginning of file Parse next 1 line(s) as column header Discard initial 0 row(s) of data Load at most 0 row(s) of data	, S
	 Parse cell text into numbers, dates, Quotation marks are used to enclose cells containing column separators 	 Store blank rows Store blank cells as nulls Store file source (file names, URLs) in each row

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 v siga la ruta que se muestra en la imagen para hacer una Faceta de texto:

💌 kingdom			
Facet	►	Text facet	
Text filter		Numeric facet	
Edit cells	►	Timeline facet	
Edit column	►	Scatterplot facet	
Transpose	►	Custom text facet	
Sort		Custom numeric facet	
View	•	Customized facets	►
Reconcile	►		

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

× kingdom	change
3 choices Sort by: name count	Cluster
Plantae 734	
Plante 2	
Plants 24	
Facet by choice counts	
_	



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.

× kingdom	change			
3 choices Sort by: name count	Cluster			
Plantae 734				
Plante 2		Plante		
Plants 24		- Caller		.1
Facet by choice counts		Apply	Cancel	
=		Enter	Esc	

Todos los valores se corregirán automáticamente.

2.2.3. Ejercicio 3. Facetas y espacios en blanco l

1. Vaya a Country col. Y haga clic en el menú columna V g 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 \bigvee luego vaya a Text facet. A continuació seleccione count. La faceta mostrará lo siguiente:

× Full name	change	
253 choices Sort by: name count	Cluster	
Guzmania lingulata 25		
Aechmea veitchii 24		
Guzmania coriostachya 22		
Guzmania lingulata 20		
Aechmea tillandsioides 17		
Aechmea penduliflora 15		
Aechmea servitensis 14		
Guzmania angustifolia 13		
Aechmea dactylina 12		
Catopsis sessiliflora 12		
Aechmea Angustifolia 11		
Aechmea.nubescens.10		

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:





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
Full name sp1 case sensitive regular expression	Edit directly in the cell	Cyperus
 Full name SP2 Case sensitive regular expression 	Edit directly in the cell, check case sensitive	Cyperus
Full name spp case sensitive regular expression	 Go to on Full name, then click Edit cells > Transform In the text box paste the formula value.replace(" spp.", "") 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]

× genus	
^[a-z]	
case sensitive	regular expression

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]

× Full name	
^[A-Z].*\s[A-Z]	
case sensitive	regular expression

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:

Cluster & Ed	dit column "C	County"					
This feature he york" are very I	lps you find grou ikely to refer to th	ps of different cell values that might be ne same concept and just have capitali	alternative r zation differe	epresentations of the s nces, and "Gödel" ar	same thing. For example, th nd "Godel" probably refer to	e two strings "New York" and the same person. Find out r	d "new nore
Method key co	ollision 🗘	Keying Function	fingerprint	0		2 cluste	rs found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Rows	s in Cluster	
2	68	Melchor de Mencos (67 rows) Mencos de Melchor (1 rows)		Melchor de Mencos			0
0	54	- La Liberted (60 mus)	-	La Liberted		51 — 68	
2	51	Libertad La (1 rows)	U	La Libertad	Averag	e Length of Choices	
						11 — 17	
Select All U	nselect All			Merg	e Selected & Re-Cluster	Merge Selected & Close	Close

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:

Mass edit 119 cells in column County Undo

- 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:

Mass edit 360 cells in column County Undo

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:

ontent	Download	Upload	Option Code	
elect and	Order Column	s to Export		Options for ID
 ✓ ID ✓ Idior ✓ Base ✓ cata ✓ occu ✓ indiv ✓ sam ✓ ever ✓ ever ✓ ever 	na e del registro logNumber urrenceRemark vidualCount plingProtocol ntDate ntTime	S		For reconciled cells, output ● Matched entity's name Cell's content ● Matched entity's ID Output nothing for unmatched cells ✓ Link to matched entity's page Output nothing for unmatched cells For date/time values, use format Output nothing for unmatched cells For date/time values, use format ISO 8601, e.g., 2011-08-24T18:36:10+08:00 ● ISO 8601, e.g., 2011-08-24T18:36:10+08:00 Short locale format ● Long locale format Full locale format ● Long locale format Full locale format ● Use local time zone Omit time
elect All	De-select A	rs 🗆	Output blank	rows

- **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

× Collector	change invert reset
54 choices Sort by: name count	Cluster
Castano N; Betancur J 1	
Clavijo L 3	
Cogollo A 2	
Colorado J 7	
Cordero Z 1	
Echeverry R 2	
Elsa P 3	exclude
Fernández JL 3	
Fonnegra R 42	
Galeano G 2	
Galindo R 4	
Giraldo D 11	

- 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')



Add column by fe	etching UF	RLs based on colu	umn Full n	ame		
-	-					
New column name Api_name				Throttle delay	250	milliseconds
On error 💿 set to bla		ink 🔘 store error				
Formulate the URLs	to fetch:					
Expression			Language	Google Refine Ex	pression Lang	juage (GREL) 🗘
"http://api.gbif.org/vl/specie		es/match?verbose=true	e&name="+val	ue	No syntax	error.
					6	
Preview Hist	orv Star	red Help				
row value		"http://api.gbif.org/v	1/species/ma	atch?verbose=t	rue&name=	"+value
29. Tillandsia ad	lpressiflora	http://api.gbif.org/v1/s adpressiflora	pecies/match	?verbose=true&	name=Tillano	dsia
688. Paspalum d	ecumbens	http://api.gbif.org/v1/s decumbens	pecies/match	?verbose=true&	name=Paspa	alum
753. Guacamaya	superba	http://api.gbif.org/v1/s superba	pecies/match	?verbose=true&	name=Guaca	amaya
OK Cancel						

- **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:

va]	<pre>lue.parseJson().get("kingdom")+</pre>
",	"+value.parseJson().get("phylum")+
۰,	"+value.parseJson().get("class")+
۰,	"+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:



Split column Rank into several columns	
How to Split Column by separator Separator Split into columns at most (leave blank for no limit) by field lengths	After Splitting Guess cell type Remove this column
List of integers separated by commas, e.g., 5, 7, 15	
OK Cancel	

- 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:

g columns to re-order	Drop columns here to remove
coordinateUncertaintyInMeters	Api_name
typeStatus	Rank 2
kingdom	Rank 3
phylum	Rank 4
class	Rank 5
order	
family	
genus	
specificEpithet	
infraspecificEpithet	
Full name	
taxonRank	
Authorship	
Other name	



4. LINKS Y REFERENCIAS ÚTILES

- Tutorial de validación de nombres: <u>https://docs.google.com/document/d/1tkDRXIYhmassYAk5T4v5oac5prF0jAiSMr_JEGTvhRo/edit</u>
- Tutorial de Taxonomía Superior: <u>https://docs.google.com/document/d/1XZ_pM9gIdQzHzl8wfUCVea-52yub5T_3tc-snBgPRa0/edit</u>
 Documentación
- https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- Lista de recursos para OpenRefine: <u>https://github.com/OpenRefine/OpenRefine/wiki/External-Resource</u>

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