

# Introducción a OpenRefine

Por: Silvia Gutiérrez @espejolento (para el taller con Rodrigo Cuéllar) [CC BY SA 4.0]

Adaptado de: <https://programminghistorian.org/es/lecciones/limpieza-de-datos-con-OpenRefine>

## Descargando tus datos

1. Ubicar datos que se usarán (para el ejercicio general se elegirá el set de datos de la producción académica de los Centros):  
[https://drive.google.com/drive/folders/1lkIH\\_E8XJFCorxEqIKLoPA\\_lhy7bRoAT](https://drive.google.com/drive/folders/1lkIH_E8XJFCorxEqIKLoPA_lhy7bRoAT)
2. Descargar los datos de CEDUA en csv (sin formato, sin filas extra)
  - ¿Qué es el formato CSV?
  - ¿Cómo descargan un archivo en un formato específico?
    - i. Tip: Archivo - Descargar como
  - Recuerden convenciones de nombrar archivo (sin acentos, sin espacios, sin caracteres especiales)

## Instalando OpenRefine

1. [Descarga](#) la última versión de *OpenRefine* . [OpenRefine](#) funciona en todas las plataformas: Windows, Mac y Linux. Estas son las instrucciones son para la descarga en **Windows**:
  - a. Descarga el **Windows kit (de OpenRefine)**
  - b. Una vez descargada la carpeta comprimida en Windows da clic en “Extraer” (Imagen 1) y luego en “Extraer todo” (imagen 2)

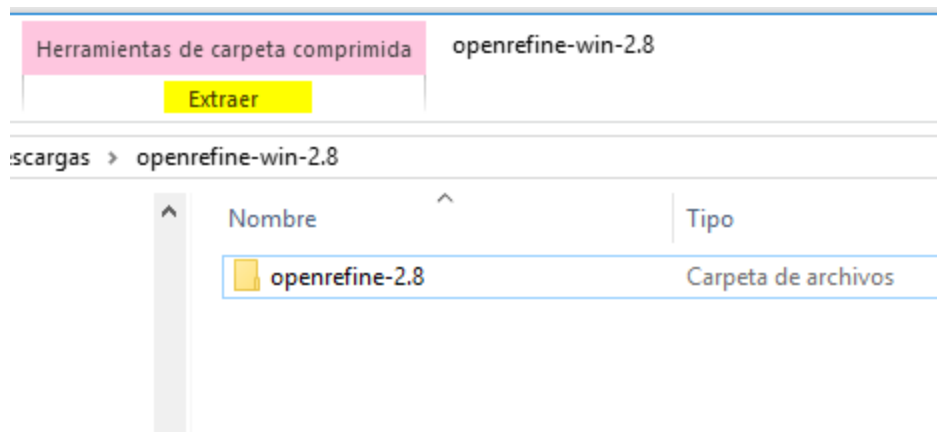


Imagen 1

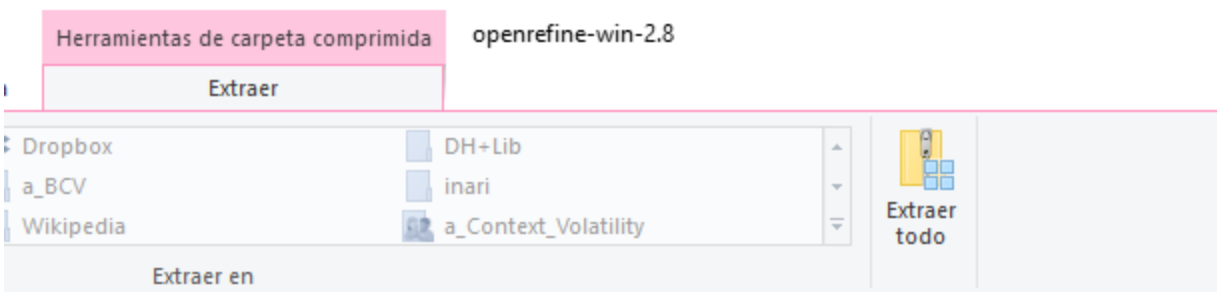
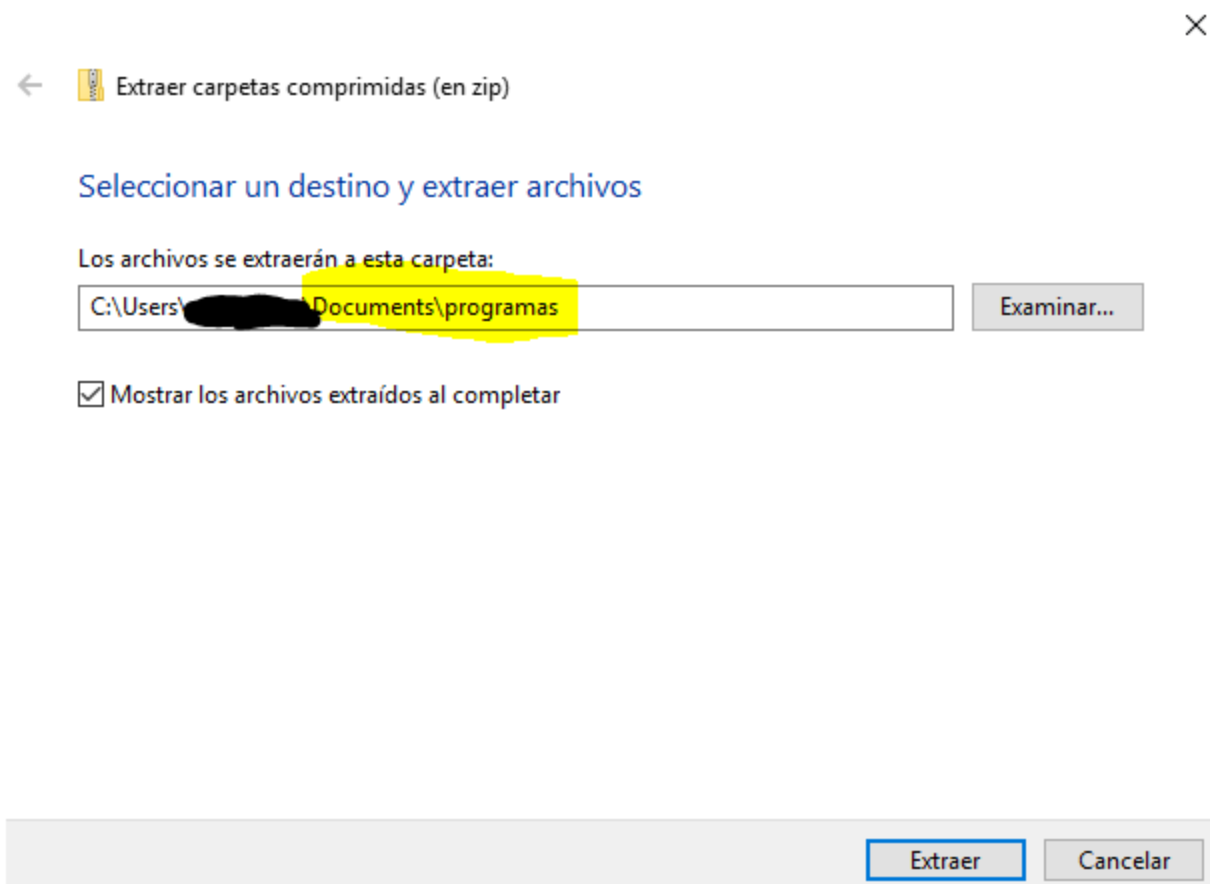


Imagen 2

c. Selecciona la carpeta en donde vas a guardar OpenRefine



## Abriendo OpenRefine

Para abrir OpenRefine da doble clic en la "aplicación", para identificar cuál es, observa la columna de Tipo (sólo en la vista de "Detalles")

💡 Tip: generalmente las aplicaciones tienen un icono distintivo (aquí es un diamante)

programas > openrefine-2.8				
Nombre	Fecha de modifica...	Tipo	Tamaño	
licenses	03/05/2018 04:34 ...	Carpeta de archivos		
server	03/05/2018 04:34 ...	Carpeta de archivos		
webapp	03/05/2018 04:34 ...	Carpeta de archivos		
LICENSE	03/05/2018 04:34 ...	Documento de tex...	4 KB	
openrefine	03/05/2018 04:34 ...	Aplicación	87 KB	
openrefine.i4j	03/05/2018 04:34 ...	Opciones de confi...	1 KB	
README	03/05/2018 04:34 ...	Documento de tex...	3 KB	
refine	03/05/2018 04:34 ...	Archivo por lotes ...	6 KB	
refine	03/05/2018 04:34 ...	Opciones de confi...	1 KB	

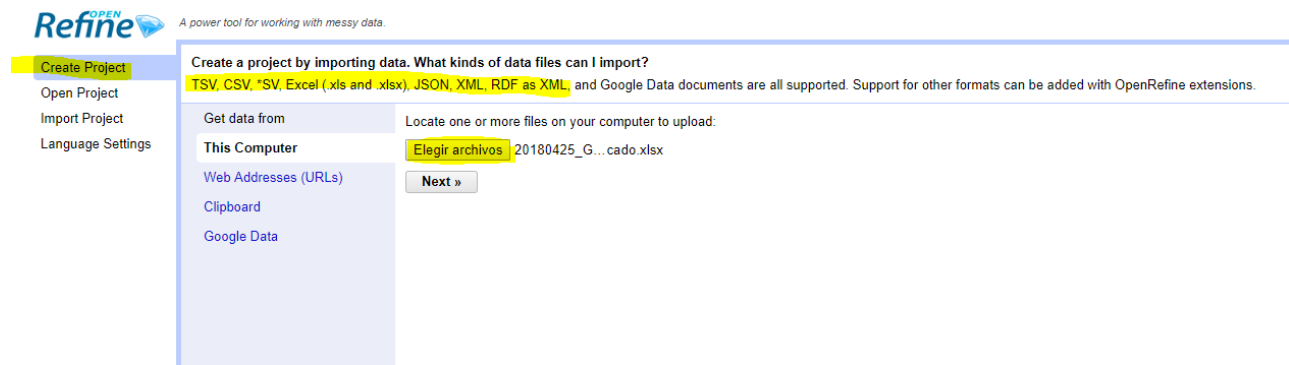
⚠ Nota: Necesitarás tener Java instalado en tu computadora, si no lo tienes, ve a: <https://www.java.com/es/download/>)

## Cargando tus datos a OpenRefine

Una vez hecho el paso anterior, *OpenRefine* se abrirá en tu navegador predeterminado (Chrome, Firefox, Opera, etc).

⚠ OJO: la aplicación se ejecuta localmente y tus datos no se almacenarán en línea.

1. Da clic en crear proyecto (“**Create Project**”) y carga la hoja de datos que quieres explorar en “**Elegir Archivos**” (puede ser **tsv**, **csv**, **excel**, **json**, **xml**, **rdf**)



2. Se verifica que se interpreten los datos como CSV, que el número de filas coincida con los datos originales y que la primera fila sea el encabezado.

⚠ OJO: Si salen letras “raras” es muy probable que el “encoding” no sea correcto. Hay que seleccionar “UTF-8” el cual corresponde a nuestro alfabeto y su forma de acentuación y puntuación.

« Start Over

Configure Parsing Options

	Autor	Título del artículo	Nombre de la revista	Editorial	Lugar de publicación	Año	Centro	Citas 2015	Citas 2016
1.	Luis Jaime Sobrino Figueroa	La urbanización en el México contemporáneo	Notas de Población	CEPAL	Chile	2012	CEDUA	3	28
2.	María del Rosario de Fátima Juárez Carcaño	Factors Associated with Abortion-Seeking and Obtaining a Safe Abortion in Ghana	Studies in Family Planning	John Wiley & Sons, Inc.	Estados Unidos	2012	CEDUA	13	28
3.	Víctor Manuel García Guerrero; José Manuel Aburto, Hiram Beltrán-Sánchez y Vladimir Canudas-Romo	Homicides In Mexico Reversed Life Expectancy Gains For Men And Slowed Them For Women, 2000-10	Health Affairs	Project HOPE — The People-to-People Health Foundation, Inc.	Estados Unidos	2016	CEDUA	0	8

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

RDF/N3 files

Wikitext

XML files

Open Document Format

Character encoding UTF-8

Columns are separated by

☒ commas (CSV)
 ☐ tabs (TSV)
 ☐ custom ,

Escape special characters with \

3. Si esto está bien: dar clic en “Create Project” (pueden darle un nombre a su proyecto)

« Start Over

Configure Parsing Options

Project name

20180503\_Gacetas\_Base

Create Project »

	NO	NO DE GACETA	FECHA	FECHA FORMATO	PAG	AUTOR	VIAF	GRUPO / ORDEN	TITULO	L DE EDICION	IMPRESOR	UBICACIÓN	AÑO	SECC	VENTA	PRECIO	MATERIA	MATERIA 2	MATERIA 3	VO
1.	1		2 Feb 1722	1722-02-01	16	Maldonado, Ángel	<a href="http://viaf.org/viaf/17174443">http://viaf.org/viaf/17174443</a>	San Bernardo	Oraciones evangelicas predicadas por el ilmo. Y Rmo. Sr. Mtro. D. Fr. Angel Maldonado... dadas a la estampa por D. Baltasar de Montoya Maldonado	México	Juan José Guillena Carrascoso, Herederos de	En la Alcayzera, Juan José Guillena Carrascoso, Herederos de	1721	LNM			Iglesia Católica -- Sermones	Oraciones		
2.	2		2 Feb 1722	1722-02-01	16	Maldonado, Ángel	<a href="http://viaf.org/viaf/17174443">http://viaf.org/viaf/17174443</a>	San Bernardo	Directorio espiritual que haze a sus queridos hijos el ilustrísimo y Reverendísimo Señor ... Angel Maldonado	México	s.n.		1717	LNM			Vida espiritual - Iglesia Católica	Perfección	Arrepentimiento	
3.	3		2 Feb 1722	1722-02-01	16	Peraña, Antonio de	<a href="http://viaf.org/viaf/310508316">http://viaf.org/viaf/310508316</a>	Jesuita	Dissertationes scholasticæ de Sacratissima Virgine Maria Genitrice Dei Nostra que etiam diecissima Mater, ac Domina. Brevi...	México	Miguel de Robles Calderón, Herederos de la Viuda de		1721	LNM			Virgen María	Escolasticismo		
4.	4		2 Feb 1722	1722-02-01	16	Castoreña y Ursúa, Juan Ignacio de	<a href="http://viaf.org/viaf/12658442">http://viaf.org/viaf/12658442</a>	Canónigo	Ocupacion angelica dolorosa, de los mil angeles Manranos y el	México	s.n.		1720	LNM	Librería de los Herederos de la Viuda de Miguel de		Virgen María -- Meditaciones	Jesucristo -- Pasión -- Oraciones y devociones		

Parse data as

Excel files

JSON files

Line-based text files

CSV / TSV / separator-based files

Fixed-width field text files

Worksheets to Import

☒ 20180503\_Gacetas\_Base.xlsx#Hoja1

1673 rows

☐ Ignore first

☒ Parse next

☐ Discard initial

☐ Load at most

0 line(s) at beginning of file

1 line(s) as column headers

0 row(s) of data

0 row(s) of data

☒ Store blank rows

☒ Store blank cells as nulls

☐ Store file source (file names, URLs) in each row

Update Preview

# Sacando jugo a las funciones de OpenRefine

## Conoce tus datos

Lo primero que debes hacer es echar un vistazo general y conocer tus datos. Puedes inspeccionar los diferentes valores de datos mostrándolos en **facetas**.

### 🤔 ¿Saben lo que es una faceta?

Se podría considerar una **faceta** como una lente a través de la cual se visualiza un subconjunto específico de los datos, basado en un criterio de su elección.

### Primer ejercicio:

Facetas de **texto**.

- Explorar la faceta de autor (click en triángulo de la columna, facet -- text facet)
  - i. Explorar los datos en orden alfabético (sort by — name)
  - ii. Explorar las celdas por orden de frecuencia (sort by —count)

🤔 ¿Quién son lo(a)s autor(e/a)s con más artículos?

- i. Por último utiliza la faceta de longitud de texto (facet - customized facet - text length facet)
- 😬 ¿Qué te llama la atención de las diferentes longitudes de los nombres de autor?
- 😬 ¿Cómo seleccionarías los autores con más de 50 caracteres de longitud?

**Refine** OPEN CEDUA\_limpiar [Permalink](#)

Facet / Filter Undo / Redo

Refresh Reset All Remove All Show as: table

**Autor** change 24 choices Sort by: name count Cluster

Castillo García Manuel Ángel;  
Movilidad transfronteriza entre Chiapas y  
Guatemala: políticas migratorias y de  
seguridad en el contexto actual 1

Echarri Cánovas Carlos Javier;  
Nancy Escalante Rivas, Roberto Ham  
Chande 1

Giorguli Saucedo Silvia Elena;  
Bryant, Jensen y Eduardo, Hernández,  
Padilla 1

Giorguli Saucedo Silvia Elena;  
Eduardo, Torre Cantalapiedra 1

Giorguli Saucedo Silvia Elena;  
Jorge Durand Douglas S. Massey, Karen

**Autor** change reset

50.00 — 150.00

Star	Comment	Count
☆	🗨️	3.
☆	🗨️	14.
☆	🗨️	36.
☆	🗨️	38.
☆	🗨️	50.

Recuerda que se puede dar "Reset all" para limpiar nuestras selecciones





- Seguir exploración con la faceta de título de artículo

Nosotros sabemos que los artículos deberían aparecer una sola vez.

😞¿Cómo darnos cuenta si hay duplicados?

😞¿Cómo visualizar únicamente solamente todos los duplicados?

- Ordenarlos por frecuencia
- Agregar faceta de duplicados (facet - customized facet - duplicates) y seleccionar todos aquellos que sí sean duplicados (true)

Estas facetas nos han permitido observar ciertas inconsistencias. Sin embargo, existen tipos de inconsistencias que no se ven tan fácil usando estas facetas. Para esto, OpenRefine ofrece otro tipo de herramientas llamadas “Agrupaciones”.

## Agrupaciones

Como se muestra en la figura siguiente el agrupamiento te permite visualizar y resolver duplicidades inexactas (registros escritos ligeramente diferente, y por lo tanto considerados como entidades diferentes, pero que hacen referencia a la misma identidad) .

*OpenRefine* detecta los valores relacionados y propone fusionarlos en la variante con mayor frecuente (aunque te permite definir un valor específico).

Primer ejercicio

Ve a la columna de “Nombre de revista” (edit cells — cluster and edit // o bien si está abierto su faceta de texto, dar clic en el botón de “cluster”)

1. La primera opción de “colisión de llaves” o “key collision” es “fingerprint”.

**Fingerprint** es un método fácil y simple. Quita todos los espacios en blanco, cambia todos los caracteres a minúsculas, remueve toda la puntuación y normaliza cualquier caracter especial a una versión estándar. Luego, parte el texto y aplica espacios en blanco. Así encuentra las coincidencias (por ejemplo, la clave de las variantes: “méxico, México, Mexico!” con fingerprint es “mexico” [todo en minúsculas sin caracteres especiales], por lo que las variantes son computadas como

## Cluster & Edit column "Nombre de la revista"

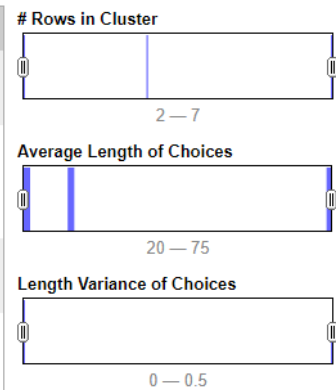
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer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

Method key collision

Keying Function fingerprint

3 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	4	<ul style="list-style-type: none"> <li>Derecho Ambiental y Ecología (3 rows)</li> <li>Derecho ambiental y ecología (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Derecho Ambiental y Ecología
2	2	<ul style="list-style-type: none"> <li>Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía (1 rows)</li> <li>Realidad, datos, espacio: Revista Internacional de Estadística y Geografía (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Realidad, Datos y Espacio. Revista
2	7	<ul style="list-style-type: none"> <li>Papeles de Población (6 rows)</li> <li>Papeles de población (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Papeles de Población



Select All

Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close

Close

Seleccionen a qué valor se deben normalizar las variaciones (recuerden que si los valores de "New Cell Value" no son correctos pueden editarlos). Cuando esté listo dar clic en "Merge Selected & Re-Cluster".

2. Ahora en "Keying Function" o "Función" selección n-gram-fingerprint

**N-Gram Fingerprint**, este método permite agrupar términos con distancias de caracteres distintas determinadas por el usuario. Por ejemplo: Rodrigo y Rodrigon tiene un caracter de distancia (la "n") y Silvia y Silviát! tiene tres caracteres de distancia (un acento, una "t" y una "!")

Modifiquen el número de caracteres de distancia y observen qué pasa

## Cluster & Edit column "Nombre de la revista"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, "New York" and "New York City" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person.

Method key collision ▼

Keying Function ngram-fingerprint ▼

Ngram Size 10

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	9	<ul style="list-style-type: none"><li>• <a href="#">Quid 16</a> (2 rows)</li><li>• <a href="#">Defensor</a> (1 rows)</li><li>• <a href="#">Elsevier</a> (1 rows)</li><li>• <a href="#">GénEros</a> (1 rows)</li><li>• <a href="#">Letra S</a> (1 rows)</li><li>• <a href="#">Soc. Sci</a> (1 rows)</li><li>• <a href="#">Springer</a> (1 rows)</li><li>• <a href="#">Water</a> (1 rows)</li></ul>	<input type="checkbox"/>	<input type="text" value="Quid 16"/>

3. Ahora en "Keying Function" o "Función" seleccionen "metaphone"

**Metaphone:** Este método no revisa los caracteres textuales sino cómo se pronunciarían.

Identifiquen qué valores sí son el mismo y agrúpenlos.

Si en algún momento necesitas desplegar los registros de una agrupación puedes utilizar "Browse this cluster" para hacer las ediciones correspondientes.

Por ejemplo en este grupo sólo dos sí son el mismo registro.

- [Revista Latinoamericana de Población](#) (3 rows)
- [Revista Latinoamericana de Recursos Naturales](#) (2 rows)
- [Revista Latinoamericana de Población RELAP](#) (1 rows)

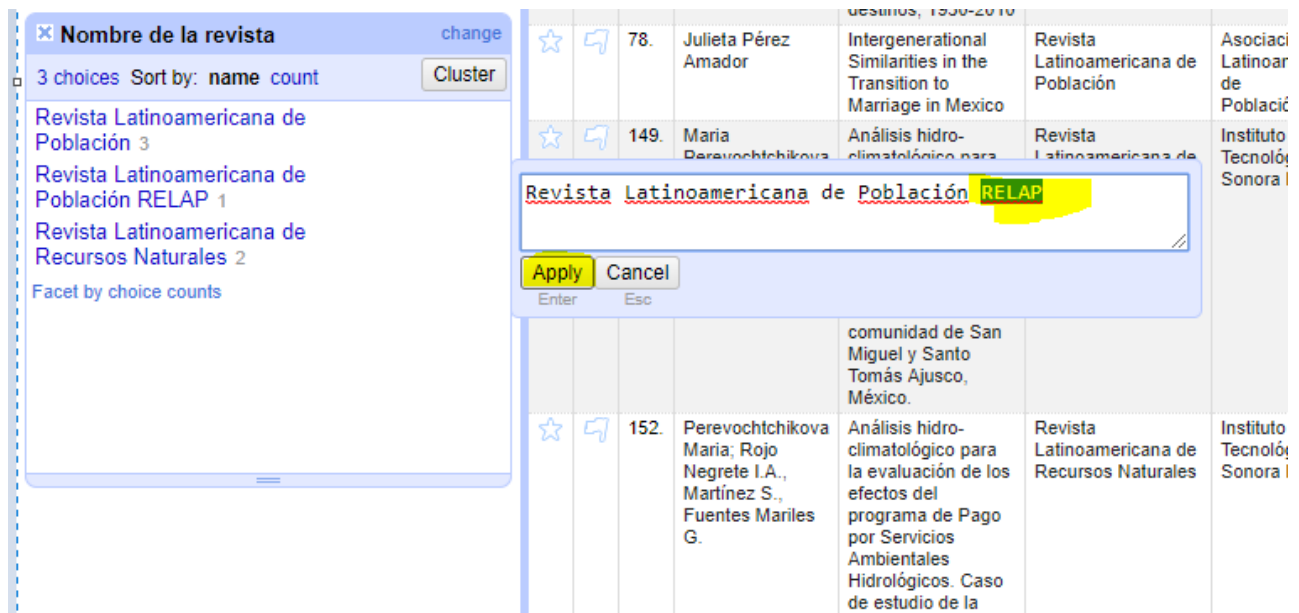
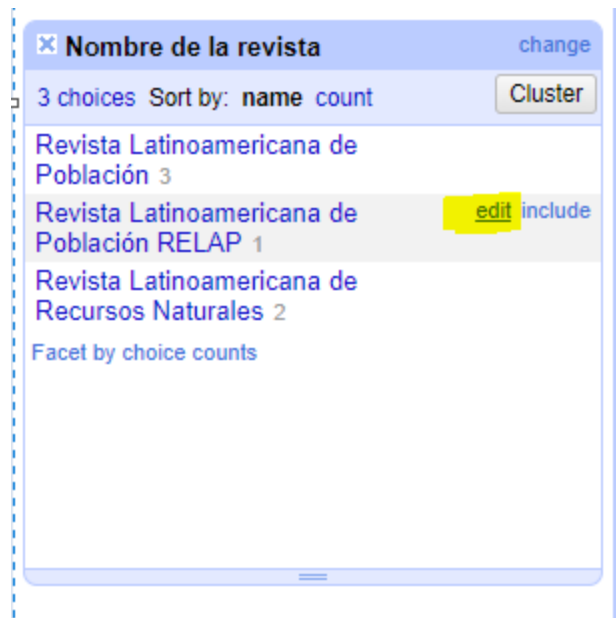
6

- [Revista Latinoamericana de Población](#) (3 rows)
- [Revista Latinoamericana de Recursos Naturales](#) (2 rows)
- [Revista Latinoamericana de Población RELAP](#) (1 rows)

☒

[Browse this cluster](#)

Esto los lleva al cluster con los seis registros y ahí, en una faceta de texto pueden editar el “Revista Latinoamericana de Población RELAP” para que aparezca como “Revista Latinoamericana de Población”:



Recuerden seleccionar **sólo** los casos que sí correspondan

**Cluster & Edit column "Nombre de la revista"**

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: **metaphone3**      **12 clusters found**

Cluster	Count	Members	Preview
2	3	<ul style="list-style-type: none"> <li>Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía (2 rows)</li> <li>Realidad, Datos y Espacios, Revista Internacional de Estadística y Geografía (1 rows)</li> </ul>	Realidad, Datos y Espacio. Revista
2	2	<ul style="list-style-type: none"> <li>Environmental Earth Sciences (1 rows)</li> <li>Environmental policies in the peri-urban area of Mexico City: The perceived effects of three environmental programs (1 rows)</li> </ul>	Environmental Earth Sciences
2	2	<ul style="list-style-type: none"> <li>Boletín Editorial del Colegio de México (1 rows)</li> <li>Boletín Editorial, El Colegio de México, núm. 170 (1 rows)</li> </ul>	Boletín Editorial del Colegio de Méx
2	2	<ul style="list-style-type: none"> <li>Acta Universitaria Multidisciplinary Scientific Journal (1 rows)</li> <li>Acta Universitaria, (1 rows)</li> </ul>	Acta Universitaria Multidisciplinary
2	2	<ul style="list-style-type: none"> <li>The Annals of the American Academy and Social Science (1 rows)</li> <li>The Annals of the American Academy of Political and Social Sciences (1 rows)</li> </ul>	The Annals of the American Acadei

**# Choices in Cluster**

**# Rows in Cluster**

**Average Length of Choices**

**Length Variance of Choices**

Repite esta misma operación con la columna de Autor y Editorial

Presentación: [https://prezi.com/ucjcfi\\_6j-qd/?utm\\_campaign=share&utm\\_medium=copy](https://prezi.com/ucjcfi_6j-qd/?utm_campaign=share&utm_medium=copy)