Introducción a OpenRefine

Por: Silvia Gutiérrez @espejolento (para el taller con Rodrigo Cuéllar)
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/1IkIH_E8XJFCorxEqlKLoPA_Ihy7bRoaT
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)

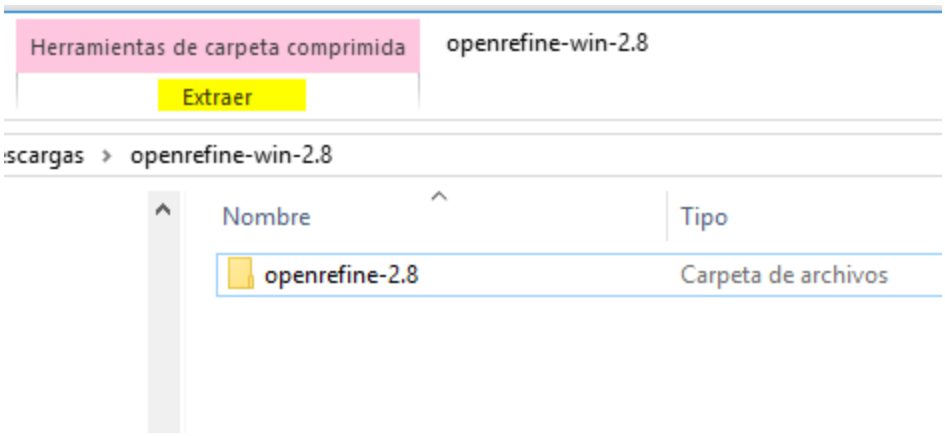


Imagen 1

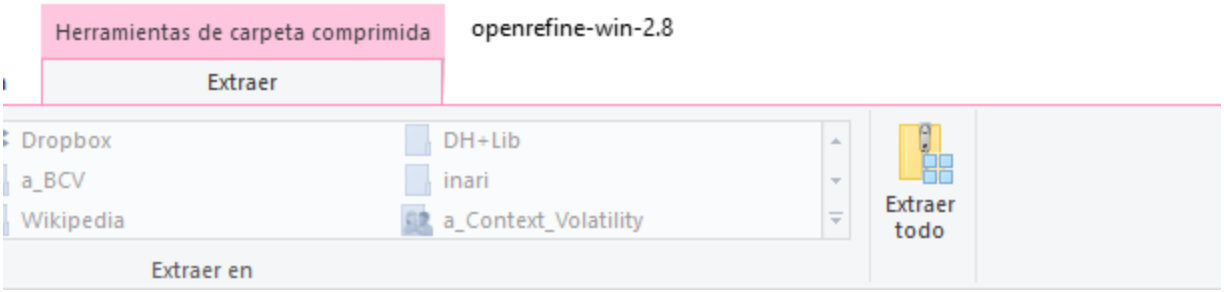
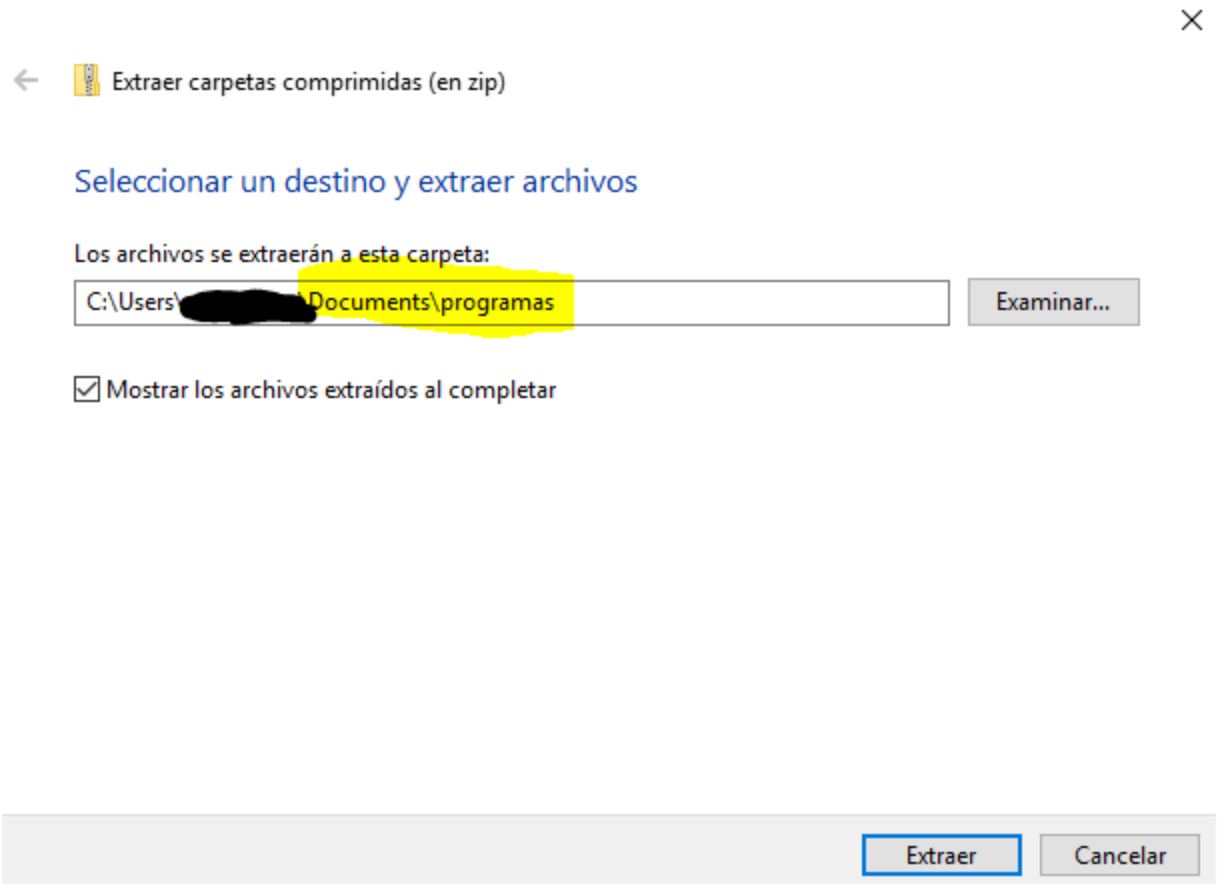


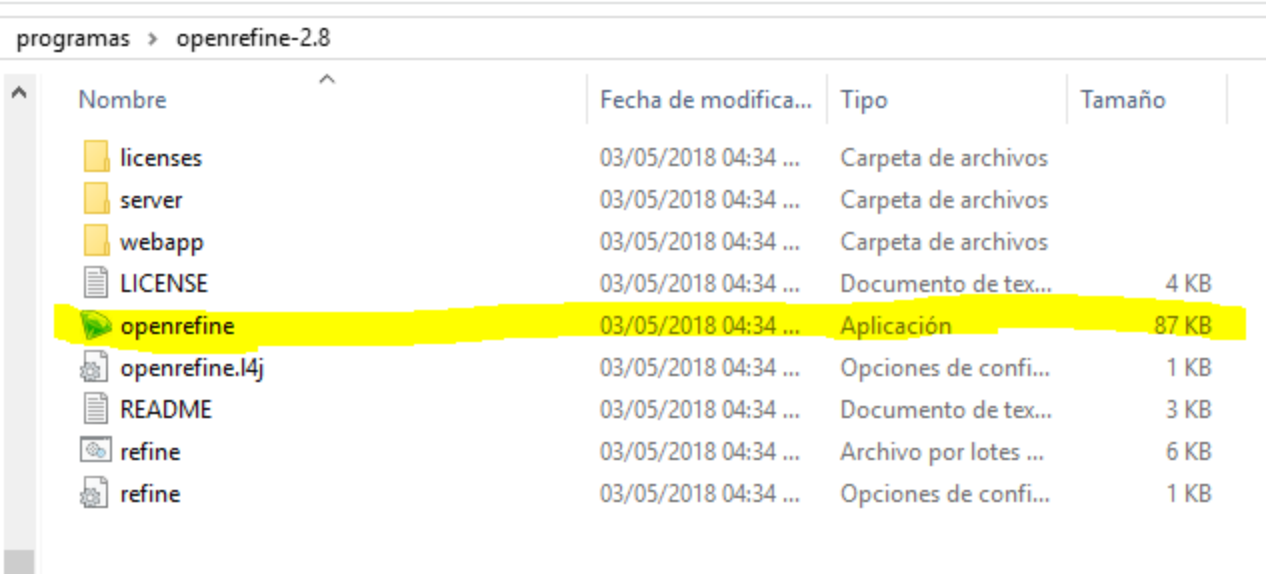
Imagen 2

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)

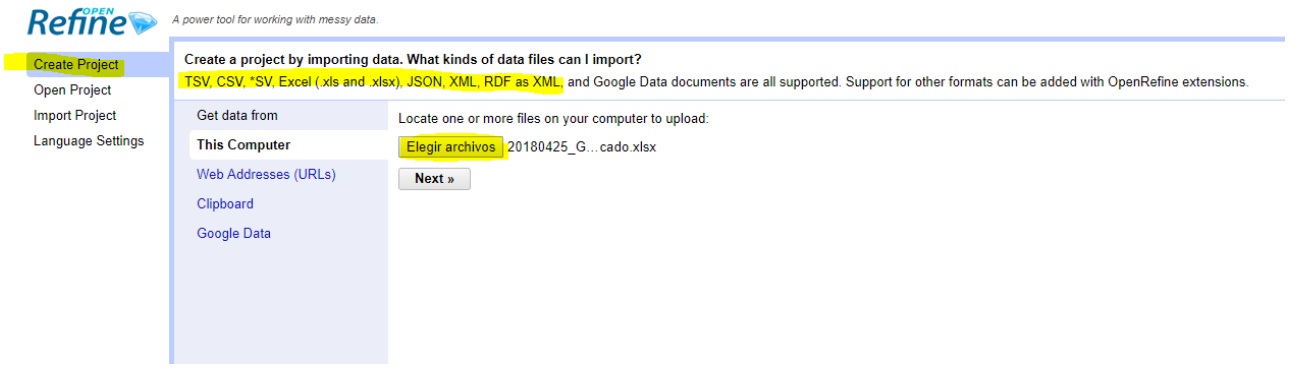


⚠️ Nota: Necesitarás tener Java instalado en tu computadora, si no lo tienes, ve a: https://www.java.com/es/download/)

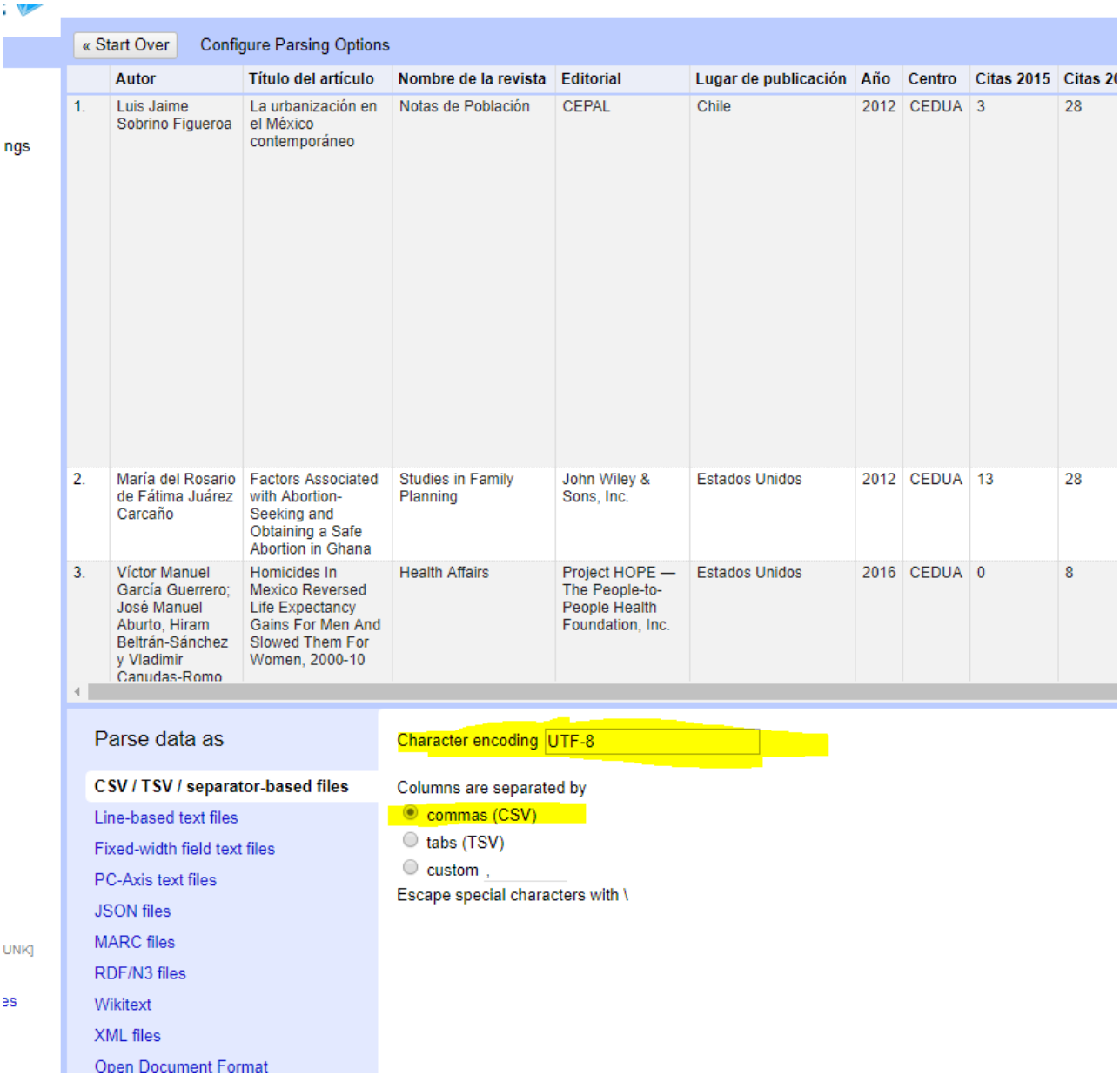
Cargando tus datos a OpenRefine

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)**

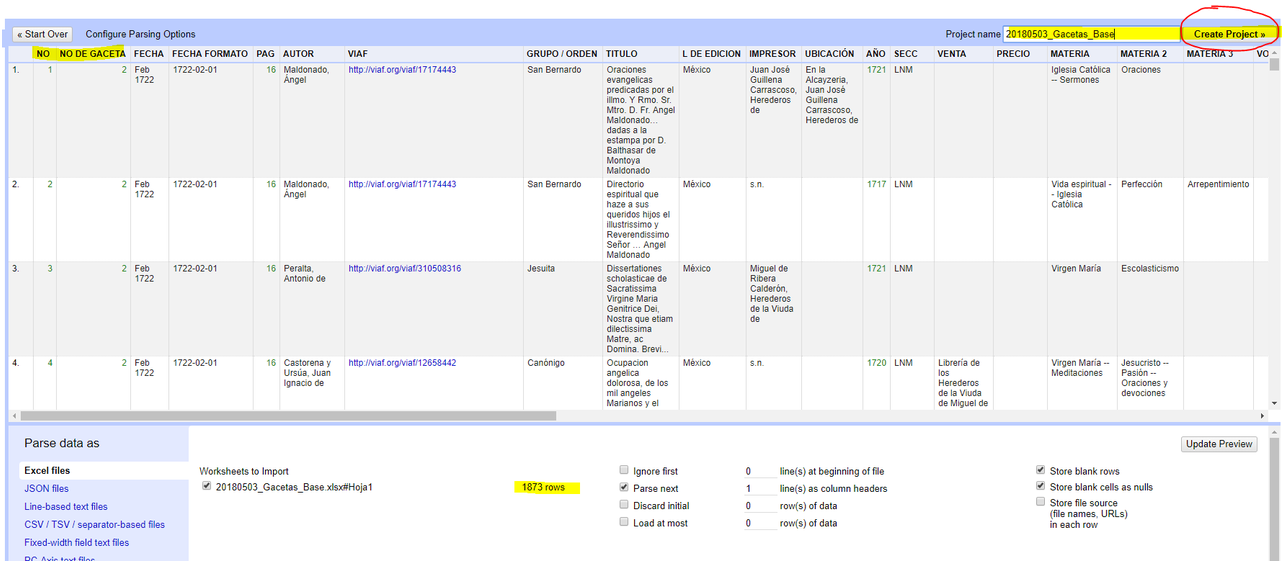


2. Se verifica que se interpreten los datos como CSV, que el número de filas coincida con los datos originales y que la primera fila sea el encabezado.
   ⚠️ OJO: Si salen letras “raras” es muy probable que el “encoding” no sea correcto. Hay que seleccionar “UTF-8” el cual corresponde a nuestro alfabeto y su forma de acentuación y puntuación.



	Autor	Título del artículo	Nombre de la revista	Editorial	Lugar de publicación	Año	Centro	Citas 2015	Citas 2(
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

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



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?

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?

i. Ordenarlos por frecuencia
ii. 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) .

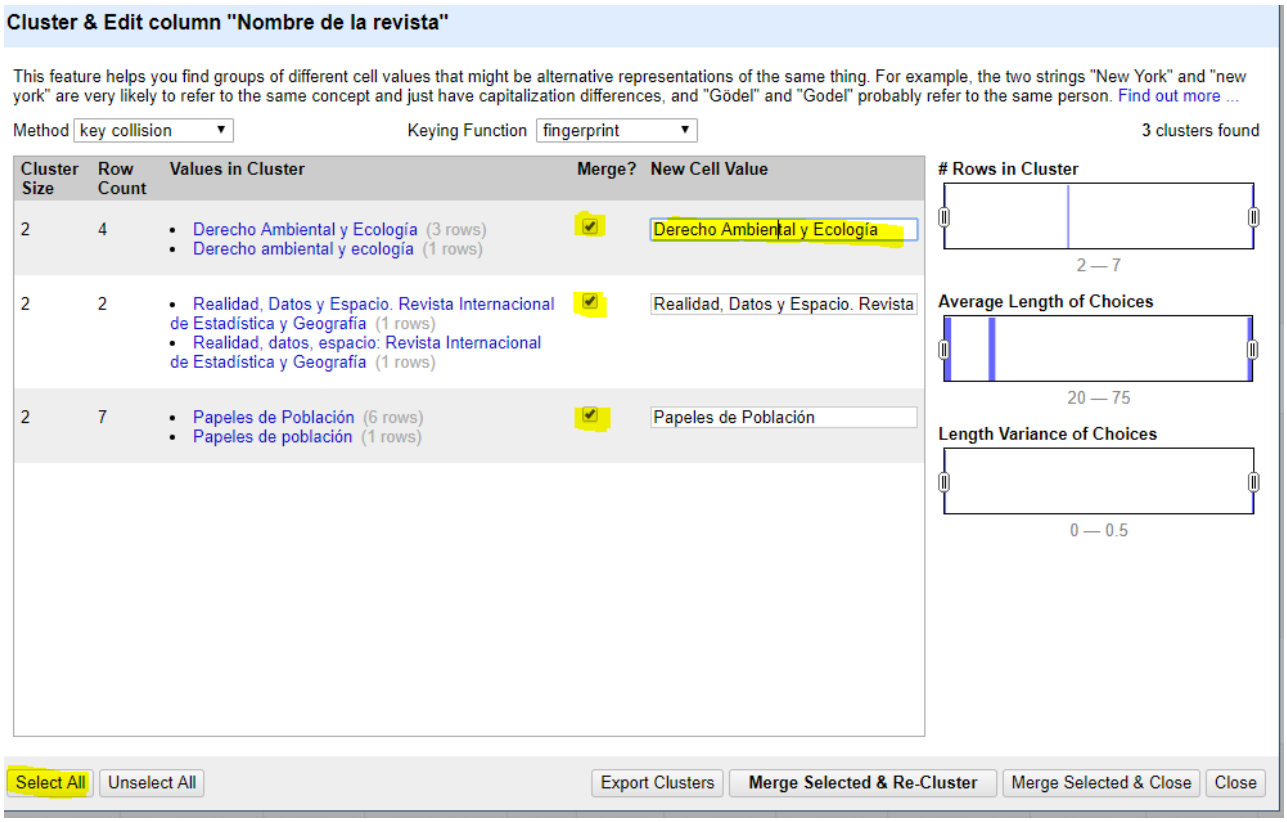
OpenRefine detecta los valores relacionados y propone fusionarlos en la variente con mayor frecuente (aunque te permite definir un valor específico).

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

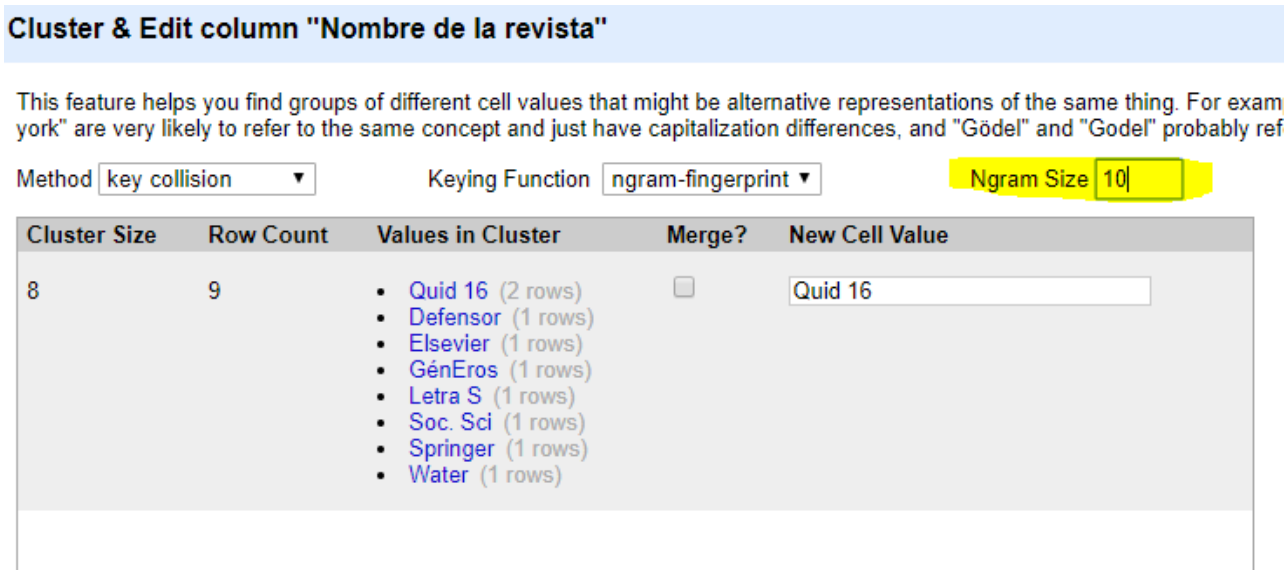


Seleccionen a qué valor se deben normalizar las variaciones (recuerden que si los valores de “New Cell Value” no son correctos pueden editarlos). Cuando esté listo dar clic en “Merge Selected & Re-Cluster”.

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



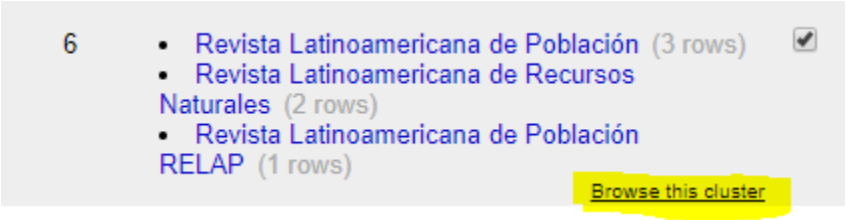
3. Ahora en “Keying Function” o “Función” seleccionen “metaphone”

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

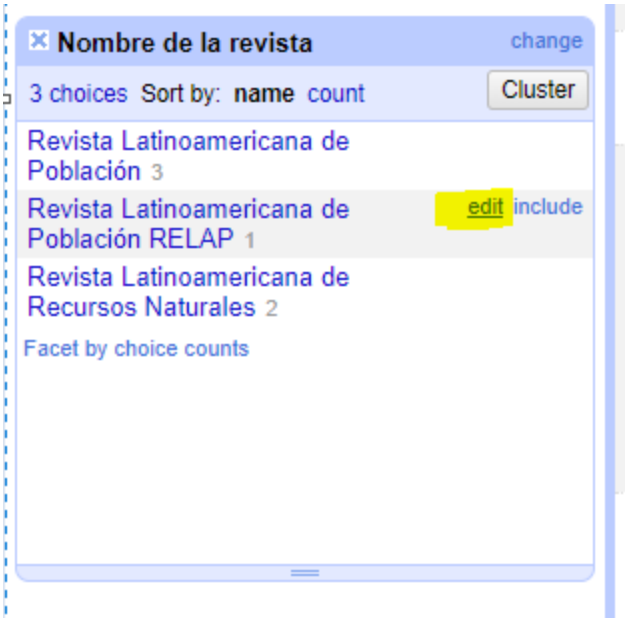
Identifiquen qué valores sí son el mismo y agrúpenlos.

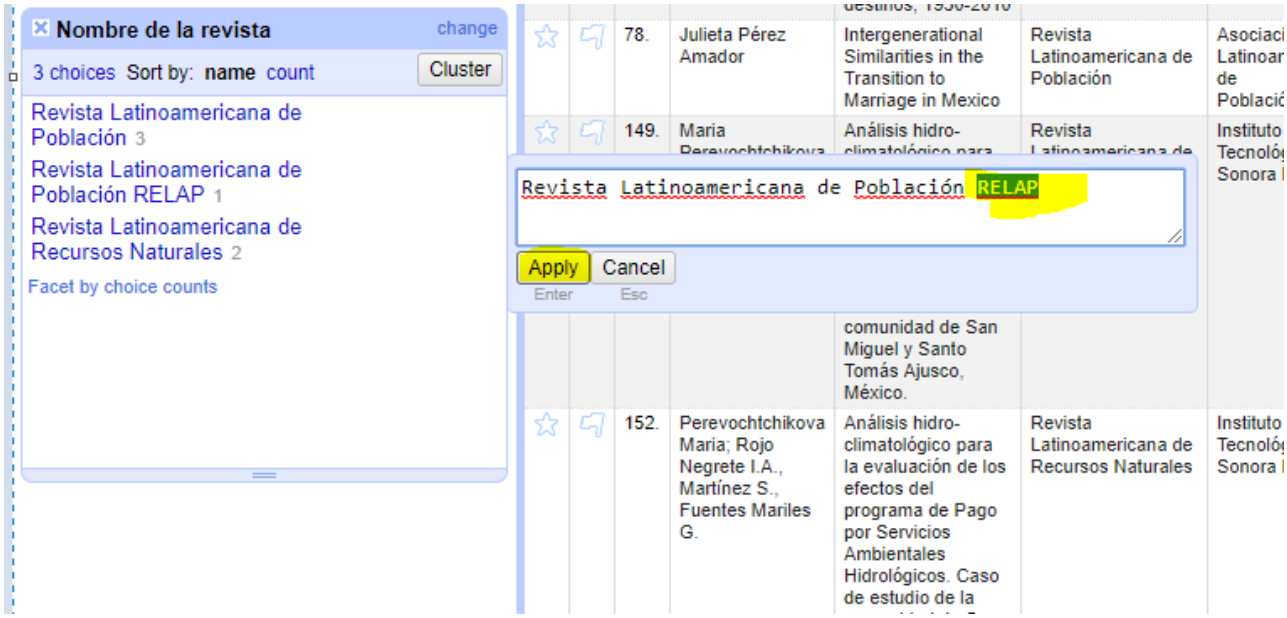
Si en algún momento necesitas despeglar los registros de una agrupación puedes utlizar “Browse this cluster” para hacer las ediciones correspondientes.
Por ejemplo en este grupo sólo dos sí son el mismo registro.

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

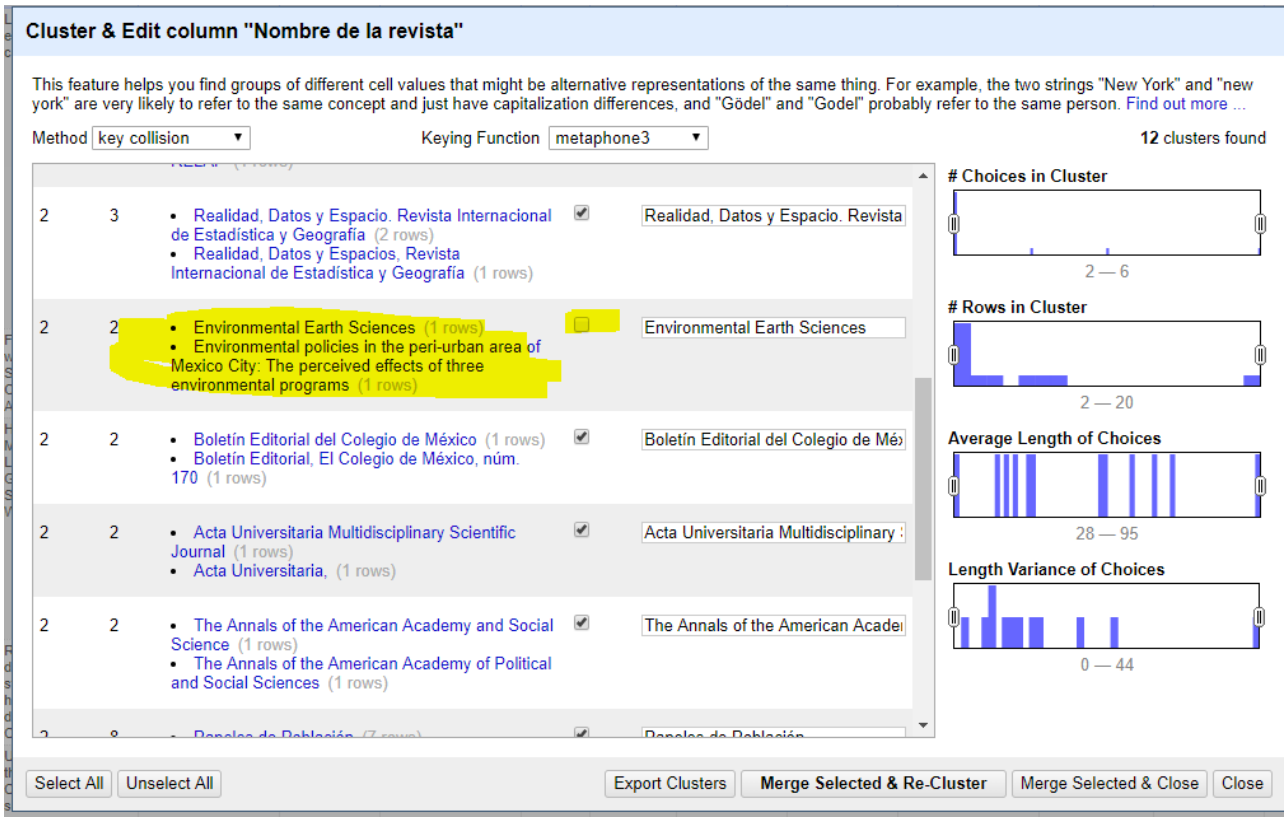


Esto los lleva al cluster con los seis registros y ahí, en una faceta de texto pueden editar el “Revista Latinoamericana de Población RELAP” para que aparezca como “Revista Latinoamericana de Población”:





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



Repite esta misma operación con la columna de Autor y Editorial

Presentación: https://prezi.com/ucjcfi_6j-qd/?utm_campaign=share&utm_medium=copy