

INSTRUCCIONES Y EJEMPLOS DE USO DE OPENREFINE EN LA BNE

(VERSIÓN 2019)

1. Datos a utilizar	2
2. Crear un proyecto.....	2
3. Uso de las facetas o filtros.....	4
4. Agrupar valores	6
4.1. Métodos de agrupamiento	7
4.1.1. Key Collision Methods o Métodos de colisión clave	7
4.1.2. Métodos vecino más cercano	9
4.2. Ejemplos de agrupamiento	10
4.2.1. Usando colisión de llaves	10
4.2.2. Usando el método del vecino más cercano / Levenshtein:	11
4.2.3. Utilizando el método del vecino más cercano / PPM:	12
5. Las celdas multivaluadas	17
6. Enriquecimiento con fuentes externas.....	22
6.1. Ejemplo 1: Autores en dominio público.....	23
6.1.1. Reconciliar datos con Wikidata.....	23
6.1.2. Reconciliar datos con fuentes propias: reconciliar con un subconjunto de Wikidata obtenido mediante consulta SPARQL.....	25
6.1.3. Crear columnas basadas en datos vinculados	27
6.2. Ejemplo 2: Videgrabaciones publicadas en 2010	28
6.2.1. Reconciliar datos con Wikidata.....	28
6.2.2. Reconciliar datos con fuentes propias: Reconciliar con catálogo de películas calificadas del ICAA 29	
6.2.3. Crear columnas basadas en datos vinculados	30

OpenRefine es una herramienta, de código abierto, utilizada para la limpieza y transformación de datos. Se puede descargar e instalar en cualquier equipo, si bien hay que considerar que para trabajar con ficheros pesados, que contengan gran cantidad de datos, es posible que el equipo requiera tener una memoria de entre 8 y 16 GB de RAM.

Enlace a la descarga: <http://openrefine.org/download.html>

Instalación: descomprimir en C:/OpenRefine y ejecutar el .exe

1. Datos a utilizar

Los conjuntos de datos que vamos a utilizar serán, en principio, los catálogos bibliográfico y de autoridades. Estos conjuntos, originariamente en formato MARC y MARC-XML, han pasado por un proceso de mapeo para ser publicados en formatos reutilizables. De este modo, cada campo y subcampo MARC aparecen ahora bajo una serie de literales que hemos definido.

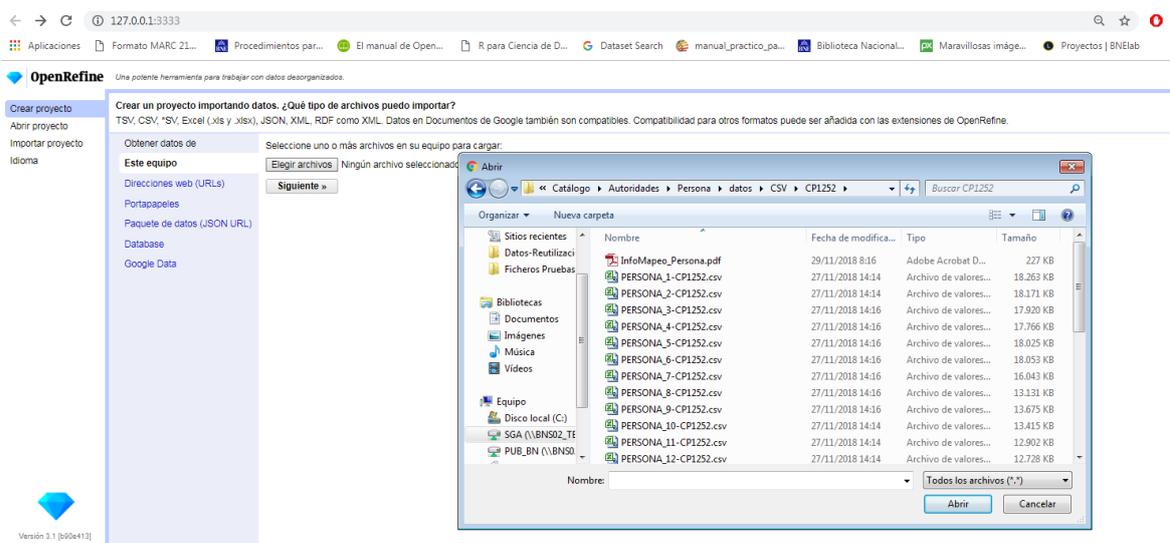
Los conjuntos de datos, la matriz de mapeo, así como las explicaciones asociadas, están ya publicadas en la web de BNElab:

- [Catálogo bibliográfico](#)
- [Catálogo de autoridades](#)

2. Crear un proyecto

Para empezar, debemos crear un proyecto, cargando los datos con los que queremos trabajar.

1. En nuestro ejemplo, utilizaríamos los ficheros en formato CSV que hemos generado como datos abiertos; concretamente vamos a cargar todo el [catálogo de autoridades de persona](#).



- Para que los datos sean manejables, hemos dividido el fichero de origen, en ficheros de 50.000 registros. Podemos seleccionar todos los ficheros de forma simultánea, de un mismo tipo. Por ejemplo, cargamos los 27 ficheros de autoridades persona.

OpenRefine Una potente herramienta para trabajar con datos desorganizados.

Crear un proyecto importando datos. ¿Qué tipo de archivos puedo importar?
TSV, CSV, *SV, Excel (.xls y .xlsx), JSON, XML, RDF como XML. Datos en Documentos de Google también

Obtener datos de: Este equipo (27 archivos), Direcciones web (URLs), Portapapeles, Paquete de datos (JSON URL), Database, Google Data

Seleccione uno o más archivos en su equipo para cargar:
Elegir archivos 27 archivos
Siguiente »

- OpenRefine nos muestra los ficheros que ha cargado, su formato, nombre, tamaño, etc. El siguiente paso será configurar las opciones de carga.

¿Importar?	Nombre	Tipo MIME	Formato	Tamaño
<input checked="" type="checkbox"/>	PERSONA_1-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	17.8 MB
<input checked="" type="checkbox"/>	PERSONA_2-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	17.7 MB
<input checked="" type="checkbox"/>	PERSONA_3-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	17.5 MB
<input checked="" type="checkbox"/>	PERSONA_4-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	17.3 MB
<input checked="" type="checkbox"/>	PERSONA_5-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	17.6 MB
<input checked="" type="checkbox"/>	PERSONA_6-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	17.6 MB
<input checked="" type="checkbox"/>	PERSONA_7-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	15.7 MB
<input checked="" type="checkbox"/>	PERSONA_8-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	12.8 MB
<input checked="" type="checkbox"/>	PERSONA_9-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	13.4 MB
<input checked="" type="checkbox"/>	PERSONA_10-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	13.1 MB
<input checked="" type="checkbox"/>	PERSONA_11-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	12.6 MB
<input checked="" type="checkbox"/>	PERSONA_12-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	12.4 MB
<input checked="" type="checkbox"/>	PERSONA_13-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	13.4 MB
<input checked="" type="checkbox"/>	PERSONA_14-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	14.9 MB
<input checked="" type="checkbox"/>	PERSONA_15-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	15.1 MB
<input checked="" type="checkbox"/>	PERSONA_16-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	14.6 MB
<input checked="" type="checkbox"/>	PERSONA_17-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	16.2 MB
<input checked="" type="checkbox"/>	PERSONA_18-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	14.8 MB
<input checked="" type="checkbox"/>	PERSONA_19-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	17.5 MB
<input checked="" type="checkbox"/>	PERSONA_20-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	18.6 MB
<input checked="" type="checkbox"/>	PERSONA_21-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	19.1 MB
<input checked="" type="checkbox"/>	PERSONA_22-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	19 MB
<input checked="" type="checkbox"/>	PERSONA_23-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	18.7 MB
<input checked="" type="checkbox"/>	PERSONA_24-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	20.2 MB
<input checked="" type="checkbox"/>	PERSONA_25-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	22.4 MB
<input checked="" type="checkbox"/>	PERSONA_26-CP1252.csv	application/vnd.ms-excel	textline-based/*sv	20.5 MB

- Ahora podemos darle un nombre al proyecto y veremos las columnas que creará, que se corresponden con los datos de nuestro mapeo¹. Pulsaríamos en “Crear proyecto”.

¹ La información sobre el mapeo realizado para pasar los catálogos del MARC-XML a los formatos reutilizables, está disponible al descargar cualquier conjunto en datos.gob.es, bajo el nombre “InfoMapeo”.

OpenRefine Una potente herramienta para trabajar con datos desorganizados.

Crear proyecto Inicio Configurar opciones de carga Nombre del proyecto Personas_catálogo_completo Tags Crear proyecto

Abrir proyecto Importar proyecto Idioma

File	id BNE	otros códigos de identificación	fecha de nacimiento	fecha de fallecimiento	nombre de persona	otros atributos de persona	lugar de nacimiento	lugar de fallecimiento	país	otros lugares asociados	dirección	campo de
PERSONA_1_ CP1252.csv	XX95384				Carnelfo (Familia)							
PERSONA_1_ CP1252.csv	XX95608				Saavedra, Cesa de							
PERSONA_1_ CP1252.csv	XX95777				Castanedo (Familia)							
PERSONA_1_ CP1252.csv	XX111504				Espareguera (Familia)							
PERSONA_1_ CP1252.csv	XX113622				Folc (Familia)							
PERSONA_1_ CP1252.csv	XX115812				Gari (Familia)							
PERSONA_1_ CP1252.csv	XX117519				Guasp (Familia)							
PERSONA_1_ CP1252.csv	XX123992				Kennedy (Familia)							
PERSONA_1_ CP1252.csv	XX127572				Moñino (Familia)							
PERSONA_1_ CP1252.csv	XX129572				Nicolau (Familia)							
PERSONA_1_ CP1252.csv	XX130843				Ostolaza (Familia)							

Abrir archivo como Codificación de caracteres Actualizar vista previa

Las columnas se encuentran separadas por comas (CSV) tabulaciones (TSV) personalizado: ; Ignorar caracteres especiales con \

Nombres de columna (separados por comas):

Ignorar primera(s) 0 línea(s) al inicio del archivo
 Seleccionar primera(s) 1 línea(s) para los nombres de las columnas
 Descartar primera(s) 0 fila(s) de datos
 Cargar al menos 0 fila(s) de datos
 Usar carácter * para encerrar celdas que contengan separadores de columnas

Detectar y transformar texto en números, fechas, ... Cargar filas en blanco Cargar celdas en blanco como nulas Cargar el origen del archivo (nombres, URLs) en cada fila

5. En este punto, ya tenemos creado nuestro proyecto en OpenRefine para poder trabajar con él.

OpenRefine Personas_catálogo_completo Enlace permanente Abrir... Exportar Ayuda

Facetas / Filtros Extensiones: Wikidata

Deshacer / Rehacer 0/0 Mostrar como: filas registros Mostrar: 5 10 25 50 filas « primera » anterior 1 - 10 siguiente » última »

Todo	File	id BNE	otros códigos de	fecha de naciomi	fecha de fallecimi	nombre de pers	otros atributos d	lugar de naciomi	lugar de fallecimi	país	otros lugares as	dirección	campo de activi	filas
1	PERSONA_1_ CP1252.csv	XX95384				Carnelfo (Familia)								
2	PERSONA_1_ CP1252.csv	XX95608				Saavedra, Cesa de								
3	PERSONA_1_ CP1252.csv	XX95777				Castanedo (Familia)								
4	PERSONA_1_ CP1252.csv	XX111504				Espareguera (Familia)								
5	PERSONA_1_ CP1252.csv	XX113622				Folc (Familia)								
6	PERSONA_1_ CP1252.csv	XX115812				Gari (Familia)								
7	PERSONA_1_ CP1252.csv	XX117519				Guasp (Familia)								
8	PERSONA_1_ CP1252.csv	XX123992				Kennedy (Familia)								
9	PERSONA_1_ CP1252.csv	XX127572				Moñino (Familia)								
10	PERSONA_1_ CP1252.csv	XX129572				Nicolau (Familia)								

Usar facetas y filtros Use las facetas y los filtros para seleccionar subconjuntos de sus datos y trabajar en ellos. Puede encontrar estas opciones en los menús de cada columna. ¿Problemas para comenzar? Vea los videos de ayuda

6. Por defecto, nos muestra los resultados de las primeras diez, pero esto se puede configurar fácilmente para adaptar la vista.

3. Uso de las facetas o filtros

Las facetas son filtros que se aplican al campo que seleccionemos. Funcionan de forma similar a un simple filtro en una hoja de cálculo, con la ventaja de que en OpenRefine podemos trabajar con los 27 ficheros unificados. Es lo que recomendamos utilizar para ver de una forma sencilla la diversidad del catálogo, especialmente en campos normalizados.

Por ejemplo, aplicando la faceta de texto al campo Género (375 \$a), veremos que nuestro catálogo ofrece 31 variantes de posibles respuestas, cuando lo lógico sería que hubiera tres, a saber, Masculino/Femenino/Vacio (campo no rellenado).

The screenshot shows the OpenRefine interface with a table of 1329370 rows. The 'género' column has a dropdown menu open, listing various filtering options like 'Faceta de texto', 'Faceta numérica', etc. The table contains names and family names such as 'Carrefo (Familia)', 'Saeveda, Casa de', 'Castanedo (Familia)', etc.

En la columna de la izquierda, se nos mostrará un cuadro con el listado de todas las opciones que el sistema encuentra para ese campo en concreto.

The screenshot shows the 'género' facet panel on the left side of the OpenRefine interface. It lists 31 choices with counts, such as 'Femenino 231375', 'Masculino 9925', and 'Maculino 7'. The main table shows the filtered results for the selected facet.

Haciendo clic sobre cada una de las opciones, la vista nos mostrará los registros que se corresponden con ese valor. Por ejemplo, "Maculino" se repite en 7 registros.



4.1. Métodos de agrupamiento

Existen dos grandes métodos de agrupamiento, a saber: métodos de colisión clave y métodos del vecino más cercano.

4.1.1. Key Collision Methods o Métodos de colisión clave

Este método se basa en la idea de crear una representación (una 'llave') que contiene solo la parte más valiosa o significativa de la cadena de texto para posteriormente asociar diferentes cadenas, basado en el hecho de que su 'llave' es la misma (de ahí el nombre de 'colisión clave'). Es uno de los métodos más rápidos, su análisis es lineal y permite agrupar, en segundos, millones de registros.

Los métodos de colisión clave disponibles en Open Refine se encuentran repartidos de la siguiente forma:

4.1.1.1. Fingerprint o Huella Digital:

Este método se caracteriza por ser rápido y simple, es el que tiene menos probabilidades de producir falsos positivos y funciona relativamente bien en una variedad de contextos. A continuación se lista el proceso que realiza el algoritmo:

- Quita espacio inicial y final de la cadena.
- Cambia todos los caracteres por su equivalente en minúsculas.
- Elimina todos los signos de puntuación y de control.
- Divide la cadena en fichas separadas por espacios en blanco.
- Ordena las fichas y elimina duplicados.
- Une las letras de nuevo.
- Normaliza caracteres occidentales extendidos a su representación ASCII; por ejemplo: 'averigüé' → 'averigue'.

Este algoritmo funciona muy bien a la hora de identificar casos en los que hay comas, puntos, tildes o mayúsculas que diferencien, en teoría, los nombres, pero que realmente representan lo mismo. Ejemplo de agrupación según este método:

```
"Biblioteca Pedro-Martínez" → "Biblioteca Pedro-Martínez"  
"Biblioteca Pedro-Martínez" → "biblioteca pedro-martínez"  
"biblioteca pedro-martínez" → "biblioteca pedro martinez"  
"biblioteca pedro martinez" → "biblioteca Martinez pedro"
```

4.1.1.2. N-Gram Fingerprint o Huella Digital N-Gram:

Este método es similar al de la huella digital anteriormente descrito, con la diferencia de que en lugar de utilizar espacios en blanco separados por fichas, usa n-gramas, donde n (o el tamaño en caracteres de la ficha) puede ser especificado por el usuario. Este algoritmo sigue el siguiente proceso:

- Cambia todos los caracteres a su equivalente en minúsculas
- Elimina todos los caracteres de puntuación, espacios en blanco y de control. Por ejemplo:

```
"León" → "león"  
"león" → "leon"
```

- Obtiene todos los n-gramas
- Ordena los n-gramas y elimina duplicados

```
(2-grama) "leon" > "le" "eo" "on"  
(1-grama) "leon" > "l" "e" "o" "n"  
(2-grama) "le" "eo" "on" > "eoleon"  
(1-grama) "l" "e" "o" "n" > "elno"
```

- Une los n-gramas ordenados de nuevo
- Normaliza caracteres occidentales extendidos a su representación ASCII

Ejemplo: la huella digital de 2-gramas de 'Paris' es "arispari" y la huella digital de 1-grama es "aiprs".

En la práctica, según el software, el uso de valores grandes de n en los n-gramas no produce ninguna ventaja sobre el método de la huella digital anterior, pero utilizando 2-gramas y 1-grama se pueden encontrar grupos que el método anterior no es capaz de encontrar incluso con cadenas que tienen pequeñas diferencias.

4.1.2. Métodos vecino más cercano

Los métodos de colisión clave, descritos anteriormente, son muy rápidos pero tienden a ser muy estrictos o muy laxos, "sin poder afinar qué tanta diferencia entre las cadenas estamos dispuestos a tolerar". Los métodos de vecino más cercano (también conocidos como kNN), proporcionan un parámetro (el radio o k) que representa un umbral de distancia, el cual sirve de referencia para agrupar un par de cadenas de texto si su distancia es cercana.

Los métodos de vecinos más cercanos disponibles en Open Refine se encuentran repartidos de la siguiente forma:

4.1.2.1. Distancia Levenshtein

En términos generales, la Distancia Levenshtein mide el número mínimo de operaciones de edición que se requieren para cambiar una cadena en otra. En la siguiente tabla se explica brevemente en qué consiste la Distancia Levenshtein:

Palabras	Explicación
'París' → 'parís'	Tienen una distancia de edición de 1 debido a que el cambio de P a p es la única operación requerida.
'Nueva York' → 'newyork'	Tiene una distancia de edición de 3: 2 sustituciones y 1 eliminación.

De acuerdo con OpenRefine, esta distancia "es útil para identificar errores tipográficos, errores de ortografía o cualquier cosa que los métodos anteriores no capturan, aunque las grandes distancias dan muchos falsos positivos (especialmente para las cadenas cortas) y no son tan útiles".

4.1.2.2. PPM o Predicción por Coincidencia Parcial

El algoritmo de Predicción por Coincidencia Parcial estima similitudes entre cadenas de texto empleando una operación que mide los recursos empleados por un computador para definir una cadena de texto. La operatividad del algoritmo surge a partir de la forma en que funcionan los compresores de texto; por ejemplo, si dos cadenas de texto A y B son idénticas, al momento de comprimirlas, es decir, al realizar la operación (A+B), se genera muy poca diferencia. Por otro lado, si A y B son muy diferentes, al momento de comprimirlas se deben producir diferencias dramáticas en longitud. La documentación del software recomienda emplear esta técnica como último recurso.

4.2. Ejemplos de agrupamiento

Una vez vistos los métodos, veremos cómo los hemos aplicado a nuestro ejemplo.

4.2.1. Usando colisión de llaves

1329370 filas

Agrupar y editar valores en la columna "género"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. [Más información ...](#)

Método | Colisión de llaves | Función | Huella | 2 clusters Encontrado

Número de valores	Número de filas	Valores en la arupación	¿Unir?	Nuevo valor de las celdas
4	9930	<ul style="list-style-type: none">Masculino (9925 rows)masculino (3 rows)MAsculino (1 rows)Másculino (1 rows)	<input type="checkbox"/>	Masculino
3	231377	<ul style="list-style-type: none">Femenino (231375 rows)Femenino (1 rows)femenino (1 rows)	<input type="checkbox"/>	Femenino

Valores en la agrupación: 3 — 4

Filas en la agrupación: 0 — 240000

Longitud promedio de los valores: 8.333 — 9

Varianza de los valores: 0 — 0.472000000000000003

Seleccionar todos | Seleccionar ninguno | Exportar agrupaciones | Unir seleccionados y reagrupar | Unir seleccionados y cerrar | Cerrar

En este caso todas las opciones propuestas se pueden agrupar, por tanto las seleccionaríamos y "Unir seleccionados y reagrupar"

Agrupar y editar valores en la columna "género"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. [Más información ...](#)

Método: Colisión de llaves Función: Huella 2 clusters Encontrado

Número de valores	Número de filas	Valores en la arupación	¿Unir?	Nuevo valor de las celdas
4	9930	<ul style="list-style-type: none"> Masculino (9925 rows) masculino (3 rows) MAsculino (1 rows) Másculino (1 rows) 	<input checked="" type="checkbox"/>	Masculino
3	231377	<ul style="list-style-type: none"> Femenino (231375 rows) Femenino (1 rows) femenino (1 rows) 	<input checked="" type="checkbox"/>	Femenino

Valores en la agrupación: 3 — 4

Filas en la agrupación: 0 — 240000

Longitud promedio de los valores: 8.333 — 9

Varianza de los valores: 0 — 0.47200000000000003

4.2.2. Usando el método del vecino más cercano / Levenshtein:

En este caso se utilizan 2 variables:

- Radio: Número de movimientos o "cambios" necesarios para que las cadenas de texto sean iguales.
- Caracteres del bloque: Número de caracteres consecutivos que compara.

En nuestro ejemplo, con radio 1 (1 cambio) y 6 caracteres consecutivos, obtenemos las siguientes agrupaciones.

Tiias

Agrupar y editar valores en la columna "género"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. [Más información ...](#)

Método **Vecino más cercano** Function **levenshtein** Radio **1.0** Caracteres del bloque **6** 6 clusters Encontrado

Número de valores	Número de filas	Valores en la agrupación	¿Unir?	Nuevo valor de las celdas
3	9945	<ul style="list-style-type: none"> Masculino (9930 rows) Masculinos (13 rows) Nasculino (2 rows) 	<input type="checkbox"/>	Masculino
2	231378	<ul style="list-style-type: none"> Femenino (231377 rows) Femenimo (1 rows) 	<input type="checkbox"/>	Femenino
2	9937	<ul style="list-style-type: none"> Masculino (9930 rows) Maculino (7 rows) 	<input type="checkbox"/>	Masculino
2	9943	<ul style="list-style-type: none"> Masculino (9930 rows) Masculinos (13 rows) 	<input type="checkbox"/>	Masculino
2	9931	<ul style="list-style-type: none"> Masculino (9930 rows) Masculijno (1 rows) 	<input type="checkbox"/>	Masculino
2	231380	<ul style="list-style-type: none"> Femenino (231377 rows) Femenina (3 rows) 	<input type="checkbox"/>	Femenino

Valores en la agrupación

Filas en la agrupación

Longitud promedio de los valores

Varianza de los valores

Seleccionar todos Seleccionar ninguno

Exportar agrupaciones **Unir seleccionados y reagrupar** Unir seleccionados y cerrar Cerrar

Como todas las propuestas de agrupación son válidas, las aceptamos.

4.2.3. Utilizando el método del vecino más cercano / PPM:

En este caso las variables significan:

- Radio: Número de cambios aceptados para considerar 2 cadenas de caracteres iguales.
- Caracteres de bloque: Número de caracteres que utiliza para partir la cadena original.

Agrupar y editar valores en la columna "género"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. [Más información ...](#)

Método Function Radio Caracteres del bloque 2 clusters Encontrado

Número de valores	Número de filas	Valores en la agrupación	¿Unir?	Nuevo valor de las celdas
2	9954	<ul style="list-style-type: none">Masculino (9953 rows)MasculinoMasculino (1 rows)	<input type="checkbox"/>	<input type="text" value="Masculino"/>
2	231386	<ul style="list-style-type: none">Femenino (231381 rows)FemeninoFemenino (5 rows)	<input type="checkbox"/>	<input type="text" value="Femenino"/>

Filas en la agrupación

Longitud promedio de los valores

Varianza de los valores

En este caso, vuelven a ser válidas las opciones propuestas.

Dada la complejidad de estos algoritmos, conviene realizar diferentes combinaciones de valores de los parámetros para intentar "cazar" el mayor número de opciones erróneas.

Por ejemplo, al ampliar el radio a 9, obtenemos:

Agrupar y editar valores en la columna "género"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. [Más información ...](#)

Método **Vecino más cercano** Funcion **levenshtein** Radio **9** Caracteres del bloque **6** 3 clusters Encontrado

Número de valores	Número de filas	Valores en la agrupación	¿Unir?	Nuevo valor de las celdas
4	9957	<ul style="list-style-type: none">Masculino (9954 rows)TorerosMasculino (1 rows)BiólogosMasculino (1 rows)ProfesoresMasculino (1 rows)	<input checked="" type="checkbox"/>	Masculino
2	9955	<ul style="list-style-type: none">Masculino (9954 rows)TorerosMasculino (1 rows)	<input checked="" type="checkbox"/>	Masculino
2	2	<ul style="list-style-type: none">TorerosMasculino (1 rows)ProfesoresMasculino (1 rows)	<input checked="" type="checkbox"/>	TorerosMasculino

Valores en la agrupación

Filas en la agrupación

Longitud promedio de los valores

Varianza de los valores

Seleccionar todos Seleccionar ninguno Exportar agrupaciones **Unir seleccionados y reagrupar** Unir seleccionados y cerrar Cerrar

En este caso, para tercera opción nos propone algo que no es lo ideal, por tanto podemos corregirlo en el momento.

Agrupar y editar valores en la columna "género"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. [Más información ...](#)

Método Function Radio Caracteres del bloque 3 clusters Encontrado

Número de valores	Número de filas	Valores en la agrupación	¿Unir?	Nuevo valor de las celdas
4	9957	<ul style="list-style-type: none"> Masculino (9954 rows) TorerosMasculino (1 rows) BiólogosMasculino (1 rows) ProfesoresMasculino (1 rows) 	<input checked="" type="checkbox"/>	<input type="text" value="Masculino"/>
2	9955	<ul style="list-style-type: none"> Masculino (9954 rows) TorerosMasculino (1 rows) 	<input checked="" type="checkbox"/>	<input type="text" value="Masculino"/>
2	2	<ul style="list-style-type: none"> TorerosMasculino (1 rows) ProfesoresMasculino (1 rows) 	<input checked="" type="checkbox"/>	<input type="text" value="Masculino"/>

Valores en la agrupación

Filas en la agrupación

Longitud promedio de los valores

Varianza de los valores

Una vez pasados los algoritmos, aún quedarán errores que se podrán corregir manualmente.

Actualizar
Restablecer todos
Remove todos

género cambiar

11 choices Ordenar por: A-Z conteo Agrupar

- Barcelona 1
- Compositores 1
- España 1
- Español 1
- Femenino 231387
- Género 1
- Hombre 1 [editar](#) [include](#)
- latita 1
- Masculino 9961
- Mijer 1
- Mujes 1
- (blank) 1088013

Facetas por conteo de opciones

	id BNE	otros códigos
1.	XX95384	
2.	XX95608	
3.	XX95777	
4.	XX111504	
5.	XX113822	
6.	XX115812	
7.	XX117519	
8.	XX123992	
9.	XX127572	
10.	XX129572	
11.	XX130843	
12.	XX130880	

OpenRefine Personas_catálogo_completo [Enlace permanente](#)

Facetas / Filtros
Deshacer / Rehacer 6 / 6

Actualizar Restablecer todos Remove todos

género cambiar
11 choices Ordenar por: A-Z conteo Agrupar

- Barcelona 1
- Compositores 1
- España 1
- Español 1
- Femenino 231387
- Género 1
- Hombre 1**
- latita 1
- Masculino 9981
- Mijer 1
- Mujes 1
- (blank) 1088013

Facetas por conteo de opciones

1329370 filas
Mostrar como: **filas** registros Mostrar: 5 10 25 50 filas

Todo	id BNE	otros códigos de identificación
1.	XX95384	
2.	XX95608	
3.	XX95777	
4.	XX111504	
5.	XX113622	
6.	XX115812	
11.	XX130843	

Masculino

Aplicar Cancelar
Aceptar Cancelar

Después de haber limpiado los casos obvios, tendríamos que detenernos en mirar algunos casos particulares. Nos queda la lista de opciones reducida a 8.

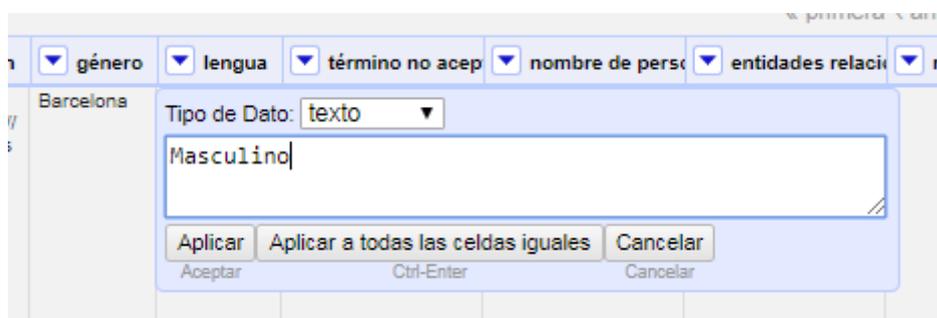
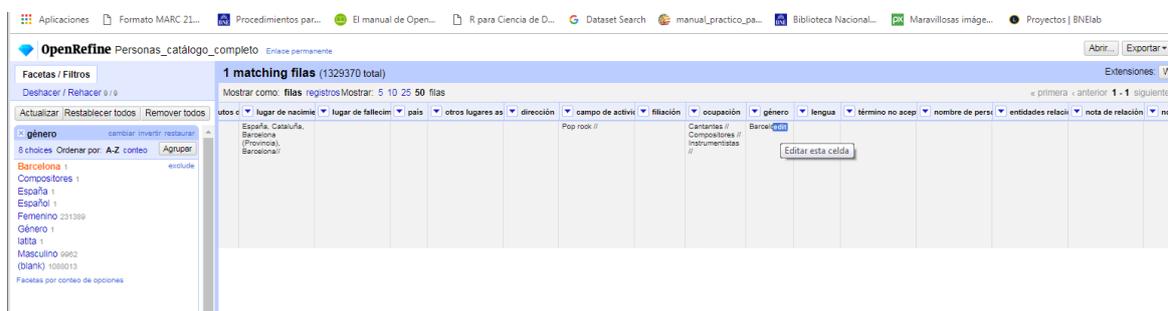
Actualizar Restablecer todos Remove todos

género cambiar
8 choices Ordenar por: A-Z conteo Agrupar

- Barcelona 1
- Compositores 1
- España 1
- Español 1
- Femenino 231389
- Género 1
- latita 1
- Masculino 9982
- (blank) 1088013

Facetas por conteo de opciones

Seleccionando cada opción, se nos mostrará el registro asociado en el panel de la derecha y podremos editarlo.



Y así lo haríamos con todos los casos en duda.

Al final, conseguimos limpiar de una forma sencilla todas las opciones hasta reducirlo a lo correcto.



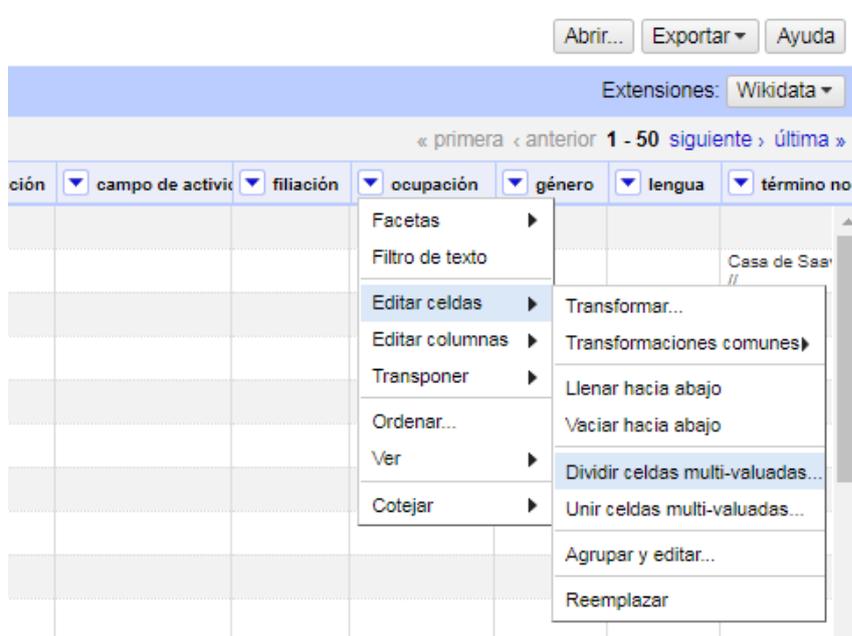
Como hecho remarcable, hay 1.088.013 registros que no tienen el campo “Género” completado. Ahora podríamos exportar el conjunto total, o en dos tandas (masculino/femenino) para poder modificar los valores en nuestro SIGB, Symphony.

5. Las celdas multivaluadas

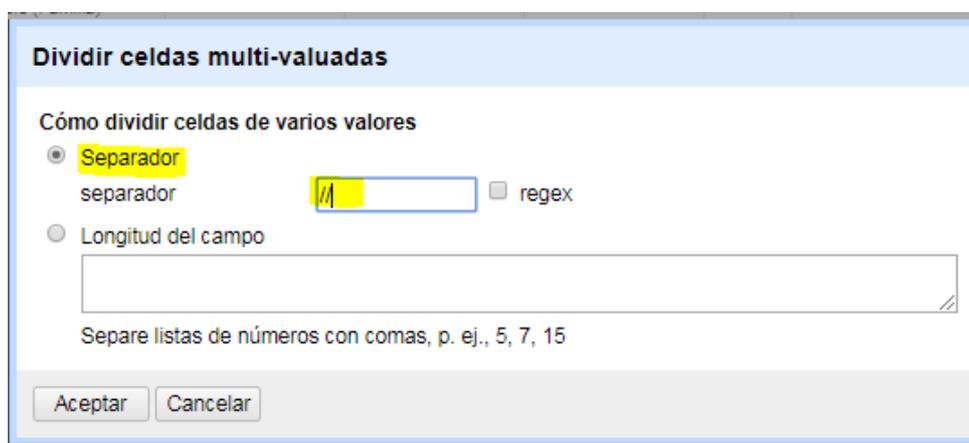
En muchos casos, el resultado de nuestro mapeo nos presenta en una misma celda valores agrupados. Esto ocurre cuando el campo MARC es repetible. Tomemos por ejemplo el campo Ocupación (374 \$a). Veremos que cuando los valores se repiten para una misma autoridad, aparecen separados por “/”.



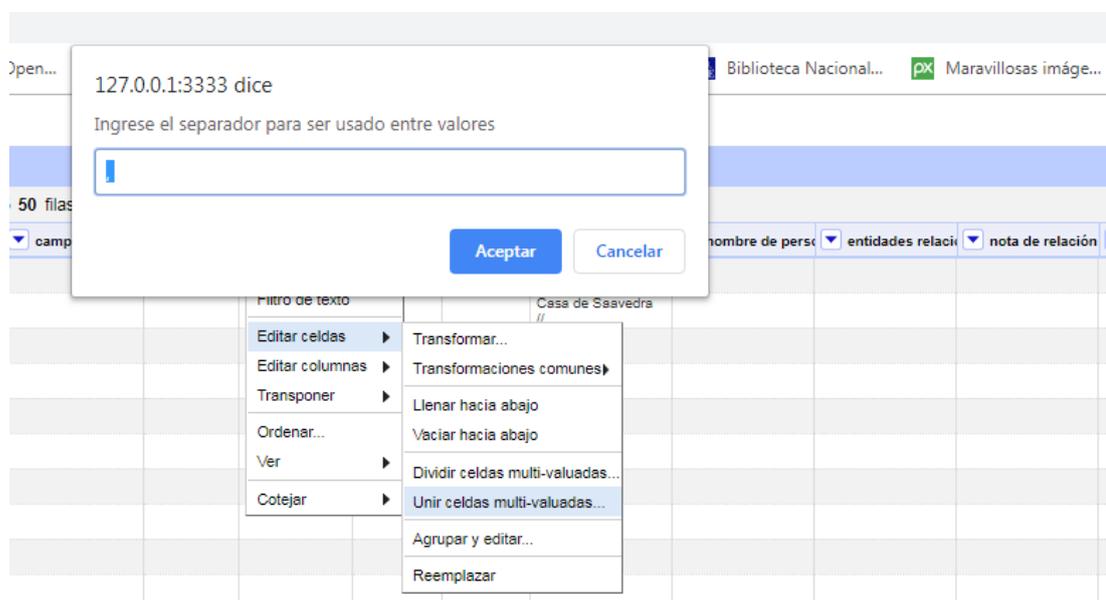
En este caso, lo primero que hay que hacer es separar las celdas multivalor en valores individuales. Para ello, desde la columna que queremos dividir: Editar celdas/Dividir celdas multi-valudadas



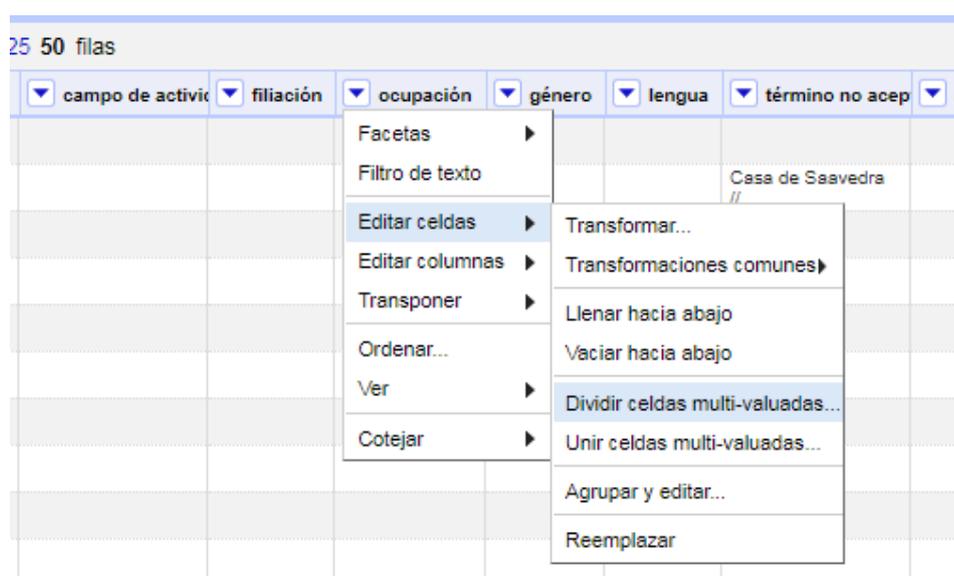
En el cuadro de diálogo debemos establecer qué caracteres queremos que use el sistema para “partir” las celdas. En nuestro caso, será “//”.



Por la estructura de nuestro mapeo, es necesario realizar esta acción 2 veces, por tanto los tenemos que unir y volver a separar. Unimos con “//”



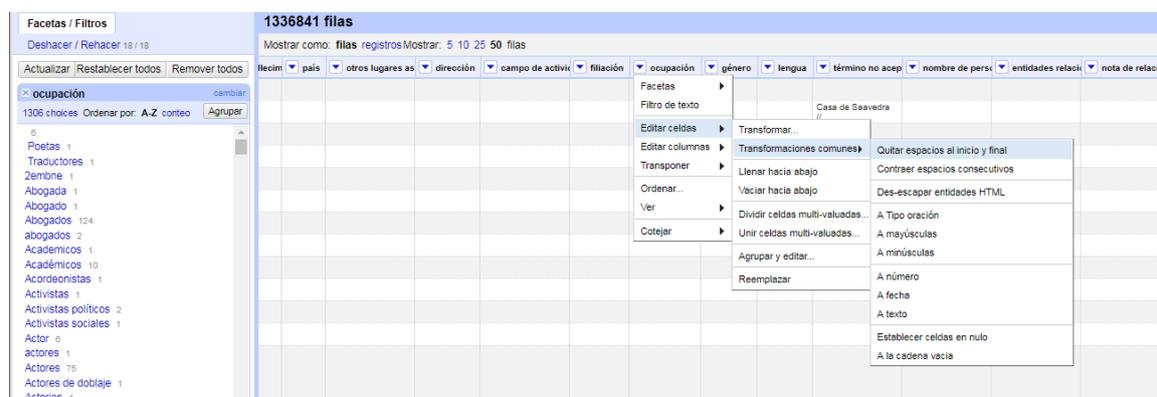
Separamos por “//”



El resultado serían 1.306 opciones:



En este caso, al haber muchas opciones, recomendamos empezar limpiando los espacios en blanco al principio y al final: Editar celdas/Transformaciones comunes/Quitar espacios al inicio y final.



Solo con esta operación, hemos reducido el número de opciones a 987.



Ahora sería el momento de repetir los diferentes métodos de agrupamiento que hemos visto anteriormente.

Al terminar de limpiar, como en este caso había celdas multivalor, debemos recordar que estamos trabajando con filas y hay que reagruparlo a nivel de registro. Eso lo haríamos utilizando de nuevo la opción unir celdas multivalor.

6. Enriquecimiento con fuentes externas

Imaginemos que, una vez detectados y corregidos los errores del catálogo, nos interesa enriquecerlo, al introducir datos que ya existen en otros catálogos y bases de datos de prestigio. OpenRefine nos ayuda también con esta tarea, ya que ofrece un servicio de “reconciliación”, que nos permite cotejar nuestros datos contra otras fuentes.

OpenRefine, por defecto, ofrece esta posibilidad contra una serie de servicios de reconciliación preestablecidos, disponibles en esta url: <http://refine.codefork.com/> . Los más importantes serían Wikidata, LC y VIAF (por países).

Los pasos a seguir serían:

- Reconciliar los datos con un servicio preestablecido, o
- Reconciliar los datos con una fuente de datos propia
- Crear columnas basadas en datos vinculados

6.1. Ejemplo 1: Autores en dominio público

6.1.1. Reconciliar datos con Wikidata

Utilizamos para el ejemplo el conjunto de autores que entraron en dominio público en 2019, fallecidos en 1938 (<https://bnelab.bne.es/dato/autores-espanoles-en-dominio-publico/>).

En OpenRefine, llamamos “reconciliar” o “cotejar” a la tarea de buscar nuestros datos en una fuente externa. Debemos elegir qué campo queremos utilizar para ir contra Wikidata. En este caso, buscaremos por “Nombre de persona” (MARC: 100 \$a \$b \$c (\$d)(\$q)). En la columna correspondiente: Cotejar/Iniciar.

The screenshot shows the OpenRefine interface with a table of 175 rows. The columns include 'id BNE', 'otros códigos de identificación', 'fecha de naciem', 'fecha de fallecim', 'nombre de pers', 'otros atributos c', 'lugar de naciem', 'lugar de fallecim', 'pais', 'otros lugares as', 'dirección', and 'campo de activi'. A context menu is open over the 'Cotejar' column, showing options like 'Iniciar', 'Facetas', 'Acciones', and 'Copiar información de cotejo...'. A tooltip points to the 'Iniciar' option with the text 'Coincidir texto en esta columna con valores de Freebase'.

En el cuadro de diálogo, elegimos los parámetros. En este ejemplo vamos a lanzar la búsqueda sin establecer una clase determinada.

Cotejar columna "nombre de persona"

» Ir a API del servicio

Cotejar cada celda con los valores de una de estas clases:

- ser humano Q5
- artículo académico Q13442814

Usar también detalles relevantes de otras columnas:

columna	¿Incluir? Como propiedad
id BNE	<input type="checkbox"/>
otros códigos de identificación	<input type="checkbox"/>
fecha de nacimiento	<input type="checkbox"/>
fecha de fallecimiento	<input type="checkbox"/>
otros atributos de persona	<input type="checkbox"/>
lugar de nacimiento	<input type="checkbox"/>
lugar de fallecimiento	<input type="checkbox"/>
país	<input type="checkbox"/>
otros lugares asociados	<input type="checkbox"/>
dirección	<input type="checkbox"/>
campo de actividad	<input type="checkbox"/>

Cotejar contra la clase:

Reconciliar contra ninguna clase en particular

Cotejar automáticamente candidatos con alta confianza

Máximo número de candidatos a devolver

Agregar servicio estándar... Cotejar Cancelar

Hemos conseguido vincular 63 registros, de los 175 que teníamos en origen. De esos 63 autores podemos traer datos que existen en Wikidata para completar nuestro conjunto, y posteriormente descargarlo para enriquecer nuestro catálogo.

OpenRefine Dominio público 2019 [Enlace permanente](#) Abrir... Exportar Ayuda

Facetas / Filtros Extensiones: Wikidata

Deshacer / Rehacer 2 / 2

Actualizar | Restablecer todos | Remover todos

nombre de persona: judgment cambiar | invertir | restaurar

2 choices Ordenar por: A-Z conteo

matched 63

none 112

Facetas por conteo de opciones

nombre de persona: best candidate's score cambiar | restaurar

Numérico Non-numérico Blank Error

63 matching filas (175 total)

Mostrar como: filas registros Mostrando: 5 10 25 50 filas

id	id BNE	otros códigos de identificación	fecha de nacimiento	fecha de fallecimiento	nombre de persona	otros atributos c	lugar de nacimiento	lugar de fallecimiento	país	otros lugares as	dirección	campo de activi	fill
2	XX1752018	viaf: http://viaf.org/viaf/87533172 **	1871	1938	Diego Alonso Nival Escoger ruota concordata		La Bañeza, León, España/	Caracas, Venezuela/					
5	XX855578	viaf: http://viaf.org/viaf/42263854 ** (n); http://www.isni.org/isni/0000000059491045 ** wikidata: https://www.wikidata.org/wiki/Q6269177 **	1854	1938	José María Alvira Escoger ruota concordata								
6	XX1182121	viaf: http://viaf.org/viaf/30423425 ** (n); http://www.isni.org/isni/0000000066354166 ** wikidata: https://www.wikidata.org/wiki/Q6141803 **	1880	1938	Teodoro de Anasagasti Escoger ruota concordata								

6.1.2. Reconciliar datos con fuentes propias: reconciliar con un subconjunto de Wikidata obtenido mediante consulta SPARQL

Además de reconciliar tus datos con las fuentes explicadas anteriormente, existe la posibilidad de crear tus propias fuentes.

Estas fuentes deben ser ficheros en formato CSV. Vamos a utilizar una herramienta libre que se llama [Reconcile-CSV](#), que nos permite convertir nuestro CSV en una fuente de reconciliación:

Por ejemplo, para el conjunto de autores españoles fallecidos en el 1938, en lugar de buscar en todo wikidata, podríamos buscar únicamente en el subconjunto de personas españolas fallecidas en el año 1938.

Hemos realizado esta consulta con SPARQL

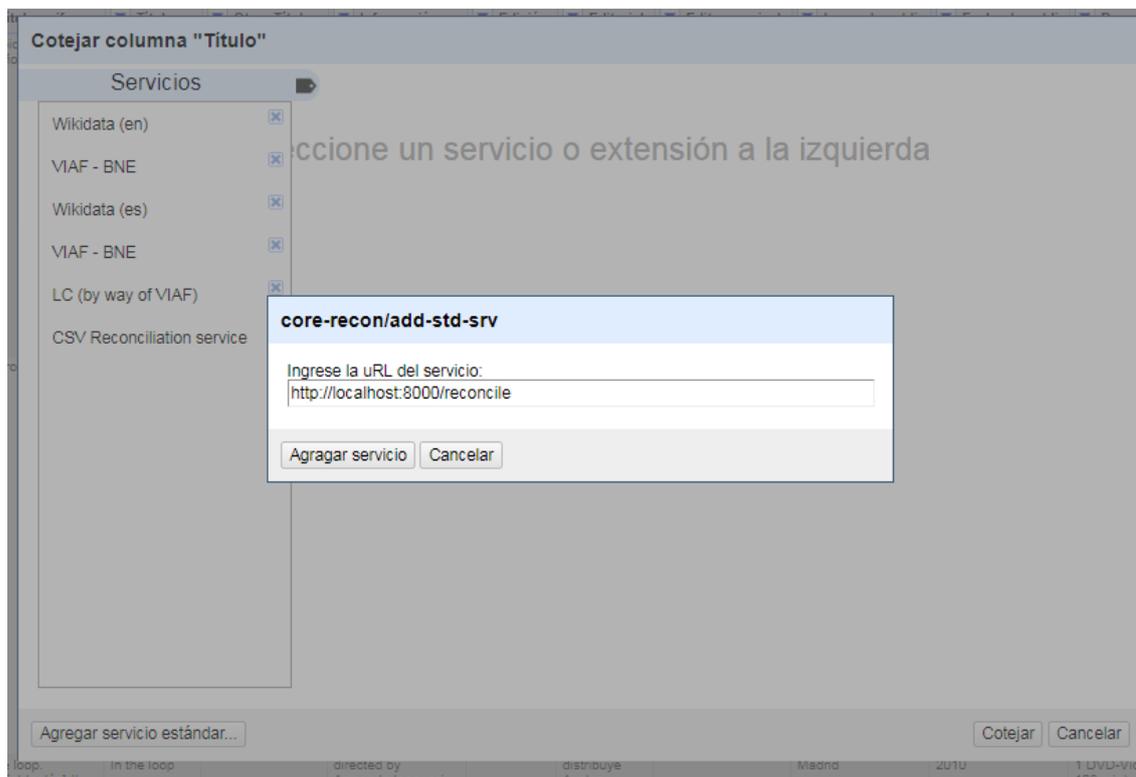
The screenshot shows the Wikidata Query Service interface. On the left, there is a filter panel with dropdown menus for 'instancia de' (set to 'ser humano') and 'pais de nacionalidad' (set to 'España'). Below this, there are several 'Mostrar' (Show) buttons for various properties: 'fecha de nacimiento', 'fecha de fallecimiento', 'lugar de nacimiento', 'lugar de fallecimiento', 'ocupación', 'sexo o género', and 'identificador BNE'. On the right, a SPARQL query is displayed in a text area, numbered 1 to 16. The query is a SELECT statement with various OPTIONAL clauses and a FILTER clause for the year 1938.

```
SELECT ?ser_humano ?fecha_de_fallecimiento ?identificador_BNE ?fecha_de_nacimiento ?lugar_de_nacimiento
?lugar_de_nacimientoLabel ?ocupaci_n ?ocupaci_nLabel ?sexo_o_g_nero ?sexo_o_g_neroLabel
?lugar_de_fallecimiento ?lugar_de_fallecimientoLabel
WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
  ?ser_humano wdt:P31 wd:Q5.
  OPTIONAL { }
  ?ser_humano wdt:P27 wd:Q29.
  FILTER((YEAR(?fecha_de_fallecimiento)) = 1938 )
  OPTIONAL { }
  OPTIONAL { ?ser_humano wdt:P569 ?fecha_de_nacimiento. }
  OPTIONAL { ?ser_humano wdt:P570 ?fecha_de_fallecimiento. }
  OPTIONAL { ?ser_humano wdt:P19 ?lugar_de_nacimiento. }
  OPTIONAL { ?ser_humano wdt:P20 ?lugar_de_fallecimiento. }
  OPTIONAL { ?ser_humano wdt:P106 ?ocupaci_n. }
  OPTIONAL { ?ser_humano wdt:P21 ?sexo_o_g_nero. }
  OPTIONAL { ?ser_humano wdt:P950 ?identificador_BNE. }
}
```

Obtenemos un conjunto de 442 registros. Podemos descargarlo en CSV y trabajarlo como fuente externa para reconciliar desde OpenRefine con nuestro conjunto origen.

Ahora reconciamos nuestro conjunto de 175 autores en dominio público con el conjunto de 442 de Wikidata.

```
C:\Windows\system32\cmd.exe - java -jar .\reconcile-csv-0.1.2.jar .\PersonasEspañolas_Fallecidas1938.csv ser_humanoLabel ser_humano
Microsoft Windows [Versión 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Reservados todos los derechos.
C:\Users\puebh010>cd ..
C:\Users>cd ..
C:\>cd reconciliarCSU
C:\reconciliarCSU>java -jar .\reconcile-csv-0.1.2.jar .\Cine-ICAA-2000-Actualidad_Españolas.csv Titulo Codigo
Starting CSU Reconciliation service
Point refine to http://localhost:8000 as reconciliation service
2019-04-25 09:44:50.454:INFO:oejs.Server:jetty-7.x.y-SNAPSHOT
2019-04-25 09:44:50.541:INFO:oejs.AbstractConnector:Started SelectChannelConnector@0.0.0:8000
C:\reconciliarCSU>java -jar .\reconcile-csv-0.1.2.jar .\PersonasEspañolas_Fallecidas1938.csv ser_humanoLabel ser_humano
Starting CSU Reconciliation service
Point refine to http://localhost:8000 as reconciliation service
2019-04-25 10:26:58.278:INFO:oejs.Server:jetty-7.x.y-SNAPSHOT
2019-04-25 10:26:58.396:INFO:oejs.AbstractConnector:Started SelectChannelConnector@0.0.0:8000
```



Resultado:

- Al intentar vincular por nombre de persona, al 100% no existe ninguna coincidencia, pero por encima del 50% obtenemos 142. Esto se debe a que nuestra forma de escribir el nombre, va en orden inverso y además incluye las fechas de nacimiento y muerte, por lo que es imposible la coincidencia total.
- Al intentar vincular por ID BNE, obtenemos 78 coincidencias al 100%. Esto puede deberse a que muchos autores de Wikidata no tienen el metadato ID BNE completo.

OpenRefine dominiopublico_2019 Permalink

Open... Export Hel

Facet / Filter Undo / Redo 4 / 4

175 rows Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

All	id BNE	otros códigos de id	fecha de naci	fecha de falleci	nombre de persona	otros atributos	lugar de naci
1.	XX1192852 <input checked="" type="checkbox"/> XX1195928 (0.75) <input checked="" type="checkbox"/> XX1195928 (0.75) <input checked="" type="checkbox"/> XX1115274 (0.533) <input checked="" type="checkbox"/> XX1389280 (0.5) <input checked="" type="checkbox"/> XX1389280 (0.5) <input checked="" type="checkbox"/> Create new item Search for match		1882	1938	Allende-Salazar, Juan (1882-1938) <input checked="" type="checkbox"/> Create new item		
2.	XX1752015 (0.625) <input checked="" type="checkbox"/> XX1720215 (0.625) <input checked="" type="checkbox"/> XX1725203 (0.625) <input checked="" type="checkbox"/> XX1725203 (0.625) <input checked="" type="checkbox"/> XX1725203 (0.625) <input checked="" type="checkbox"/> XX1725203 (0.625) <input checked="" type="checkbox"/> Create new item Search for match	viaf: http://viaf.org/viaf/87533172 **	1871	1938	Alonso Nistal, Diego 1871-1938 Choose new match		La Bañeza, León, España//
3.	XX852706 <input checked="" type="checkbox"/> XX852656 (0.571) <input checked="" type="checkbox"/> XX855578 (0.462) <input checked="" type="checkbox"/> XX855578 (0.462) <input checked="" type="checkbox"/> XX855578 (0.462) <input checked="" type="checkbox"/> XX855578 (0.462) <input checked="" type="checkbox"/> Create new item Search for match	viaf: http://viaf.org/viaf/87687473 ** isni: http://www.isni.org/isni/00000000968787918 **	1882	1938	Alonso Valdrés, Emilio 1882-1938 Choose new match		
4.	XX856136 Choose new match	viaf: http://viaf.org/viaf/88977969 ** isni: http://www.isni.org/isni/0000000129965975 ** wikidata: https://www.wikidata.org/wiki/Q11948260 **	1871	1938	Álvarez Quintero, Serafín, 1871-1938 Choose new match		España, Andalucía, Sevilla (Provincia), Ultramar//

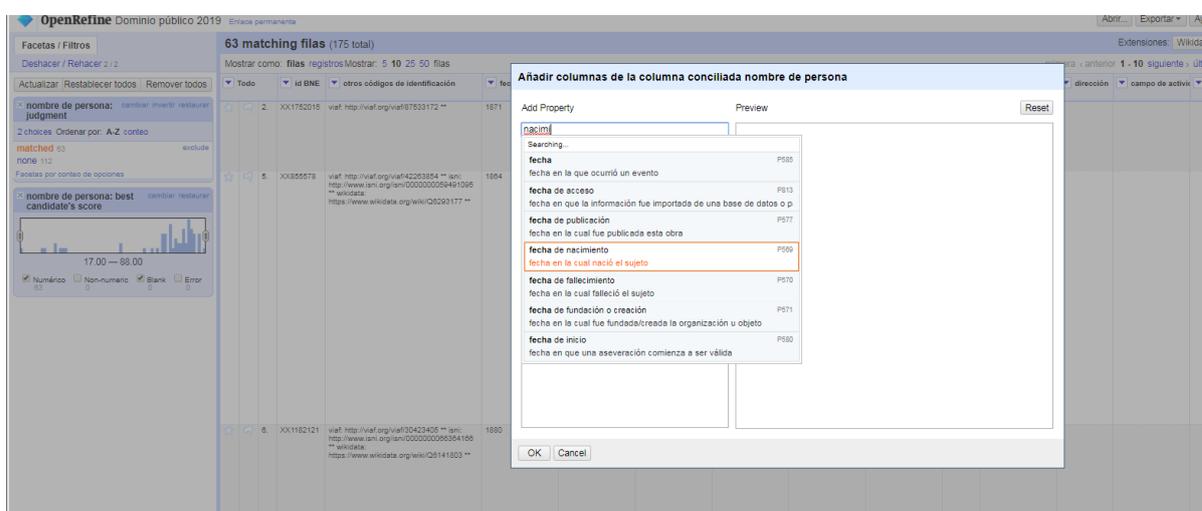
Vistos estos ejemplos, podemos decir, que si se realiza una limpieza previa de la fuente que queremos usar como repositorio de reconciliación, podemos mejorar el porcentaje de coincidencias automáticas.

6.1.3. Crear columnas basadas en datos vinculados

Una vez que hemos logrado “casar” nuestros datos con otras fuentes, podremos ahora importar los campos que nos interese incorporar a nuestro catálogo.

Del conjunto de registros, pensamos que hay una serie de datos de interés para enriquecer las autoridades:

- Fecha de nacimiento
- Fecha de fallecimiento
- Lugar de nacimiento
- Lugar de muerte
- Ocupación
- Género
- ID BNE



Ejemplo de un autor concreto del que hemos recuperado información de Wikidata:

	Fuente: BNE	Fuente: Wikidata
Nombre de persona	José María Alvira	
id BNE	XX855578	XX855578
Fecha de nacimiento	1864	1864-01-01T00:00:00Z
Fecha de fallecimiento	1938	1938-01-01T00:00:00Z
Lugar de nacimiento		Zaragoza
Lugar de fallecimiento		Madrid
Ocupación		compositor // pianista // director de orquesta // profesor
Género/sexo		Masculine

6.2. Ejemplo 2: Videograbaciones publicadas en 2010

6.2.1. Reconciliar datos con Wikidata

Otro ejemplo podría ser reconciliar [registros bibliográficos de videograbaciones](#), a través del campo "Título" (MARC: 245 \$a, \$b, \$n, \$p). Para tener un conjunto manejable, hemos filtrado videograbaciones con fecha de publicación 2010 y que tengan algún intérprete (MARC: 511 \$a). Este conjunto queda reducido a 3.808 registros.

Después hemos llamado a Wikidata. En este caso, en lugar de buscar contra todo Wikidata, hemos preferido elegir directamente la clase Película (Q11424). Tras el proceso, se han vinculado 1.674 películas, de las que podemos traer datos de interés para el registro.

OpenRefine videos [Enlace permanente](#)

1674 matching files (126789 total)

Mostrar como: **filas** registros Mostrar: 5 10 25 50 filas

Todo	File	idBNE	Autor Personas	Autor Entidades	Título uniforme	Título	Otros Títulos	Información aso	Edición	Editorial	Editor musical	Lugar de publico	Fecha de publico	Descr
	4442	VIDEO_1-CPI1252.csv	biv0000004929	Chabrol, Claude, 1930-2010 // Rabier, Jean, 1927-2019 // Jensen, Pierre // Génovés, André, 1941- // Gégauff, Paul, 1922-1993 // Tompkins, Jean-Louis, 1930- // Sessard, Jacqueline, 1940- // Audran, Stéphane, 1932-2018 //		Les broches. Español-Francés //	Las ciencias biológicas conciencia		dirigida por Claude Chabrol / guión, Claude Chabrol y Paul Gégauff / fotografía, Jean Rabier / música, Pierre Jensen / (productor, André Génovés)	Regie Films	Regie Films	[Barcelona]	2010	1 DVD-Vid (201 min) ; son. n.
	83018	VIDEO_2-CPI1252.csv	s4592838	Berlioz, Hector, 1803-1869 // Melano, Fabrizio // Weiler, Peter, 1930- // Gell, Peter // Domingo, Plácido, 1941- on. // Monk, Allan, 1942- // Chase, John // Levine, James, 1943- // Silverman, David, 1933-1990 //	Metropolitan Opera (Nueva York) Ballet // Metropolitan Opera (Nueva York) Opera // Metropolitan Opera (Nueva York) Orquesta //	[Les troyens] //	Catálogo de las troyanas Escoger nueva conciencia	Hector Berlioz		Altaya		[Barcelona]	2010	2 DVD-Vid (201 min) ; 2 folletos
	84016	VIDEO_2-CPI1252.csv	s4596257	Iannucci, Armando, 1963- // Loader, ... //	The Elysian Quartet //	[In the loop. Español-Ingles] //	[In the loop. Escoger nueva conciencia		directed by Armando Iannucci ;	distribuye Avalon		[Madrid]	2010	1 DVD-Vid (106 min) ;

6.2.2. Reconciliar datos con fuentes propias: Reconciliar con catálogo de películas calificadas del ICAA

El ICAA dispone de una [base de datos](#) de películas calificadas desde los años 40 a la actualidad. Este repositorio está muy completo con información muy detallada en campos que nosotros no tenemos, por ejemplo: nacionalidad o año de producción.

Hemos podido acceder a un [servicio web](#), que nos ofrece la posibilidad de descargar algunos de esos datos de las películas calificadas en un año concreto. Con esto podemos generar un CSV que utilizaremos como repositorio de reconciliación.

Nuestro conjunto origen son registros bibliográficos de videograbaciones (BNE) con fecha de publicación 2010 e intérpretes no nulos (3.308 registros)

Como repositorio de reconciliación vamos a tomar el fichero csv de películas del ICAA desde el año 2000 a la actualidad (7.790 películas).

Ejecutamos el reconcile-csv para nuestro fichero fuente, con las películas del ICAA.

```

C:\Windows\system32\cmd.exe - java -jar .\reconcile-csv-0.1.2.jar .\Cine-ICAA-2000-Actualidad_Espanolas.csv Titulo Codigo
Microsoft Windows [Versión 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Reservados todos los derechos.

C:\Users\peweb010>cd ..
C:\Users>cd ..
C:\>cd reconciliarCSU
C:\reconciliarCSU>java -jar .\reconcile-csv-0.1.2.jar .\Cine-ICAA-2000-Actualidad_Espanolas.csv Titulo Codigo
Starting CSU Reconciliation service
Point refine to http://localhost:8000 as reconciliation service
2019-04-25 09:44:50.454:INFO:oe.js.Server:jetty-7.x.y-SNAPSHOT
2019-04-25 09:44:50.541:INFO:oe.js.AbstractConnector:Started SelectChannelConnector@0.0.0:8000

```

Intentamos reconciliar desde OpenRefine y como resultado de intentar vincular por título, obtenemos 70 coincidencias al 70%. Esto significa que son 70 ediciones de película que podemos confirmar que son de nacionalidad Española.

ID	File	ICAAE	Autor Personal	Autor Entidades	Título uniforme	Título	ICAA_codigo	ICAA_Director	ICAA_Fecha_Pr	ICAA_Genero	Otros Titulos	Información aso	Edición	Ed
84312	VIDEO_2_CPI1232.csv	#4712411	Morante, Alexis 1975- // Bolo, María // Cid Troya, Jaime // Sosa Segura, Daniel. 1974- // Tomes, Miguel. 1975- // Santos, Rauli // Cevallos, Pater // Rosso, María Alberta // O'Donoherty, Alex. 1973- //	Los Delinquentes (Grupo musical) //	VOLTERETA	El grupo nueva barcelonesa	43600	MARIO ALEXIS MORANTE PORTILLO	2010	Comedia		dirigido por Alexis Morante, producido por María Bolo, Jaime Cid Troya, dirección de fotografía: Daniel Sosa - música, Miguel Tomes con la colaboración especial de Los Delinquentes - guión: Alexis Morante, Rauli Santos		7000 F
84313	VIDEO_2_CPI1232.csv	#4712500	Morante, Alexis 1975- // Bolo, María // Cid Troya, Jaime // Sosa Segura, Daniel. 1974- // Tomes, Miguel. 1975- // Santos, Rauli // Cevallos, Pater // Rosso, María Alberta // O'Donoherty, Alex. 1973- //	Los Delinquentes (Grupo musical) //	VOLTERETA	El grupo nueva barcelonesa	43609	MARIO ALEXIS MORANTE PORTILLO	2010	Comedia		directed by Alexis Morante, produced by María Bolo, Jaime Cid Troya - direction of photography: Daniel Sosa - music Miguel Tomes with the special collaboration of Los Delinquentes written by Alexis Morante, Rauli Santos		7000 F

Sobre estas películas coincidentes, podemos traer sus datos asociados para completar nuestro catálogo.

6.2.3. Crear columnas basadas en datos vinculados

Si por ejemplo, quisiéramos enriquecer nuestro catálogo con los campos año de producción, género y director del fichero de datos del ICAA, necesitaríamos tener un campo en común, que debe ser el identificador único del ICAA, para así posteriormente, poder recuperar su información asociada.

Desde OpenRefine, esto se haría trabajando con 2 proyectos simultáneamente (proyecto A: datos originales y proyecto B: el archivo usado como repositorio de reconciliación)

Para hacer esto recuperamos en el proyecto A el ID único del fichero CSV que utilizamos como fuente de reconciliación (en este caso, el código ICAA).

Agregar columna basada en la columna Título

Nuevo nombre de la columna:

core-views/addasdasd cambiar a en blanco guardar error copiar valor de la columna original

Expresión: Lenguaje: No hay error de sintaxis.

Vista previa Historial Con estrella Ayuda

row	value	cell.recon.match.id
84312.	Voltereta	43609
84313.	Voltereta	43609
84530.	Bullying	135208
84544.	Adam	null
84817.	La noche La notte	null
84923.	Tras el cristal	null
84973.	Sube y baja	null

Gracias a este código relacionamos los 2 proyectos y podemos ir trayendo columnas seleccionadas del proyecto B al proyecto A, que es el que queremos enriquecer.

Ejemplo: `cell.cross("ICAA_SPAIN", "Codigo").cells["Genero"].value[0]`

Agregar columna basada en la columna ICAA_codigo

Nuevo nombre de la columna

core-views/addasdasd cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje No hay error de sintaxis.

Vista previa [Historial](#) [Con estrella](#) [Ayuda](#)

row	value	cell.cross("ICAA_SPAIN","Codig ...
84312.	43609	Comedia
84313.	43609	Comedia
84530.	135208	Drama
84544.	null	Error: cross expects a string or cell, a project name to join with, and a column name in that project
84817.	null	Error: cross expects a string or cell, a project name to join with, and a column name in that project
84923.	null	Error: cross expects a string or cell, a project name to join with, and a column name in that project

El resultado lo podemos ver en OpenRefine:

70 matching rows (128539 total) Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last

Autor Personas	Autor Entidades	Título uniforme	Título	Codigo_ICAA	ICAA_Director	ICAA_año_prod	Genero ICAA	Otros Títulos	Información adicional
ante, Alexis , 3- // Boix, María Troya, Jaime // Segura, Daniel 74- // Torres, iel, 1976- // tos, Raúl // allos, Peter // so, María nsa // gherty, Álex , 3- //	Los Delinquentes (Grupo musical) //		VOLTERETA <small>Choose new match</small>	43609	MARIO ALEXIS MORANTE PORTILLO	2010	Comedia		dirigido por Alexis Morante ; producido por María Boix, Jaime Cid Troya ; dirección de fotografía, Daniel Sosa ; música, Miguel Torres con la colaboración especial de Los Delinquentes ; guión, Alexis Morante, Raúl Santos
ante, Alexis , 3- // Boix, María Troya, Jaime // Segura, Daniel 74- // Torres, iel, 1976- // tos, Raúl // allos, Peter // so, María nsa // gherty, Álex , 3- //	Los Delinquentes (Grupo musical) //		VOLTERETA <small>Choose new match</small>	43609	MARIO ALEXIS MORANTE PORTILLO	2010	Comedia		directed by Alexis Morante ; produced by María Boix, Jaime Cid Troya ; direction of photography, Daniel Sosa ; music, Miguel Torres with the special collaboration of Los Delinquentes ; written by Alexis Morante, Raúl Santos

También podemos exportarlo en diferentes formatos como por ejemplo CSV. Aquí veremos un ejemplo de campos enriquecidos: Directo, año de producción o género.

	B	C	D	E	F	G	H	I	J	K	L
1	idBNE	Autor Personas	Autor Entidades	Título uniforme	Título	Codigo_ICAA	ICAA_Director	ICAA año producción	Genero ICAA	Otros Títulos	Información asociada al título
2	a4712411	Morante, Alexis	Los Delinquentes (Grupo musical) //	VOLTERETA	VOLTERETA	43609	MARIO ALEXIS MOR	2010	Comedia		dirigido por Alexis Morante
3	a4712520	Morante, Alexis	Los Delinquentes (Grupo musical) //	VOLTERETA	VOLTERETA	43609	MARIO ALEXIS MOR	2010	Comedia		directed by Alexis Morante
4	a4720394	San Mateo, José María // García Ro	[Bullying. Español-Cata	BULLYING	BULLYING	135208	JOSETXO SAN MATE	2009	Drama		dirigida por Josecho San M
5	a4756847	Cameron, James	Academy Awards [Avatar (Película cinema	AVATAR	AVATAR	107804	LLUIS QUILEZ SALA	2005	Drama		guión y dirigida por James
6	a4770447	Huerga, Manuel, 1957- // Escribano, Francesc, 1958- // Lla	SALVADOR PU			111505	MANUEL HUERGA G	2006	Drama		dirigida por Manuel Huerg
7	a4770486	Balagueró, Jaume	Premis Gaudi 2º, 2010 Barcelona // Festiv	REC 2	REC 2	155908	JAUME BALAGUERO	2009	Terror	Rec dos //(Rec) 2	dirigida por Jaume Balaguer
8	a4780581	Ibáñez, Gabe, 19	Festival Internacional de Cinema de Catalu	HIERRO	HIERRO	97208	GABRIEL IBÁÑEZ RO	2009	Thriller		dirigida por Gabe Ibáñez ; i
9	a4790102	Cameron, James	Academy Awards [Avatar (Película cinema	AVATAR	AVATAR	107804	LLUIS QUILEZ SALA	2005	Drama		guión y dirigida por James
10	a4792600	Reeve, Geoffrey, 1932-2010 // Tam	[Souvenir (Geoffrey Ree	SOUVENIR	SOUVENIR	153111	GERARDO CARRERAS	2013	Documental		directed by Geoffrey Reeve
11	a4793398	Agresti, Alejandro, 1961- // Mairal, Pedro, 1970- // Bossi,	UNA NOCHE C			117699	ALEJANDRO AGRESTI	1999	Drama		un film de Alejandro Agrest
12	a4797908	Monzón, Daniel,	Premios Goya 24º, 2010 Madrid // Meda	CELDA 211	CELDA 211	127008	DANIEL MONZON JE	2009	Thriller		dirigida por Daniel Monzón
13	a4800176	Serra, Albert, 19	Mostra Internazionale d'Arte Cinematogra	HONOR DE CA	HONOR DE CA	9906	ALBERT SERRA JUAN	2006	Drama	Albert Serra //	una película escrita, produ
14	a4803338	Pons, Ventura, 1945- // Cases, Carles, 1958- // Minguell, J	A LA DERIVA			85209	BONAVENTURA PON	2009	Drama		dirección y producción, Ve
15	a4813211	Recha, Marc, 1970- // Lamari, Nadine // Vidal, Jérôme // L	PETIT INDI			77307	MARC RECHA BATAL	2009	Drama		dirección, Marc Recha ; gui
16	a4826400	Aliaga, Adán // H	Semana Internacional de Cine de Valladolid	ESTIGMAS	ESTIGMAS	72807	Adán Aliaga Pasto	2009	Drama		guión y dirección, Adán Ali
17	a4840325	Moccia, Federico, 1963- // Rusic, F	[Scusa ma ti chiamo am	PERDONA SI T	PERDONA SI T	105413	JOAQUIN LLAMAS	2014	Comedia		dirigida por Federico Mocc
18	a4843161	Las Heras, Pepe d	Mojinos Escozios (Grupo musical) //	MUCHA SANG	MUCHA SANG	97199	PEPE DE LAS HERAS	2002	Comedia de terror	Paul Naschy //	dirigida por Pepe de las He
19	a4843227	Yuzna, Brian, 1951- // Fernández, Julio, 1947- // Vázquez-F	ROTWEILER			126003	BRIAN YUZNA	2005	Thriller	Paul Naschy //	dirigida por Brian Yuzna ; g
20	a4847010	Torras, Carles, 1	Premis Gaudi 2º, [Trash (Carles Torras)	TRASH	TRASH	136007	CARLES TORRAS PERE	2009	Drama		dirigida por Carles Torras
21	a4847385	Costa Perdomo, Renate // Andreu, Marta // Benito, Susana	CUCHILLO DE			75608	RENATE COSTA PER	2010	Documental	108 Cuchillo de palo	written and directed by Ren
22	a4847982	Folk, Abel, 1959-	Dept. (Grupo musical) //	XTREMS	XTREMS	187309	ABEL FOLK GILSANZ	2009	Thriller		dirección, Abel Folk, Joan Ri

Como hemos podido ver, la herramienta es muy útil para mantener limpios los catálogos así como para enriquecerlos con datos de otras fuentes externas.