# Práctica OpenRefine

Big Data, Máster universitario de Tecnologías del sector financiero

uc3m | Universidad Carlos III de Madrid

Andoni Alcelay Izarzugaza
18-10-2018

# Objetivo

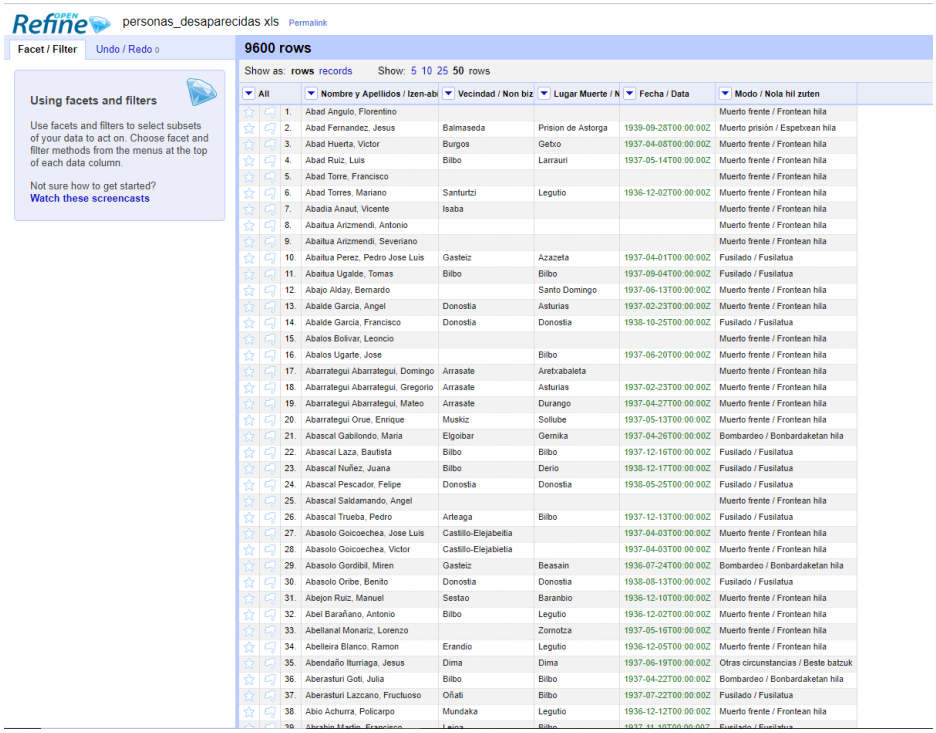
El objetivo de esta práctica es utilizar el software OpenRefine para la limpieza de datos y enriquecer esta información para Wikidata.

# Fichero de datos

Para esta práctica se utilizará el fichero de datos de las personas desaparecidas o asesinadas durante la guerra civil. El fichero está dividido en columnas que tienen información sobre su nombre y apellidos, dónde vive, el lugar de su muerte, la fecha y el modo en el que lo mataron, y está compuesto por 9600 líneas.

# Limpieza de datos

Con la información adquirida inicialmente, se procede a la limpieza de sus datos, nada más importar el fichero de datos la información que aparece en él es la siguiente:




| All | Nombre y Apellidos / Izen-abi | Vecindad / Non biz | Lugar Muerte / N | Fecha / Data | Modo / Nola hil zuten |
|---|---|---|---|---|---|
| 1. | Abad Angulo, Florentino | | | | Muerto frente / Frontean hila |
| 2. | Abad Fernandez, Jesus | Balmaseda | Prision de Astorga | 1939-09-28T00:00:00Z | Muerto prisión / Espetxean hila |
| 3. | Abad Huerta, Victor | Burgos | Getxo | 1937-04-08T00:00:00Z | Muerto frente / Frontean hila |
| 4. | Abad Ruiz, Luis | Bilbo | Larrauri | 1937-05-14T00:00:00Z | Muerto frente / Frontean hila |
| 5. | Abad Torre, Francisco | | | | Muerto frente / Frontean hila |
| 6. | Abad Torres, Mariano | Santurtzi | Legutio | 1936-12-02T00:00:00Z | Muerto frente / Frontean hila |
| 7. | Abadia Anaut, Vicente | Isaba | | | Muerto frente / Frontean hila |
| 8. | Abaitua Arizmendi, Antonio | | | | Muerto frente / Frontean hila |
| 9. | Abaitua Arizmendi, Severiano | | | | Muerto frente / Frontean hila |
| 10. | Abaitua Perez, Pedro Jose Luis | Gasteiz | Azazeta | 1937-04-01T00:00:00Z | Fusilado / Fusilatua |
| 11. | Abaitua Ugalde, Tomas | Bilbo | Bilbo | 1937-09-04T00:00:00Z | Fusilado / Fusilatua |
| 12. | Abajo Alday, Bernardo | | Santo Domingo | 1937-06-13T00:00:00Z | Muerto frente / Frontean hila |
| 13. | Abalde Garcia, Angel | Donostia | Asturias | 1937-02-23T00:00:00Z | Muerto frente / Frontean hila |
| 14. | Abalde Garcia, Francisco | Donostia | Donostia | 1938-10-25T00:00:00Z | Fusilado / Fusilatua |
| 15. | Abalos Bolivar, Leoncio | | | | Muerto frente / Frontean hila |
| 16. | Abalos Ugarte, Jose | | Bilbo | 1937-06-20T00:00:00Z | Muerto frente / Frontean hila |
| 17. | Abarrategui Abarrategui, Domingo | Arrasate | Aretxabaleta | | Muerto frente / Frontean hila |
| 18. | Abarrategui Abarrategui, Gregorio | Arrasate | Asturias | 1937-02-23T00:00:00Z | Muerto frente / Frontean hila |
| 19. | Abarrategui Abarrategui, Mateo | Arrasate | Durango | 1937-04-27T00:00:00Z | Muerto frente / Frontean hila |
| 20. | Abarrategui Orue, Enrique | Muskiz | Sollube | 1937-05-13T00:00:00Z | Muerto frente / Frontean hila |
| 21. | Abascal Gabilondo, Maria | Elgoibar | Gernika | 1937-04-26T00:00:00Z | Bombardeo / Bonbardaketan hila |
| 22. | Abascal Laza, Bautista | Bilbo | Bilbo | 1937-12-16T00:00:00Z | Fusilado / Fusilatua |
| 23. | Abascal Nuñez, Juana | Bilbo | Derio | 1938-12-17T00:00:00Z | Fusilado / Fusilatua |
| 24. | Abascal Pescador, Felipe | Donostia | Donostia | 1938-05-25T00:00:00Z | Fusilado / Fusilatua |
| 25. | Abascal Saldamando, Angel | | | | Muerto frente / Frontean hila |
| 26. | Abascal Trueba, Pedro | Arteaga | Bilbo | 1937-12-13T00:00:00Z | Fusilado / Fusilatua |
| 27. | Abasolo Goicoechea, Jose Luis | Castillo-Elejabeitia | | 1937-04-03T00:00:00Z | Muerto frente / Frontean hila |
| 28. | Abasolo Goicoechea, Victor | Castillo-Elejabietia | | 1937-04-03T00:00:00Z | Muerto frente / Frontean hila |
| 29. | Abasolo Gordibil, Miren | Gasteiz | Beasain | 1936-07-24T00:00:00Z | Bombardeo / Bonbardaketan hila |
| 30. | Abasolo Oribe, Benito | Donostia | Donostia | 1938-08-13T00:00:00Z | Fusilado / Fusilatua |
| 31. | Abejon Ruiz, Manuel | Sestao | Baranbio | 1936-12-10T00:00:00Z | Muerto frente / Frontean hila |
| 32. | Abel Barañano, Antonio | Bilbo | Legutio | 1936-12-02T00:00:00Z | Muerto frente / Frontean hila |
| 33. | Abellanal Monariz, Lorenzo | | Zornotza | 1937-05-16T00:00:00Z | Muerto frente / Frontean hila |
| 34. | Abelleira Blanco, Ramon | Erandio | Legutio | 1936-12-05T00:00:00Z | Muerto frente / Frontean hila |
| 35. | Abendaño Iturriaga, Jesus | Dima | Dima | 1937-06-19T00:00:00Z | Otras circunstancias / Beste batzuk |
| 36. | Aberasturi Goti, Julia | Bilbo | Bilbo | 1937-04-22T00:00:00Z | Bombardeo / Bonbardaketan hila |
| 37. | Aberasturi Lazcano, Fructuoso | Oñati | Bilbo | 1937-07-22T00:00:00Z | Fusilado / Fusilatua |
| 38. | Abio Achurra, Policarpo | Mundaka | Legutio | 1936-12-12T00:00:00Z | Muerto frente / Frontean hila |

*Ilustración 1: Información inicial*

## Examinar los datos con Facet/Text Facet. Por ejemplo, en Vecindad o Lugar de Muerte. Identifica posibles errores

Se empezará con examinar los datos de la Vecindad, clicando en el triangulo desplegable de esa columna y clicando en Facet->Text facet. Como se puede observar en la Ilustración 2, se agrupa toda la información y sus repeticiones en un view en la zona izquierda de pantalla.


*Ilustración 2: Facet de la columna Vecindad*

Estos datos podrían no ser del todo fiables por lo que se hará un análisis cluster para detectar posibles erroes. Con este análisis se resolverán pequeñas diferencias entre nombres de pueblos y ciudades como se puede observar en la Ilustración 3. Aun así hay casos de pueblos como Palencia y Valencia que tienen un parecido muy grande pero que se tratan de pueblos diferentes, de ahí que se de la opción a seleccionar el cluster.

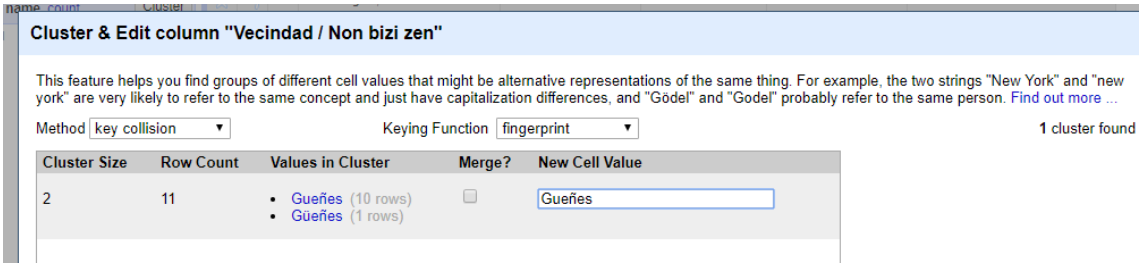



*Ilustración 3: Ejemplo de cluster*

Después de analizarlo con el cluster de palabras parecidas, se han encontrado parecidos entre más pueblos a los que le faltaban guiones o -a al final, algo muy común en pueblos de Euskadi como son Legazpi y Legazpia, que se tratan del mismo pueblo, uno redactado en Euskera y el otro en castellano.

A continuación se procede a hacer lo mismo con el lugar de su muerte, ya que pueden haber más diferencias.

## Cluster en la columna de personas y lugares

Para hacer el cluster sobre personas se ha clicado sobre Edit cells -> Edit and cluster y nos aparece una tabla como la que se muestra en la ilustración 4.

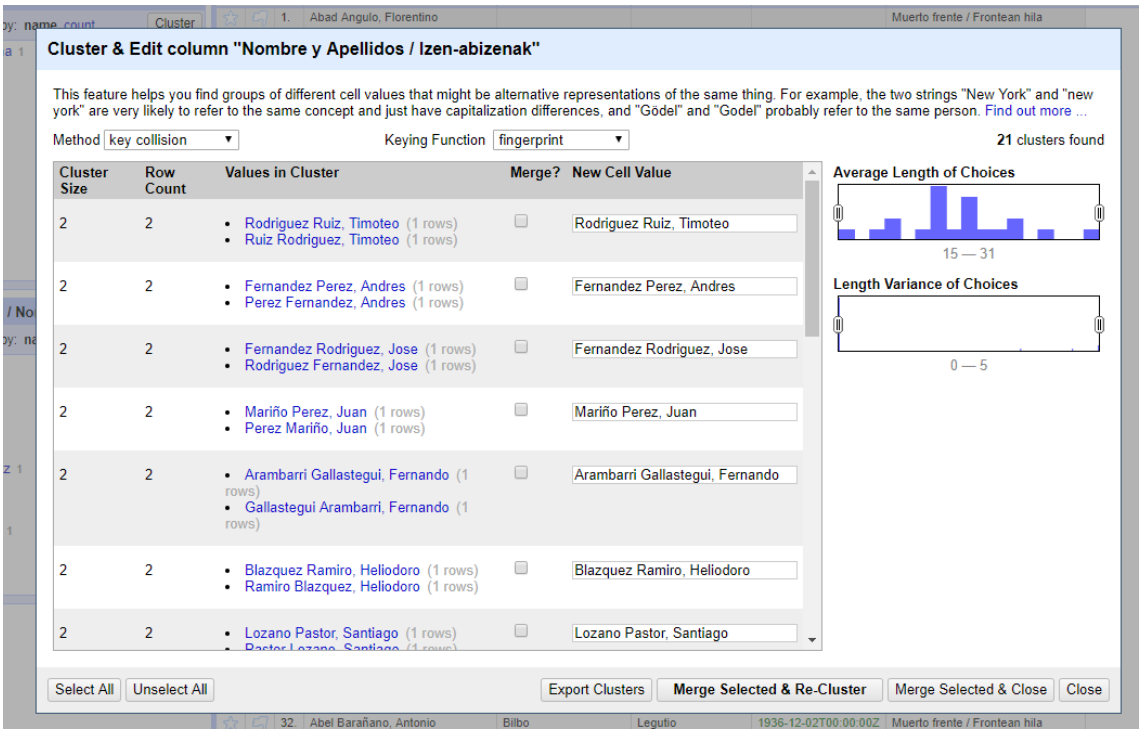



*Ilustración 4: Ejemplo de cluster sobre la columna personas*

En esta ventana se pueden ver los nombres y apellidos de personas que pueden coincidir entre orden o apellidos parecidos. Esto hay que tenerlo en cuenta ya que se ha podido meter al revés y realmente se trata de una misma persona. Ya que no se puede tener el criterio de nombre y apellidos para saber si se trata de una misma persona, en cada fila hay una opción "Browse this cluster" que dará una información más detallada sobre la persona sobre la que se hará el cluster. Al clicar en la primera fila, se obtiene la información mostrada en la ilustración 5.

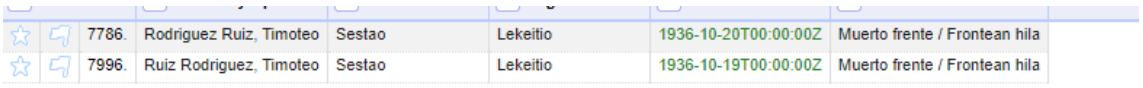



*Ilustración 5: Información detallada sobre el primer dato del cluster*

Al analizar estos datos, se puede observar que definitivamente sí se trata de esa misma persona, por lo que se le pondrá el nombre en común a los dos.

Hay casos en los que, aunque la persona sea la misma, aparentemente, hay datos ausentes, como el caso de Fernando Arambarri, al que le faltan los datos como la Vecindad, aun así, ya que todos los demás campos de las filas coinciden, se tomarán como una misma persona.

## Buscar errores ortográficos y registros duplicados

Para eliminar errores ortográficos y registros duplicados se ordena por nombre de las personas de la tabla y se empiezan a clusterizar para evitar duplicados, todo esto se ha realizado en el punto anterior Cluster en la columna de personas y lugares

## Reconciliar la columna lugares con Wikidata

Si se quiere reconciliar una columna con wikidata, se clica en el desplegable de esa columna-> Reconcile->Start reconciling…

Una vez clicado ahí se abrirá la una ventana como la que se muestra en la Ilustración 6.

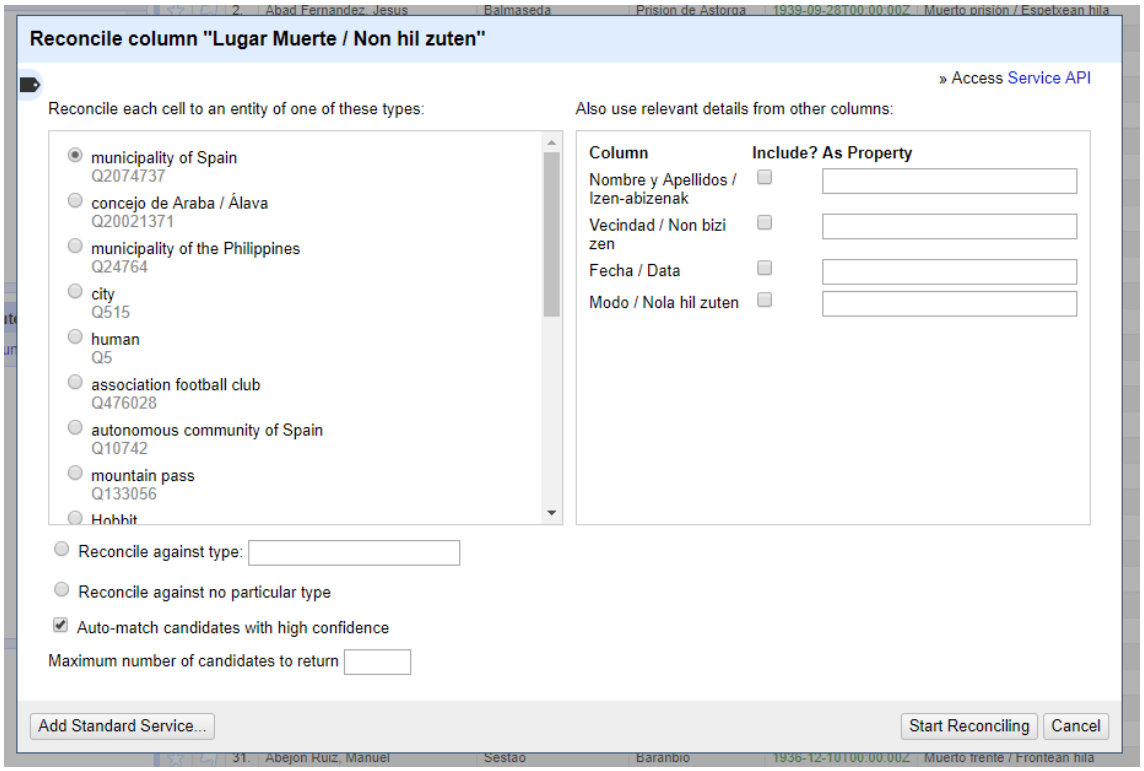



*Ilustración 6: Reconcilio de datos con Wikidata*

En este caso reconciliaremos la información por la columna municipios de España. El resultado es el mostrado en la Ilustración 7.


*Ilustración 7: Resultado de reconciliación de pueblos de España*

Como se puede observar, 1284 entradas de han quedado en blanco y 5549 se han matcheado. El resto, 2767 no se han podido reconciliar debido a que algunos de ellos, como por ejemplo “Prisión de Astorga”, que no corresponde a un municipio, sino que, a un establecimiento, en este caso una prisión. Podría especificarse que esa prisión se ubica en Astorga, pero sería pérdida de información. Por otro lado, hay otros que no se informan en Wikidata por lo que no se han podido matchear, como por ejemplo Larrauri o Azazeta.

## Enriquecer el conjunto de datos a partir de Wikidata

Ya que hay datos en los lugares de nacimiento que, como se ha visto en el punto anterior, Reconciliar la columna lugares con Wikidata, se han reconciliado. A partir de estos datos se puede sacar otra información y así enriquecer la información actual. Para ello se clica en el desplegable de Lugar muerte->Edit Column->Add columns from reconciled values. Se enriquecerán los datos de sus coordenadas y su área, como se muestra en la Ilustración 8.

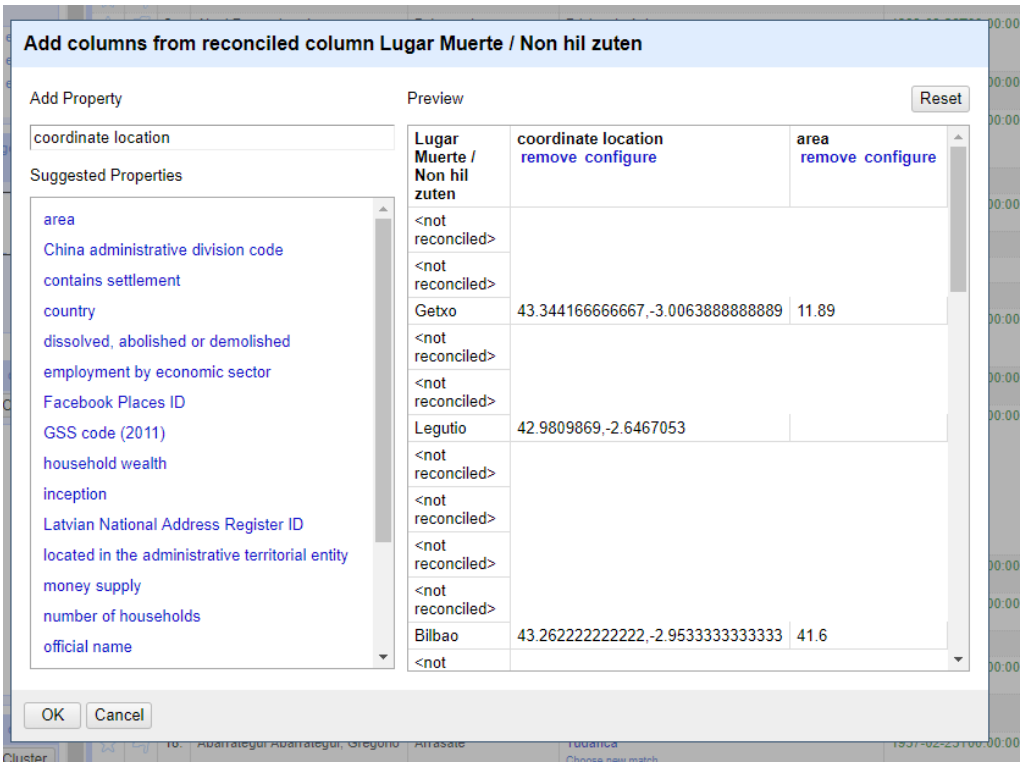



*Ilustración 8: Añadir columnas de coordenadas y área*

Una vez terminado el proceso, se puede observar como a la tabla actual se le han añadido dos columnas nuevas, que son las coordenadas de los municipios y su área. Esta información se ha recibido solo para las filas en las que se ha conseguido reconciliar con la información de Wikidata. Todas las demás se quedan en blanco, ya que no hay información acerca de ellas. El resultado final se puede ver en la ilustración 9, filtradas también por municipio, cogiendo el municipio Durango, Bizkaia.

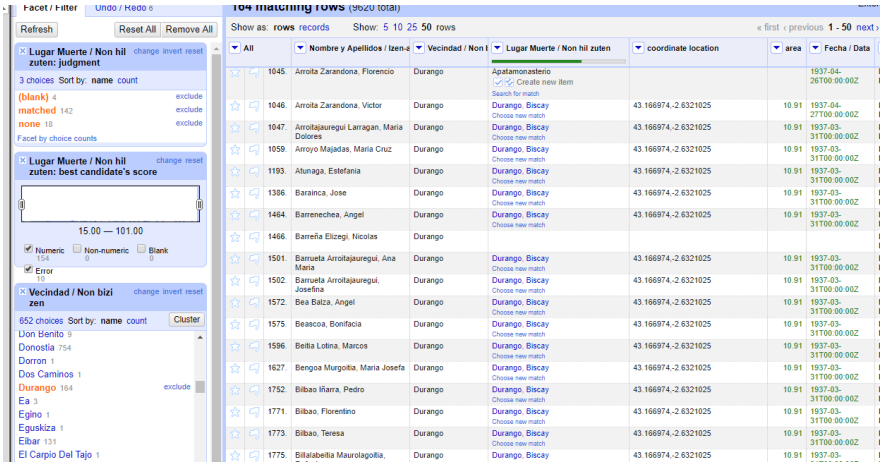


*Ilustración 9: Ejemplo del enriquecimiento con wikidata*