

Introducción a Google Refine

Por David Cabo

Google Refine es una herramienta gratuita y muy poderosa, diseñada con dos objetivos en mente: ayudar a entender la estructura y calidad de unos datos, y permitir corregir determinados tipos de errores comunes en ellos. Si bien algunas de las tareas de análisis pueden hacerse a través de otras herramientas (como Excel), su enfoque en la “limpieza de datos” de forma cómoda e intuitiva hacen de Google Refine una herramienta muy valiosa en cualquier procesos que maneje datos.

Un caso práctico: desaparecidos de la Guerra Civil

Vamos a mostrar algunas de las principales funcionalidades de Google Refine a través de un caso concreto y real: el análisis de los datos sobre personas muertas, fusiladas o desaparecidas durante la Guerra Civil y la dictadura en Euskadi.

Los datos sobre desaparecidos y fosas comunes están disponibles en [Open Data Euskadi](#), y en concreto nosotros trabajaremos con el fichero Excel que contiene la [lista de personas desaparecidas](#). Como sucede en prácticamente todos los casos, este fichero contiene una serie de erratas e inconsistencias que deben ser corregidas antes de analizar los datos y utilizarlos en un artículo, visualización o una estadística.

Instalar Google Refine

Existen instrucciones detalladas para instalar Google Refine, [en inglés](#), pero básicamente se reducen a descargar la última versión del programa [en este enlace](#) y:

- en Windows, descomprimir el fichero .zip en una carpeta, y ejecutar el programa google-refine.exe con el icono del diamante. En caso de que el ordenador no tenga instalado Java la instalación comenzara automáticamente; también se puede instalar Java [manualmente](#). Si todo es correcto aparecerá una pantalla con fondo negro (la línea de comandos de Windows), y a los pocos segundos se abrirá una página web en el navegador de Internet.
- en Mac OS X, es necesario descargar el fichero .dmg, abrirlo, y copiar el programa a la carpeta “Aplicaciones” (Applications), arrastrando el icono del diamante a esa carpeta. Una vez copiado, para arrancar el programa hay que hacer doble click en el diamante y entonces se te abrirá en el navegador.

En cualquiera de los dos casos es recomendable que se utilicen los navegadores Firefox o Chrome ya que Internet Explorer a veces da problemas con este programa.

Google Refine se ejecuta localmente en el ordenador en el que se instala, pero no tiene un interfaz de usuario propio, sino que utiliza un navegador de Internet para ello, a través de la dirección:

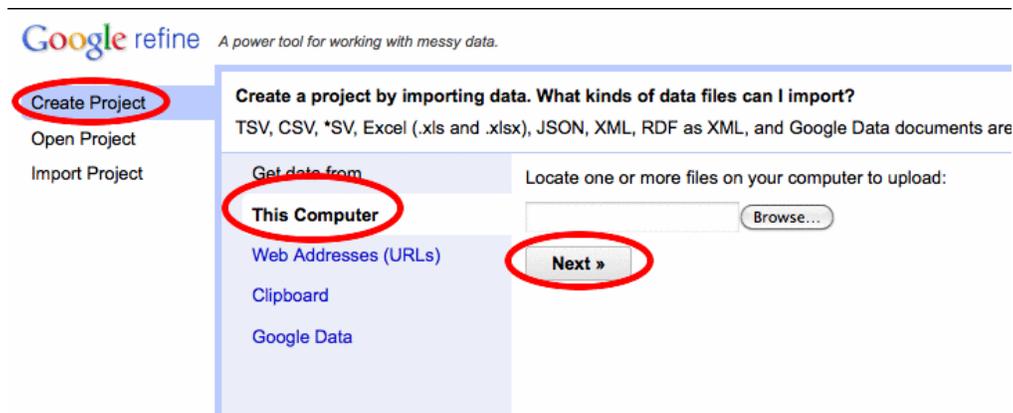
<http://127.0.0.1:3333>

dónde “127.0.0.1” significa, en la web, “la dirección del ordenador en el que estoy”, es decir, “este ordenador”. Esto es importante para darse cuenta de que los datos que se usan en Google Refine en ningún momento se suben a la red, ni pueden ser vistos por Google u otras personas.

Crear un proyecto

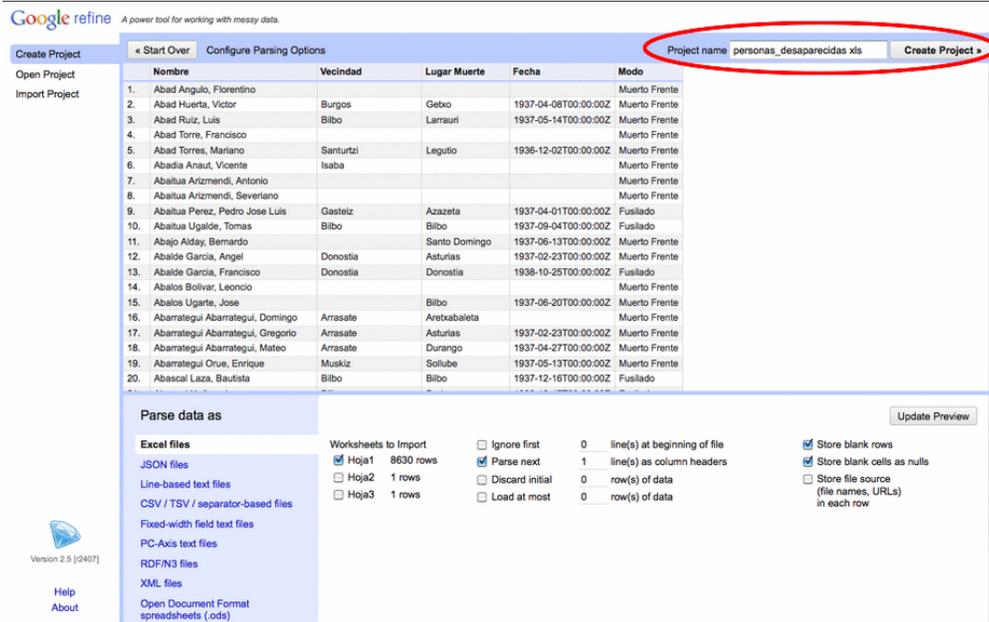
Al usar Google Refine llamamos “proyecto” (project) a un fichero que queremos explorar y modificar. En nuestro caso, por ejemplo, el fichero de personas desaparecidas.

La pantalla de inicio de Google Refine nos permite abrir proyectos creados anteriormente (“Open Project” – “Abrir Proyecto”), o crear uno nuevo utilizando datos que tengamos en nuestro ordenador (“Get data from... this computer” – “Obtener datos de... este ordenador”), en una dirección web (“...Web Addresses (URLs)”) o en el portapapeles (“...clipboard”). En nuestro caso seleccionaremos el fichero ‘personas_desaparecidas.xls’ que habíamos descargado previamente.



*Pantalla de arranque de Google Refine.
Pestaña de creación de un nuevo proyecto.*

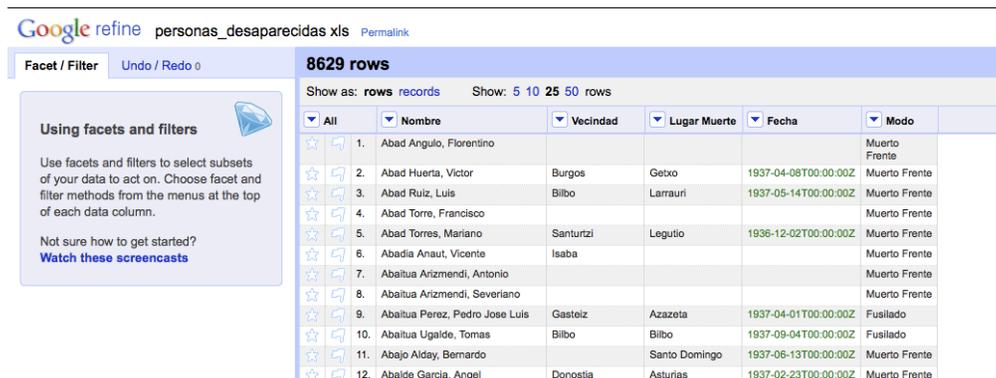
En la siguiente pestaña Refine nos muestra un subconjunto de los datos, y nos permite ajustar una serie de opciones (para por ejemplo limitar el número de registros a tratar, o indicar si existe una primera línea con el nombre de las columnas), que en nuestro caso no tocaremos. Podemos ajustar el nombre del proyecto, que por defecto es el nombre del fichero, y pulsar el botón “Create Project” (“Crear Proyecto”):



Pantalla de creación de un proyecto, mostrando los datos a utilizar.

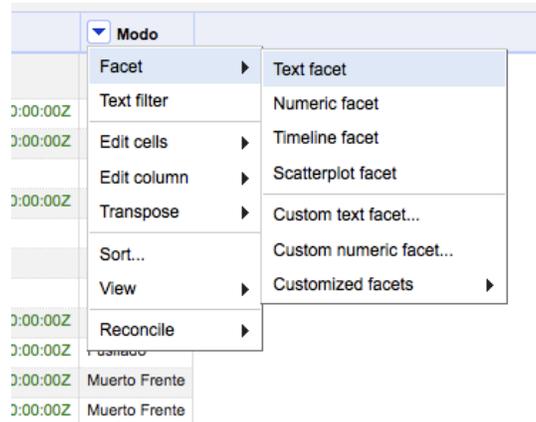
Explorar los datos

Una vez creado el proyecto vemos la pantalla de trabajo de Google Refine, que se compone de dos partes fundamentales: la sección derecha muestra las primeras líneas de los datos (5, 10, 25 o 50 líneas, configurable), mientras que la columna de la izquierda muestra los filtros aplicados, que se explican a continuación:



Pantalla de trabajo en Refine, con los datos a la derecha, y filtros a la izquierda.

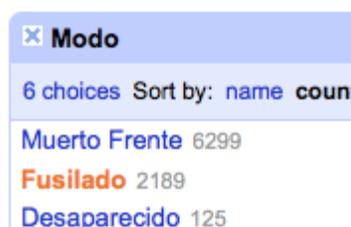
Google Refine permite analizar y filtrar los datos mediante el uso de “facetas” (‘facets’ en inglés), es decir, filtros en cada una de las columnas de nuestro fichero. Como ejemplo, vamos a crear una faceta en la columna “Modo” para ver las causas de muerte en nuestros datos. Hacemos click en el botón que parece un triángulo azul que apunta para abajo a la izquierda de la columna “Modo”. Esto desplegará un menú. Elegimos “Facet > Text Facet” (“Faceta > Faceta de Texto”):



Ahora la columna izquierda de Google Refine nos muestra una “faceta” o filtro con todos los valores distintos de la columna “Modo”, así como el número de repeticiones de cada valor:



Pulsando en uno de los valores mostrados solo veremos en la sección derecha las personas que murieron por esa causa. Por ejemplo, podemos ver solamente los fusilados:



Para volver a ver todos los datos, podemos volver a hacer click en la palabra “Fusilado”, o seleccionar “reset” (restaurar) en la parte superior derecha de la faceta:



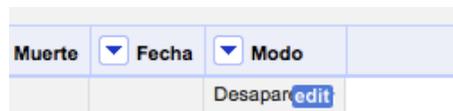
También podemos ordenar los valores por número de repeticiones, en vez de por orden alfabético, seleccionando la opción “count” (recuento). Vemos así que la causa de muerte más común es “Muerto Frente” (6299 casos), seguido de “Fusilado” (2189 casos):



Podemos ver también que existen trece casos en los que no se conoce la causa de la muerte (“blank” significa “vacío”), un caso inconsistente en el que se registra el lugar y no la causa de muerte (“Miranda de Ebro”), así como erratas que trataremos a continuación.

Editar los datos

Vamos a corregir una de las erratas: en la faceta de “Modo” elegimos la palabra “Desaparecido”, y vemos el registro que contiene el error en la sección de la derecha. Ahora podemos corregirlo de dos formas, pulsando en el botón “edit” (editar) que aparece al pasar el cursor por encima del error en la parte derecha de la pantalla:



O pulsando “edit” al pasar el cursor por encima del valor en la faceta de la izquierda:



Por cualquiera de los dos métodos podemos corregir la errata y guardar el cambio dándole al botón “Apply” (aplicar):



Si la errata se repitiera más de una vez Refine nos daría la opción de aplicar el cambio en todos los casos (“Apply to All Identical Cells” – “Aplicar a Todas las Celdas Idénticas”). Posteriormente veremos también cómo Refine nos permite buscar y corregir erratas automáticamente.

Google Refine guarda automáticamente todos los cambios realizados, y permite retroceder un número ilimitado de pasos en cualquier momento, a través de la pestaña “Undo/Redo” (Deshacer/Rehacer) situada en la parte superior izquierda:



Pestaña “Undo/Redo” mostrando todos los cambios realizados, y permitiendo retroceder a cualquiera de ellos.

Antes de pasar a la siguiente sección cerramos la faceta que hemos creado usando la cruz situada en la parte superior izquierda de ésta:



Separar nombre de apellidos

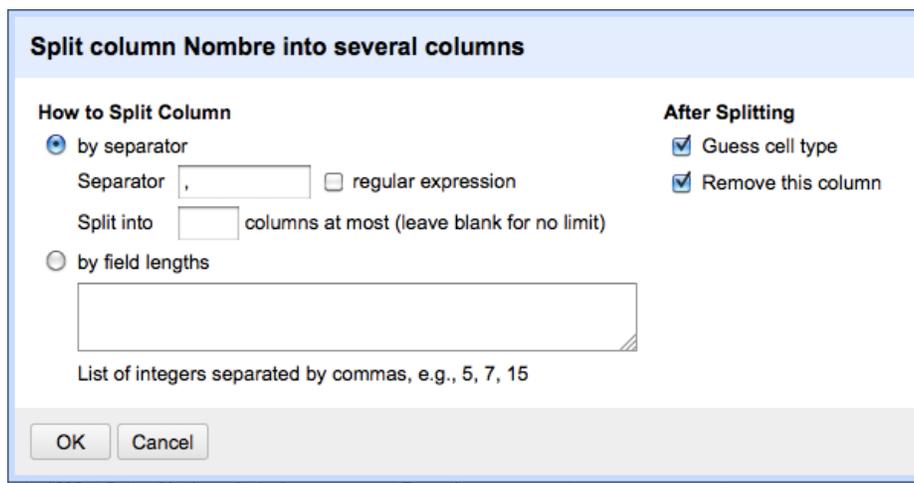
Refine permite aplicar una serie de transformaciones a los datos, desde las más sencillas (eliminar espacios innecesarios al final de las columnas) a las más sofisticadas (cambiar el formato de fechas o de direcciones utilizando expresiones regulares).

Ahora, la columna “Nombre” tiene formato “Apellidos, Nombre” y el contenido en la misma celda. Vamos a partirla en dos partes para tener “Apellidos” y “Nombre” en dos columnas

diferentes y ver qué apellidos se repiten más frecuentemente. Para ello seleccionamos “Edit column > Split into several columns...” (“Editar columna > Dividir en varias columnas...”) del menú de la columna “Nombre”:



En nuestro caso queremos partir la columna utilizando la coma que separa el nombre de los apellidos como separador (separator). Esta es la opción por defecto, pero Refine nos permitiría usar cualquier otro separador:



Podemos cambiarles el nombre a las columnas resultantes por claridad, de forma que “Nombre 1” sea “Apellidos”, y “Nombre 2” simplemente “Nombre”. Pulsa en el menú, “Edit column > Rename this column” (“Editar columna > Renombrar esta columna”):

Nombre 1	Nombre 2	Nombre 3	Vecindad
Facet	Florentino		
Text filter	Victor		Burgos
Edit cells	Luis		Bilbo
Edit column	Split into several columns...		
Transpose	Add column based on this column...		
Sort...	Add column by fetching URLs...		
View	Add columns from Freebase ...		
Reconcile	Rename this column		

Vemos también que se ha creado una columna “Nombre 3” de forma inesperada. Creando una faceta de texto en “Nombre 3” descubrimos que solo un registro contiene información:

Nombre 3 change invert reset

1 choices Sort by: name count Cluster

Victoriana 1 exclude

(blank) 8628

Facet by choice counts

Uno de los nombres de los desaparecidos contenía dos comas, por lo que Refine ha creado tres columnas:

All	Apellidos	Nombre	Nombre 3	Vecindad	Lugar Muerte	Fecha	Modo
3749	Gonzalez De Larralde	Lopez De Suso	Victoriana	Gasteiz	Gasteiz	1936-12-01T00:00:00Z	Fusilado

Lo más sencillo es corregir el error a mano, como vimos anteriormente:

All	Apellidos	Nombre	Nombre 3	Vecindad	Lugar Muerte	Fecha	Modo
3749	Gonzalez De Larralde Lopez De Suso	Victoriana	Victoriana	Gasteiz	Gasteiz	1936-12-01T00:00:00Z	Fusilado

y borrar la columna “Nombre 3”. Haz click en el menú, “Edit column > Remove this column” (“Editar columna > Eliminar esta columna”):

Nombre 3	Vecindad	Lugar Muerte	Fecha
Facet		Gasteiz	1936-12-01T00:00:00Z
Text filter			
Edit cells			
Edit column		Split into several columns...	
Transpose		Add column based on this column...	
Sort...		Add column by fetching URLs...	
View		Add columns from Freebase ...	
Reconcile		Rename this column	
		Remove this column	

Con los nombres separados correctamente de los apellidos, ya podemos continuar. Google Refine nos permite crear varias facetas simultáneamente, y filtrar por todas ellas a la vez, así que podemos crear facetas para las columnas “Modo”, “Apellidos” y “Lugar Muerte”, e investigar por ejemplo los fusilamientos en Bilbao filtrando por “Fusilado” en la faceta de “Modo” y “Bilbo” en la faceta “Lugar Muerte”. Podemos recorrer entonces la lista de apellidos repetidos para encontrar por ejemplo el caso de dos personas posiblemente hermanos fusilados el mismo día:

Modo change invert reset

1 choices Sort by: name count Cluster

Fusilado 2 exclude

Facet by choice counts

Lugar Muerte change invert reset

1 choices Sort by: name count Cluster

Bilbo 2 exclude

Facet by choice counts

Apellidos change invert reset

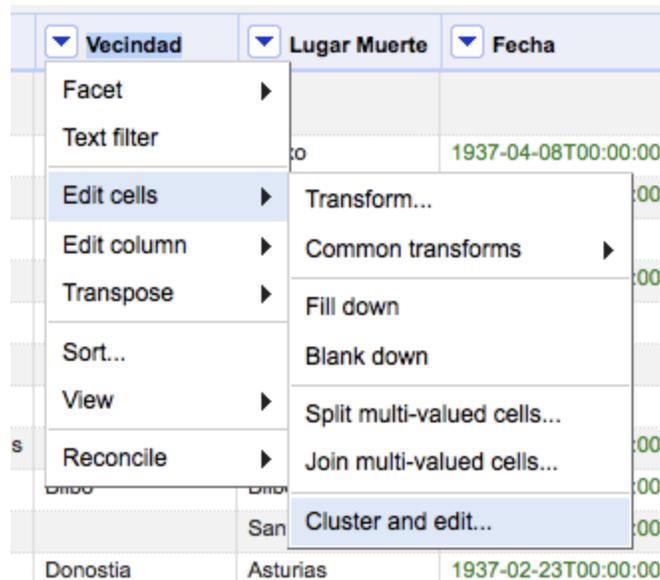
502 choices Sort by: name count Cluster

- Cueto Ibañez 3
- Artaza Llantada 2
- Curiel Cordon 2
- Fernandez Garcia 2
- Fernandez Ruiz 2
- Flores Lazcano 2
- Mirones Garcia 2
- Perez Garcia 2
- Torres Bonet** 2 exclude
- Villa Ateca 2
- Abaitua Ugalde 1

All	Apellidos	Nombre	Vecindad	Lugar Muerte	Fecha	Modo
☆	7851. Torres Bonet	Crescencio	Bilbo	Bilbo	1938-03-12T00:00:00Z	Fusilado
☆	7852. Torres Bonet	Vicente	Bilbo	Bilbo	1938-03-12T00:00:00Z	Fusilado

Clustering

Google Refine nos permite hallar erratas e inconsistencias de forma automática, usando lo que en inglés se conoce como “clustering”, que consiste en detectar conjuntos de valores muy similares entre sí. En nuestro caso vamos a buscar posibles erratas en la columna “Vecindad” que, como veremos, contiene algunos fallos. Para ello seleccionamos “Edit cells > Cluster and edit...” (“Editar celdas > Agrupar y editar...”) en el menú de dicha columna:



Vecindad	Lugar Muerte	Fecha
Donostia	Asturias	1937-02-23T00:00:00.

Obtenemos una pantalla desde la que podemos configurar y revisar el funcionamiento del clustering:

Cluster & Edit column "Vecindad"

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 1 cluster found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	9	<ul style="list-style-type: none"> Gueñes (8 rows) Güeñes (1 rows) 	<input type="checkbox"/>	Gueñes

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

En nuestro ejemplo Google Refine ha detectado que existen dos pueblos cuyos nombres son tan similares que parecen ser una errata, "Gueñes" y "Güeñes":

Cluster Size	Row Count	Values in Cluster
2	9	<ul style="list-style-type: none"> Gueñes (8 rows) Güeñes (1 rows)

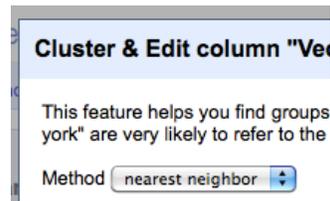
Vemos que "Gueñes" (la grafía vasca) se repite ocho veces, mientras que "Güeñes" (la grafía castellana) solo aparece una vez. Podemos decidir usar una u otra, en función de nuestras preferencias, pero debemos ser consistentes y utilizar siempre la misma, para evitar confusiones y estadísticas equivocadas. Pulsando sobre uno de los nombres le estamos diciendo a Google Refine que en efecto se ha detectado una errata, y que queremos utilizar ese nombre en todos los casos. Decidimos usar la grafía en euskera, así que hacemos click sobre la palabra "Gueñes". Una pequeña marca a la derecha de los nombres nos recordará entonces que le hemos pedido a Refine que consolide ("merge" en inglés) los distintos valores:

Values in Cluster	Merge?	New Cell Value
<ul style="list-style-type: none"> Gueñes (8 rows) Güeñes (1 rows) 	<input checked="" type="checkbox"/>	Gueñes

Para aplicar los cambios seleccionamos el botón “Merge Selected & Re-Cluster” (“Consolidar las selecciones y reagrupar”) en la parte inferior de la pantalla. Los registros que contenían la palabra “Güeñes” ahora han sido modificados para usar la grafía vasca, por lo que Google Refine no encuentra más erratas en este momento.

Por defecto Google Refine utiliza un método de búsqueda de erratas conservador, es decir, poco agresivo, y que puede no detectar todos los errores existentes. El proceso (“key collision - fingerprint” en inglés) consiste en convertir el texto a minúsculas, quitar espacios y signos de puntuación y ordenar las palabras alfabéticamente: si el resultado de aplicar estos pasos a dos textos distintos es el mismo, se presume que es una errata, como hemos visto en el caso de “Gueñes” y “Güeñes”. También podría haber detectado erratas como “Garcia” en vez de “García”, o incluso “Lopez, Antonio” en vez de “Antonio López”.

Un método alternativo de búsqueda de errores se conoce como “el vecino más cercano” (o “nearest neighbor” en inglés), y consiste en calcular el número de cambios que hay que hacer a una palabra para convertirla en otra: si un par de palabras se encuentran a una “distancia” menor que un límite fijado por nosotros, se considera que es una errata. Para probar este método en nuestro ejemplo, elegimos “nearest neighbor” en el desplegable “Method” situado en la parte superior izquierda:



Vemos inmediatamente que Refine encuentra ahora una serie de inconsistencias como “Legazpia” y “Legazpi” (que se diferencian en la ‘a’ final) o erratas como “Gallarta” (con tres eles) en vez de “Gallarta”. Podemos ahora corregir estos errores como acabamos de ver en el caso de “Gueñes”, pulsando en el nombre correcto en cada caso. Pero debemos estar atentos para no “corregir” aquellos casos en los que no hay ningún error, como “Palencia” y “Valencia”:

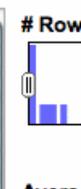
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none"> Castillo-Elejabeitia (1 rows) Castillo Elejabeitia (1 rows) 	<input checked="" type="checkbox"/>	Castillo-Elejabeitia
2	17	<ul style="list-style-type: none"> Legazpia (11 rows) Legazpi (6 rows) 	<input checked="" type="checkbox"/>	Legazpia
2	76	<ul style="list-style-type: none"> Gallarta (75 rows) Gallarta (1 rows) 	<input checked="" type="checkbox"/>	Gallarta
2	14	<ul style="list-style-type: none"> Palencia (11 rows) Valencia (3 rows) 	<input type="checkbox"/>	Palencia
2	8	<ul style="list-style-type: none"> Zigoitia (5 rows) Rigoitia (3 rows) 	<input type="checkbox"/>	Zigoitia
2	2	<ul style="list-style-type: none"> Errigoiti (1 rows) Errigoitia (1 rows) 	<input checked="" type="checkbox"/>	Errigoitia

Podemos ser más agresivos en la búsqueda de errores aumentando el “radio” (radius) o “distancia” por debajo del cual consideramos que dos palabras son iguales. Si aumentamos la distancia a “2”, por ejemplo, Refine encuentra 13 posibles errores, frente a los seis anteriores:

all values that might be alternative representations of the same thing. For example, tit and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person.

Function Radius Block Chars

Cluster	Merge?	New Cell Value	# Row
(rows) (rows)	<input type="checkbox"/>	Rigoitia	3
(rows) name)	<input type="checkbox"/>	Murrieta	2



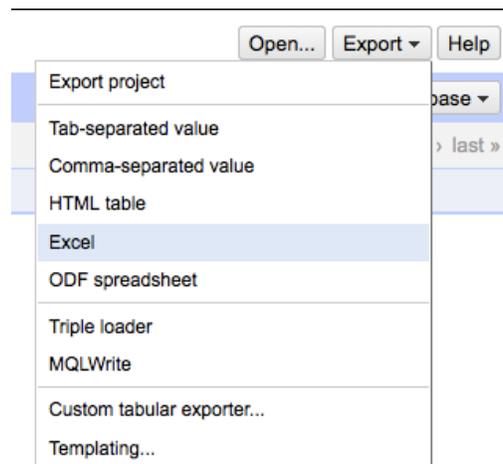
La búsqueda de erratas es un proceso exploratorio que depende mucho del tipo de errores que tengan los datos. Lo más recomendable es comenzar por un método “conservador”, como el de “key collision” que nos muestra Google Refine por defecto, y pasar después a uno más agresivo (“vecino más cercano”), incrementado progresivamente el parámetro “Radio” (Radius). Estos últimos métodos encuentran más erratas, pero también muestran “falsos positivos” (como es el caso de “Valencia” y “Palencia”), por lo que requieren mayor atención al utilizarlos. La documentación de Google Refine [sobre clustering](#) contiene una sección en la que se explica [en mayor detalle los distintos algoritmos disponibles](#).

Exportar el resultado

Todos los cambios que hemos ido haciendo son registrados internamente por Refine sin modificar el fichero original. Por ello, una vez hayamos terminado de analizar y limpiar nuestros datos debemos exportarlos para utilizarlos con otros programas como Excel/ OpenOffice o Google Fusion Tables.

Para ello utilizamos el botón “Export” en la esquina superior derecha de la pantalla, y elegiremos uno de los formatos disponibles. Los más comunes son:

- “Comma-separated value” (CSV), un fichero de texto con los valores separados por comas.
- “Excel” (XLS)
- “ODF spreadsheet”, utilizado por Open Office (que también puede leer ficheros Excel).



Más información

Este tutorial describe solo algunas de las funciones más básicas de Google Refine, que contiene herramientas avanzadas muy potentes, como el “clustering”, una amplia librería de funciones, expresiones regulares y “reconciliación” con bases de datos externas.

Los siguientes enlaces contienen información de referencia, pero solo están disponibles en inglés:

- [Guía del usuario: ¿qué es Google Refine?](#)
- [Documentación para usuarios](#)
- [Vídeos de ejemplo](#)