

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 <u>Open Data Euskadi</u>, y en concreto nosotros trabajaremos con el fichero Excel que contiene la <u>lista de personas</u> <u>desaparecidas</u>. 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, <u>en inglés</u>, pero básicamente se reducen a descargar la última versión del programa <u>en este enlace</u> 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 <u>manualmente</u>. 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.

Google refine	A power tool for working with messy data.					
Create Project Open Project	Create a project by importing data. What kinds of data files can I import? TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are					
Import Project	Get date from This Computer Web Addresses (URLs) Clipboard Google Data	Locate one or more files on your computer to upload: Browse Next »				



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"):



eate Project	« !	Start Over	Configure Parsing Opt	ions					Proje	ect name personas_desaparecidas xls	Create Project
en Project		Nombre		Vecindad	Lugar	Muerte	Fecha		Modo		
art Project	1.	Abad Ang	ulo, Florentino						Muerto Frente		
	2.	Abad Huer	rta, Victor	Burgos	Getxc		1937-04-081	00:00:00Z	Muerto Frente		
	3.	Abad Ruiz	, Luis	Bilbo	Larrau	iri	1937-05-141	00:00:00Z	Muerto Frente		
	4.	Abad Torre	e, Francisco						Muerto Frente		
	5.	Abad Torre	es, Mariano	Santurtzi	Legut	0	1936-12-021	00:00:00Z	Muerto Frente		
	6.	Abadia An	aut, Vicente	Isaba					Muerto Frente		
	7.	Abaitua Ar	rizmendi, Antonio						Muerto Frente		
	8.	Abaitua Ar	rizmendi, Severiano						Muerto Frente		
	9.	Abaitua Pe	erez, Pedro Jose Luis	Gasteiz	Azaze	ta	1937-04-011	00:00:00Z	Fusilado		
	10.	Abaitua Ug	galde, Tomas	Bilbo	Bilbo		1937-09-041	00:00:00Z	Fusilado		
	11.	Abajo Alda	ay, Bernardo		Santo	Domingo	1937-06-131	00:00:00Z	Muerto Frente		
	12.	Abalde Ga	arcia, Angel	Donostia	Asturi	as	1937-02-231	00:00:00Z	Muerto Frente		
	13.	Abalde Ga	ircia, Francisco	Donostia	Donor	itia	1938-10-251	00:00:00Z	Fusilado		
	14.	Abalos Bo	livar, Leoncio						Muerto Frente		
	15.	Abalos Ug	arte, Jose		Bilbo		1937-06-201	00:00:00Z	Muerto Frente		
	16.	Abarrateg	ui Abarrategui, Domingo	Arrasate	Areba	ibaleta			Muerto Frente		
	17.	Abarrategu	ui Abarrategui, Gregorio	Arrasate	Asturi	85	1937-02-231	00:00:00Z	Muerto Frente		
	18.	Abarrateg	ui Abarrategui, Mateo	Arrasate	Duran	go	1937-04-271	00:00:00Z	Muerto Frente		
	19.	Abarrateg	ui Orue, Enrique	Muskiz	Sollut	0	1937-05-131	00:00:00Z	Muerto Frente		
	20.	Abascal La	aza, Bautista	Bilbo	Bilbo		1937-12-161	00:00:00Z	Fusilado		
	F	Parse dat	a as								Update Preview
	E	xcel files		Worksheets	to Import	🗌 Igr	ore first	0 1	ine(s) at beginnir	ng of file 🗹 Store blan	k rows
	J	SON files		🗹 Hoja1	8630 rows	🗹 Pa	rse next	1 1	ine(s) as column	headers Store blan	k cells as nulls
	L	ine-based to	ext files	Hoja2	1 rows	Dis	card initial	0	row(s) of data	Store file s (file name)	ource s. URLs)
	c	SV/TSV/	separator-based files	🖂 нојаз	Trows		ad at most	0	row(s) of data	in each ro	N
	F	ixed-width f	field text files								
ersion 2.5 (r2407)		C-Axis text	files								
	5	ML files									
Help	ć	pen Docum	nent Format								

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:

Google refine personas_desap	arecidas xls	Per	malink					
Facet / Filter Undo / Redo 0	8629 r	row	s					
	Show as	Show as: rows records Show: 5 10 25 50 rows						
Using facate and filtors			Nombre	Vecindad	Lugar Muerte	Fecha	Modo	
Line facets and filters to select subsets	* 5	1.	Abad Angulo, Florentino				Muerto Frente	
of your data to act on Choose facet and	☆되	2.	Abad Huerta, Victor	Burgos	Getxo	1937-04-08T00:00:00Z	Muerto Frente	
filter methods from the menus at the top	\$ 5	3.	Abad Ruiz, Luis	Bilbo	Larrauri	1937-05-14T00:00:00Z	Muerto Frente	
of each data column.	☆듸	4.	Abad Torre, Francisco				Muerto Frente	
Not sure how to get started?	☆되	5.	Abad Torres, Mariano	Santurtzi	Legutio	1936-12-02T00:00:00Z	Muerto Frente	
Watch these screencasts	☆되	6.	Abadia Anaut, Vicente	Isaba			Muerto Frente	
	\$ 57	7.	Abaitua Arizmendi, Antonio				Muerto Frente	
	\$ 57	8.	Abaitua Arizmendi, Severiano				Muerto Frente	
	\$ 5	9.	Abaitua Perez, Pedro Jose Luis	Gasteiz	Azazeta	1937-04-01T00:00:00Z	Fusilado	
	\$ 57	10.	Abaitua Ugalde, Tomas	Bilbo	Bilbo	1937-09-04T00:00:00Z	Fusilado	
	☆되	11.	Abajo Alday, Bernardo		Santo Domingo	1937-06-13T00:00:00Z	Muerto Frente	
	5 57	12	Abalde Garcia, Angel	Donostia	Asturias	1937-02-23T00:00:00Z	Muerto Frente	

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 triangulo azul que apunta para abajo a la izquierda de la columna "Modo". Esto desplegará un menú. Elegimos "Facet > Text Facet" ("Faceta > Faceta de Texto"):



	Modo		
	Facet	•	Text facet
):00:00Z	Text filter		Numeric facet
D:00:00Z	Edit cells	•	Timeline facet
	Edit column	•	Scatterplot facet
0:00:00Z	Transpose	•	Custom text facet
	Sort		Custom numeric facet
	View	•	Customized facets
D:00:00Z	Reconcile	•	
0:00:00Z	- uanado	,	
D:00:00Z	Muerto Frente		
0:00:00Z	Muerto Frente		

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:

× Modo	change
6 choices Sort by: name count	Cluster
Desaparecdio 1	
Desaparecido 125	
Desparecido 1	
Fusilado 2189	
Miranda De Ebro 1	
Muerto Frente 6299	
(blank) 13	
Facet by choice counts	

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:

× Modo					
6 choices Sort by: name count					
Muerto Frente 6299					
Fusilado 2189					
Desaparecido 125					

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 "Desaparecdio", 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:

Muerte	Fecha	Modo
		Desaparedit

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

Desaparecido 120	
Desaparecdio 1	<u>edit</u> e

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

Desapa	recido	
Apply Enter	Cancel Esc	



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:

Google	e refine pers	onas_de			
Facet / Filter Undo / Redo 1					
Defreeb	Reco				
Google re	efine persona	s_desapare			
Facet / Filter	Undo / Redo 1				
	Extrac	t Apply			
Filter:					
0. Create project					
1. Mass edit 1 cells in column Modo					

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:

Modo
5 choices Sort by: nan
Desaparecido 126

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":

Nombre		Vecindad	💌 Lugar Mue
Facet	•	Donostia	Ordizia
Text filter		Bilbao	Otxandio
Edit colle	us	Zumaia	
Luit della	•	Hueto De Arriba	
Edit column	Split int	o several columns	
Transnosa			

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:

Split column Nombre into several columns	
How to Split Column by separator Separator , regular expression Split into columns at most (leave blank for no limit) by field lengths List of integers separated by commas, e.g., 5, 7, 15	After Splitting ✓ Guess cell type ✓ Remove this column
OK Cancel	

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 fr	om Freebase			
Reconcile	►	Rename this column				
Abaltua Ogalue						

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:

	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:

	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	Vec	indad	Lugar Muerte	Fecha		
Facet	•		Gasteiz	1936-12-01T00:00:00Z		
Text filter						
Edit cells	•					
Edit column	•	Split	into several column	IS		
Transpose	•	Add column based on this column				
Sort		Add o	column by fetching	URLs		
View	•	Add o	columns from Freet	base		
Reconcile	•	Rena	me this column			
		Rem	ove 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 cour	nt 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	A •
AL 11	LA



		Apellidos	Nombre	Vecindad	💌 Lugar Muerte	Fecha	Modo
\$ 5	7851.	Torres Bonet	Crescencio	Bilbo	Bilbo	1938-03-12T00:00:00Z	Fusilado
\$ 5	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:



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



Cluster & Edit	t column "Vec	indad"					
This feature help york" are very like	s you find groups of ely to refer to the s	of different cell values that r ame concept and just have	night be altern capitalization	ative representat differences, and	ions of the same thing. For exa "Gödel" and "Godel" probably	ample, the two strings "New York' refer to the same person. Find ou	and "new t more
Method key coll	ision 🛟	Keying F	unction finge	rprint 🛟		1 c	luster found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Valu	Ie		
2	9	 Gueñes (8 rows) Güeñes (1 rows) 		Gueñes			
Select All De	eselect All				Merge Selected & Re-Clu	ster Merge Selected & Clos	e 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	 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
Gueñes (8 rows)		Gueñes
 Güeñes (1 rows) 		



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:

	Cluster & Edit column "Ver
¢	This feature helps you find groups york" are very likely to refer to the
	Method nearest neighbor 🛟

Vemos inmediatamente que Refine encuentra ahora una serie de inconsistencias como "Legazpia" y "Legazpi" (que se diferencian en la 'a' final) o erratas como "Galllarta" (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	Castillo-Elejabeitia (1 rows)Castillo Elejabeitia (1 rows)		Castillo-Elejabeitia
2	17	Legazpia (11 rows)Legazpi (6 rows)	V	Legazpia
2	76	Gallarta (75 rows)Gallarta (1 rows)	V	Gallarta
2	14	Palencia (11 rows)Valencia (3 rows)		Palencia
2	8	Zigoitia (5 rows)Rigoitia (3 rows)		Zigoitia
2	2	 Errigoiti (1 rows) Errigoitia (1 rows) 	V	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:



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 <u>en mayor detalle los distintos algoritmos disponibles</u>.

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:

- <u>Guía del usuario: ¿qué es Google Refine?</u>
- <u>Documentación para usuarios</u>
- <u>Vídeos de ejemplo</u>