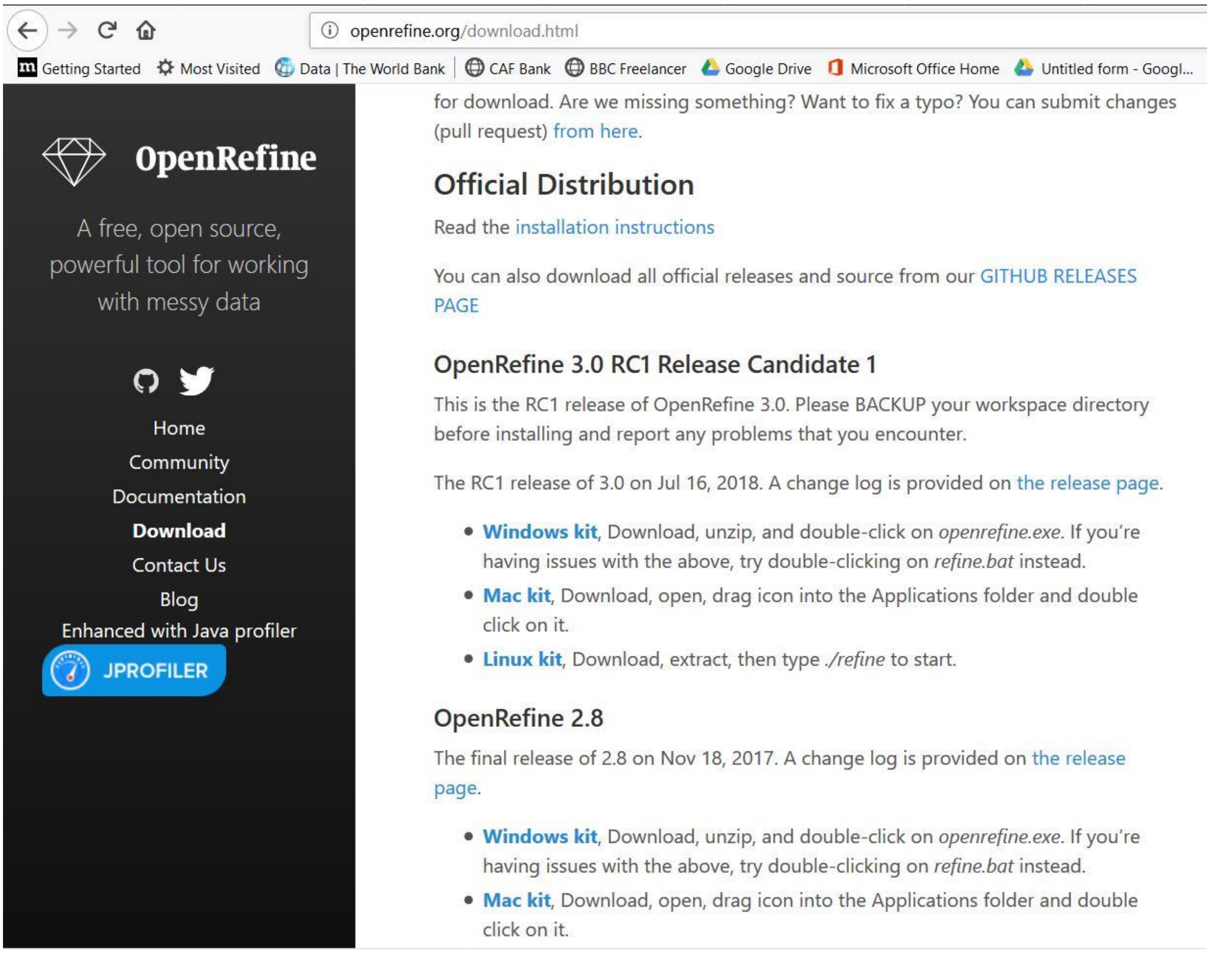
OPEN REFINE – CLEAN DATA PAINLESSLY

Open Refine is a powerful tool for cleaning large amounts of data quickly and efficiently. And it's free.

You download a copy from openrefine.org/download.html – at the time of writing the most recent stable version is 2.8



for download. Are we missing something? Want to fix a typo? You can submit changes (pull request) from here.

Official Distribution

Read the installation instructions

You can also download all official releases and source from our GITHUB RELEASES PAGE

OpenRefine 3.0 RC1 Release Candidate 1

This is the RC1 release of OpenRefine 3.0. Please BACKUP your workspace directory before installing and report any problems that you encounter.

The RC1 release of 3.0 on Jul 16, 2018. A change log is provided on the release page.

- **Windows kit**, Download, unzip, and double-click on *openrefine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type *./refine* to start.

OpenRefine 2.8

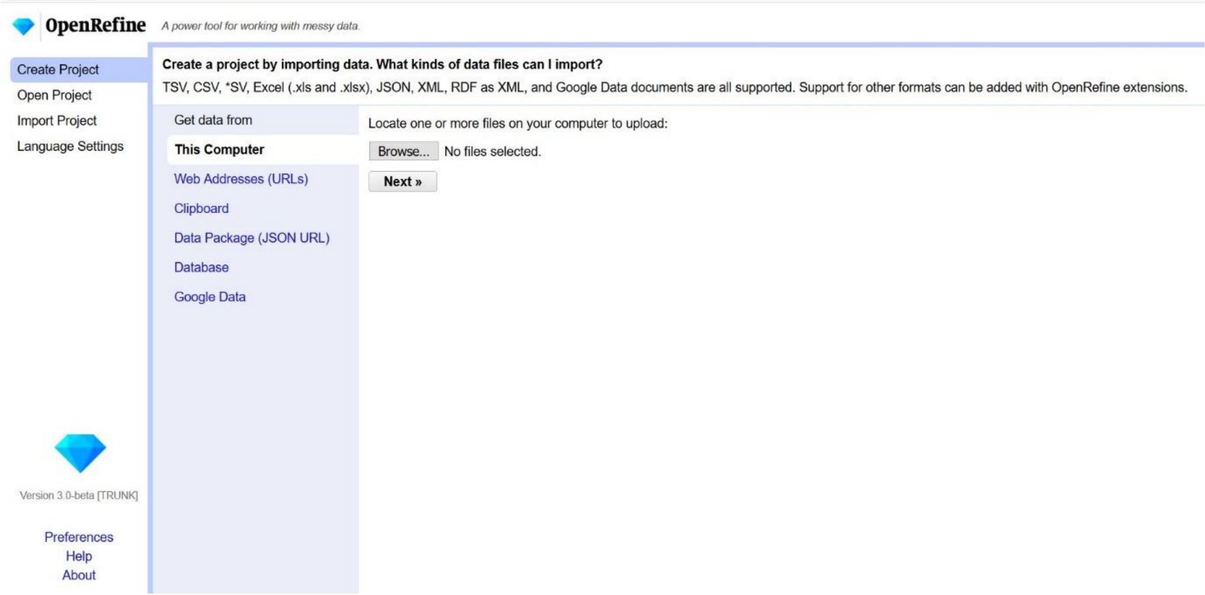
The final release of 2.8 on Nov 18, 2017. A change log is provided on the release page.

- **Windows kit**, Download, unzip, and double-click on *openrefine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.

Follow the instructions for installing onto your system – Mac/PC/Linux. As a program, OpenRefine runs within a browser window, and so will force your default browser to open (if it's already open, it will force a new tab to open). It works well with Firefox or Chrome, but not Microsoft Edge or Explorer.

When you run the program for the first time you may get error messages – your virus software may not like being forced to open a browser, or you may need to update your version of Java. Just do what it asks.

When you go to the new tab it will look like this



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 all supported. Support for other formats can be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

No files selected.

Version 3.0-beta [TRUNK]

The main feature is a fairly standard file choosing menu `>browse>select file > next`. But take a closer look when you have a moment sometime – there are different ways of importing data, and on the far left, there are options to create projects, open existing projects, or import projects.

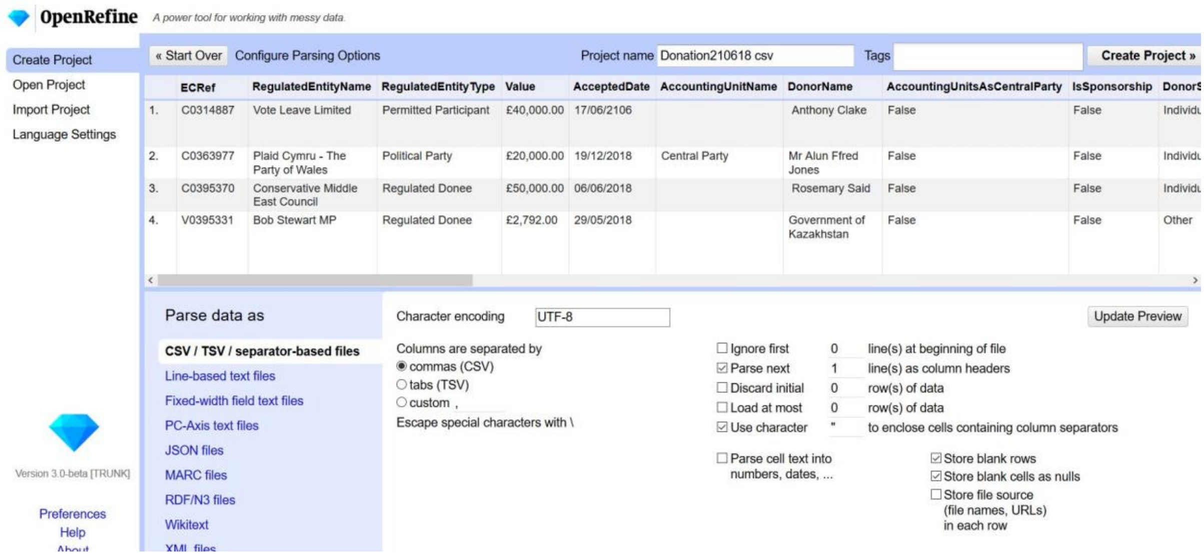
Every file you work on is seen as a “project”. OpenRefine does all its work on the file as a virtual copy, which you never actually open, but make edit decisions, which are recorded by the program, viewed within your browser. This means you never change the original data. It also means that when you have finished cleaning the dataset, and want to work on it back in Excel or R or whatever, then you will need to export it (in whatever format you need) – it then gets a new title, and you save it just as you would any other file you downloaded or imported to your hard drive.

By the way – the fact that you are using a web browser does not mean your data is on the internet: you don’t need a connection to the web, and your data remains on your hard drive, not in the cloud. It’s just Refine’s way of presenting the data to you.

Projects can be shared too. The 3rd option – `import project` – is very useful in a newsroom or team context. Projects can be exported from one person’s hard drive and imported by another – the 2nd user acquires the original file, and the editing record created by the 1st user, so when they import a project, they can see what their colleague has already done, and take on the file and all the editing decisions as if they had done it all along. Very powerful.

Back to creating a project (ie opening a dataset in OpenRefine for the first time):

The first thing you see after selecting your file, and clicking next is this preview



OpenRefine *A power tool for working with messy data.*

Create Project · Open Project · Import Project · Language Settings

« Start Over · Configure Parsing Options · Project name: Donation210618 csv · Tags: · Create Project »

	ECRef	RegulatedEntityName	RegulatedEntityType	Value	AcceptedDate	AccountingUnitName	DonorName	AccountingUnitsAsCentralParty	IsSponsorship	DonorS
1.	C0314887	Vote Leave Limited	Permitted Participant	£40,000.00	17/06/2106		Anthony Clake	False	False	Individu
2.	C0363977	Plaid Cymru - The Party of Wales	Political Party	£20,000.00	19/12/2018	Central Party	Mr Alun Ffred Jones	False	False	Individu
3.	C0395370	Conservative Middle East Council	Regulated Donee	£50,000.00	06/06/2018		Rosemary Said	False	False	Individu
4.	V0395331	Bob Stewart MP	Regulated Donee	£2,792.00	29/05/2018		Government of Kazakhstan	False	False	Other

Parse data as

CSV / TSV / separator-based files
Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
MARC files
RDF/N3 files
Wikitext
XML files

Character encoding: UTF-8 · Update Preview

Columns are separated by
● commas (CSV)
○ tabs (TSV)
○ custom ,
Escape special characters with \

□ Ignore first 0 line(s) at beginning of file
☑ Parse next 1 line(s) as column headers
□ Discard initial 0 row(s) of data
□ Load at most 0 row(s) of data
☑ Use character " to enclose cells containing column separators

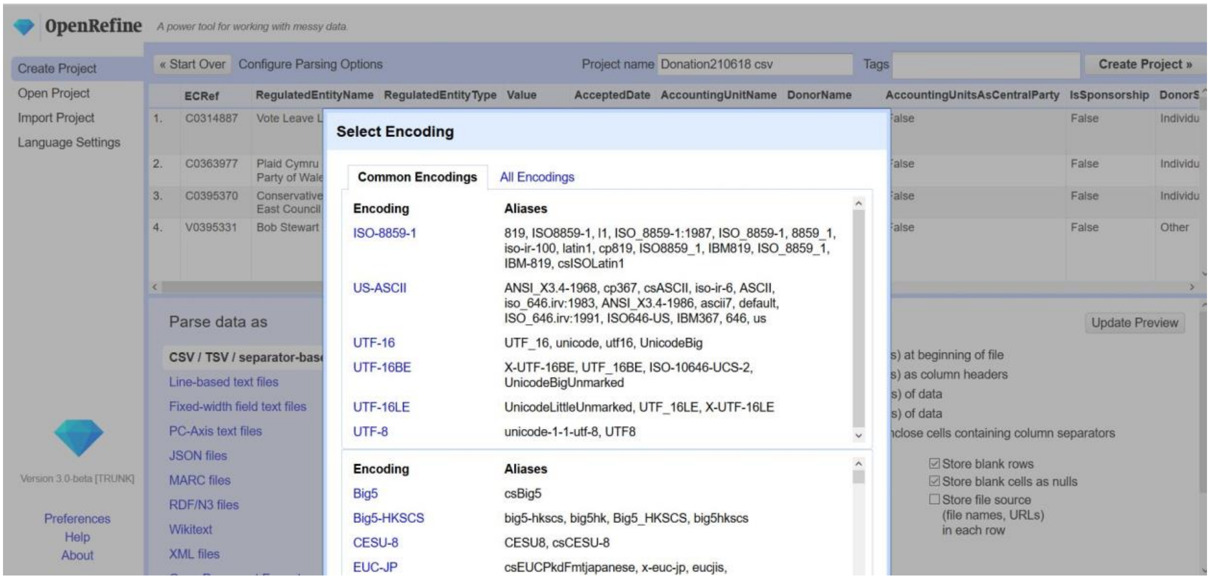
□ Parse cell text into numbers, dates, ...

☑ Store blank rows
☑ Store blank cells as nulls
□ Store file source (file names, URLs) in each row

Version 3.0-beta [TRUNK]

Preferences
Help
About

You have various options – OpenRefine usually (99.9% of the time) guesses what the file format is and opens it correctly. You can choose the formatting (in this case UTF-8 is the only one which reads the £ correctly – just click on the encoding to see the other options:



Select Encoding

Common Encodings · All Encodings

Encoding	Aliases
ISO-8859-1	819, ISO8859-1, l1, ISO_8859-1:1987, ISO_8859-1, 8859_1, iso-ir-100, latin1, cp819, ISO8859_1, IBM819, ISO_8859_1, IBM-819, csISOLatin1
US-ASCII	ANSI_X3.4-1968, cp367, csASCII, iso-ir-6, ASCII, iso_646.irv:1983, ANSI_X3.4-1986, ascii7, default, ISO_646.irv:1991, ISO646-US, IBM367, 646, us
UTF-16	UTF_16, unicode, utf16, UnicodeBig
UTF-16BE	X-UTF-16BE, UTF_16BE, ISO-10646-UCS-2, UnicodeBigUnmarked
UTF-16LE	UnicodeLittleUnmarked, UTF_16LE, X-UTF-16LE
UTF-8	unicode-1-1-utf-8, UTF8

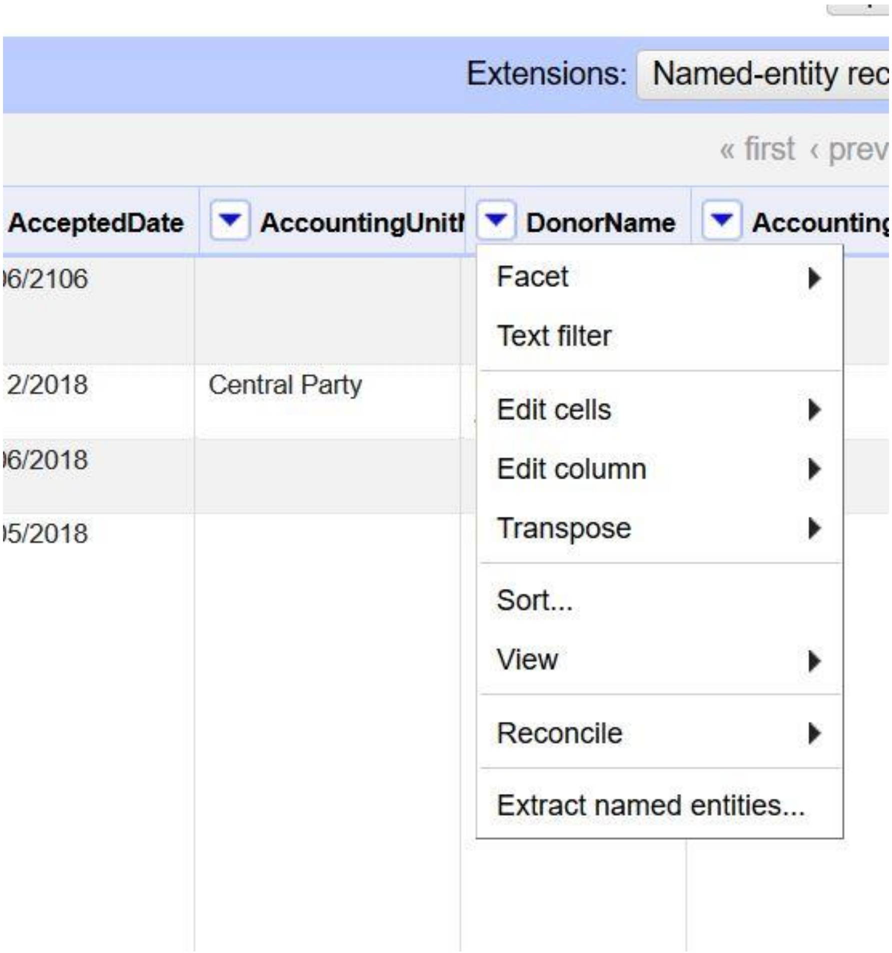
Encoding	Aliases
Big5	csBig5
Big5-HKSCS	big5-hkscs, big5hk, Big5_HKSCS, big5hkscs
CESU-8	CESU8, csCESU-8
EUC-JP	csEUCPkdFmtjapanese, x-euc-jp, eucjis,

You can also choose options to help deal with blank cells, or column names, or ignore lines with logos or other redundant data in them. Remember, you're not editing the original file, and you can go back to this step at any point and restart a project – either from scratch, or by using the undo function, which we will come to shortly.

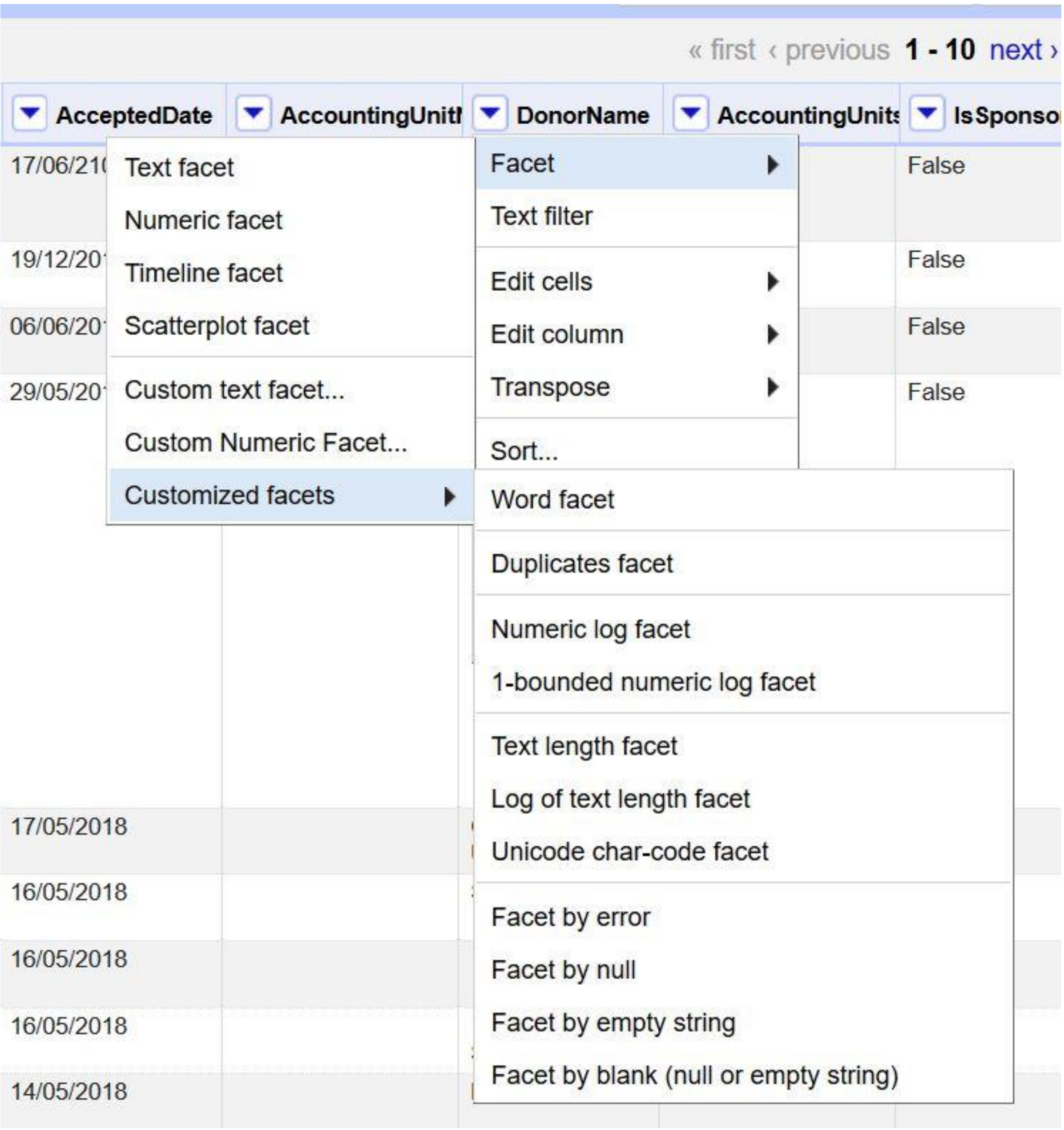
OpenRefine is very powerful and has many built-in functions which you may never use. When you have time, just play with it to see what happens. In this case we are going to use the most powerful function for datajournalists – cleaning one column of data by finding all the anomalies and getting names to be recorded in a uniform way.

As you will have seen, OpenRefine's logo and imagery is centred on a diamond, and the polishing of rough diamonds into perfect jewels. So we are going to use a function called "faceting" (polishing one side or facet of a diamond) and we are going to polish the text.

Every column has a dropdown arrow selection tool. We are working on the "Donor Name" column:

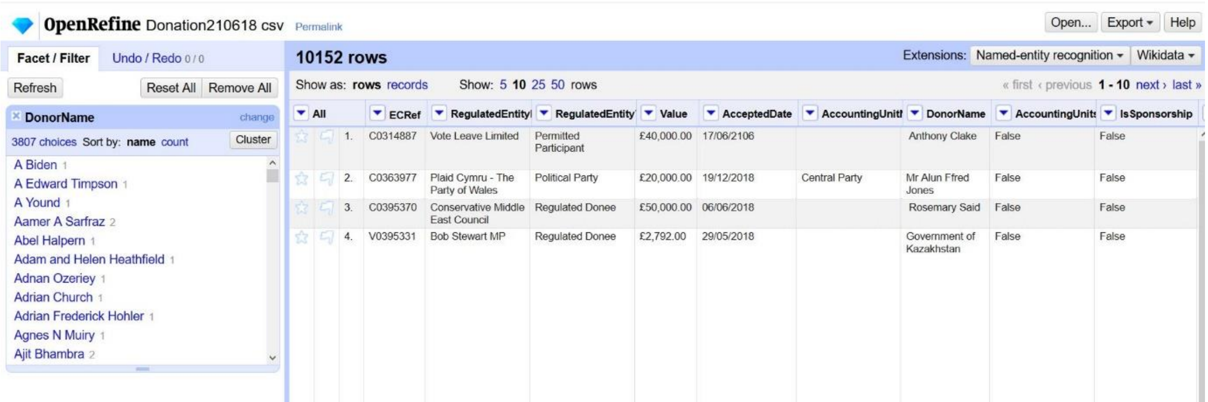


We choose "Facet" which brings up the next set of options:



We choose “Text Facet” – but do come back and take a look at some of the others. There are many options which may come in handy one day. Documentation is pretty good, but OpenRefine is a unique search string and you will be able to find a lot of tutorials and solutions by googling these options (including looking for them on youtube).

Choosing “text facet” makes OpenRefine do some work, and a few seconds later your screen will look like this – with a new box on the left, taking the name of the column you chose “Donor Name”. It shows the names once each, in alphabetical order. You can also click on “count” and see them in order of frequency (slightly faster than pivoting them in a spreadsheet if you just want to see whose name pops up the most).



OpenRefine Donation210618 csv Permalink

10152 rows

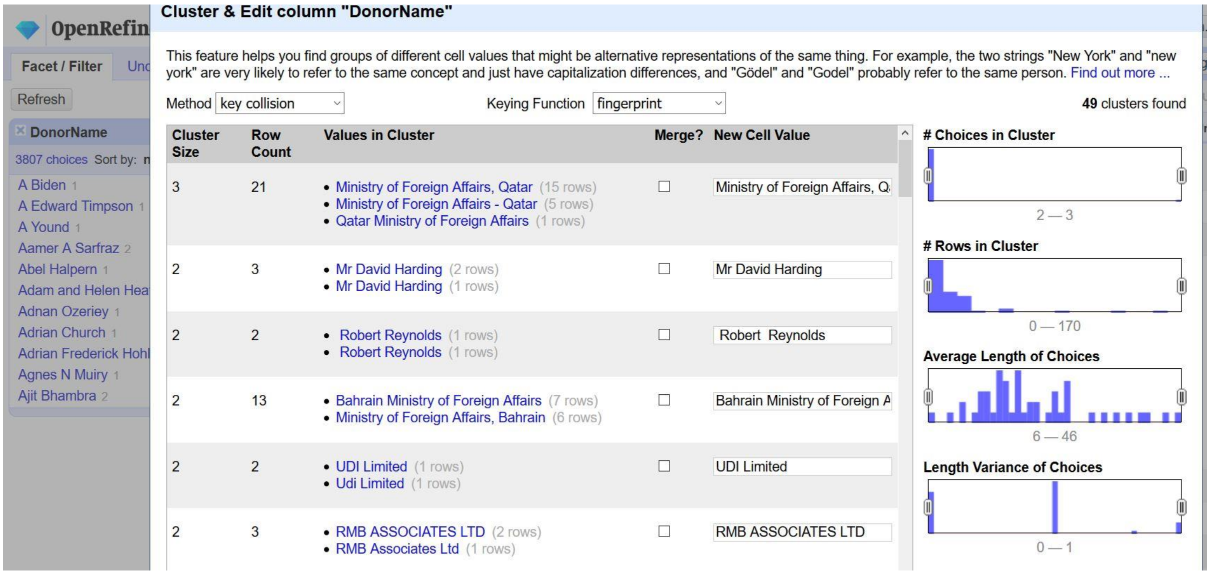
Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 10 next › last »

DonorName — 3807 choices Sort by: name count — Cluster

- A Biden 1
- A Edward Timpson 1
- A Yound 1
- Aamer A Sarfraz 2
- Abel Halpern 1
- Adam and Helen Heathfield 1
- Adnan Ozeriey 1
- Adrian Church 1
- Adrian Frederick Hohler 1
- Agnes N Muiry 1
- Ajit Bhambra 2

All	ECRef	RegulatedEntityI	RegulatedEntity	Value	AcceptedDate	AccountingUnitI	DonorName	AccountingUnits	IsSponsorship
1.	C0314887	Vote Leave Limited	Permitted Participant	£40,000.00	17/06/2106		Anthony Clake	False	False
2.	C0363977	Plaid Cymru - The Party of Wales	Political Party	£20,000.00	19/12/2018	Central Party	Mr Alun Ffred Jones	False	False
3.	C0395370	Conservative Middle East Council	Regulated Donee	£50,000.00	06/06/2018		Rosemary Said	False	False
4.	V0395331	Bob Stewart MP	Regulated Donee	£2,792.00	29/05/2018		Government of Kazakhstan	False	False

To see the real power of faceting, click on the "cluster" button on the right of the facet box. Now you see all the ways in which OpenRefine identifies entries which could be the same – but are entered differently ("dirty data") – the first one in this example shows the power of this tool – not only does it spot that "Ministry of Foreign Affairs, Qatar" and "Ministry of Foreign Affairs – Qatar" are probably the same, but it even finds another less obvious variant – "Qatar Ministry of Foreign Affairs" which is yet another way of saying the same thing.



Cluster & Edit column "DonorName"

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 ...

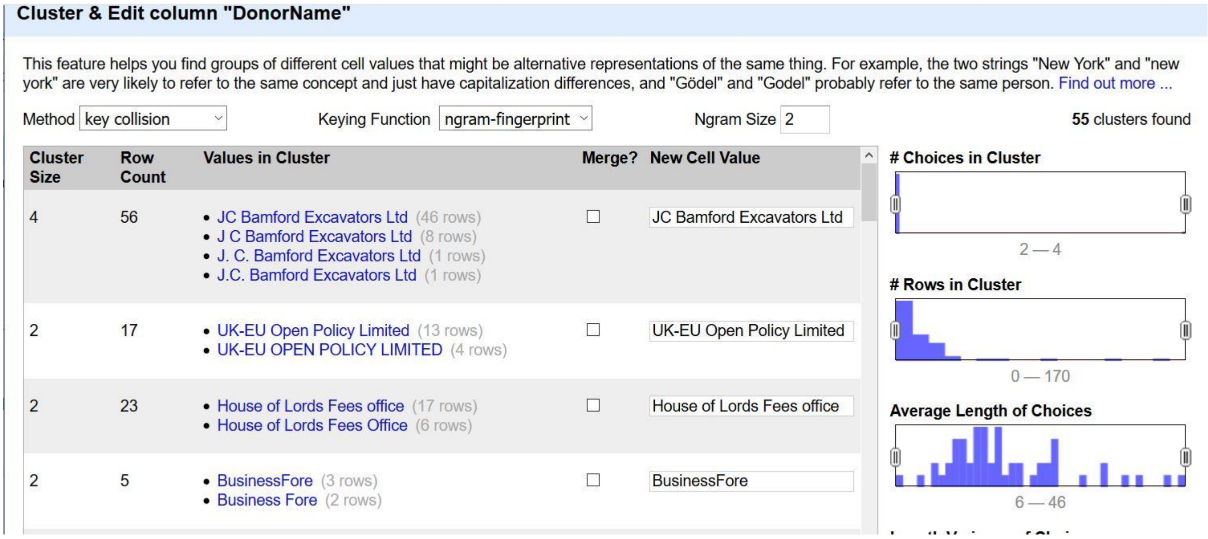
Method key collision   Keying Function fingerprint   49 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	21	• Ministry of Foreign Affairs, Qatar (15 rows) • Ministry of Foreign Affairs - Qatar (5 rows) • Qatar Ministry of Foreign Affairs (1 rows)	☐	Ministry of Foreign Affairs, Q
2	3	• Mr David Harding (2 rows) • Mr David Harding (1 rows)	☐	Mr David Harding
2	2	• Robert Reynolds (1 rows) • Robert Reynolds (1 rows)	☐	Robert Reynolds
2	13	• Bahrain Ministry of Foreign Affairs (7 rows) • Ministry of Foreign Affairs, Bahrain (6 rows)	☐	Bahrain Ministry of Foreign A
2	2	• UDI Limited (1 rows) • Udi Limited (1 rows)	☐	UDI Limited
2	3	• RMB ASSOCIATES LTD (2 rows) • RMB Associates Ltd (1 rows)	☐	RMB ASSOCIATES LTD

Just above the table showing the clustered entries you will see two dropdown selectors – "Method" and "Keying function". These are the different ways – 6 in all – which OpenRefine assesses and groups possible errors in data entry which you can then use your own knowledge of the subject area to decide which entries to merge, and under which version to merge them (if you don't like the options you can just type your version into the box under the heading "New Cell Value".

Under the method "key collision" there are 4 different keying functions. Work through them in the order they are listed. The next one is "ngram-fingerprint" – which brings up a whole different set of options. As before you work through them, selecting or ignoring them depending on the accuracy of OpenRefine's assessment. When you're happy with a set of selections, click "Merge selected and re-cluster" and wait for OpenRefine to do its magic. Then carry on, either with a new batch, or a new keying function. (I tend to do about 10-15 mergers in one go, as the actual edit can take several minutes. I did once crash OpenRefine by doing too many “mass edits” – as they’re called – in one go!)

Take a moment to look at how these groups of results differ slightly from each other. Sometimes you can’t see the difference, and it’s usually down to an extra space somewhere in the string – white space is “seen” differently by the computer, and would become a different entry in a pivot table, for example.



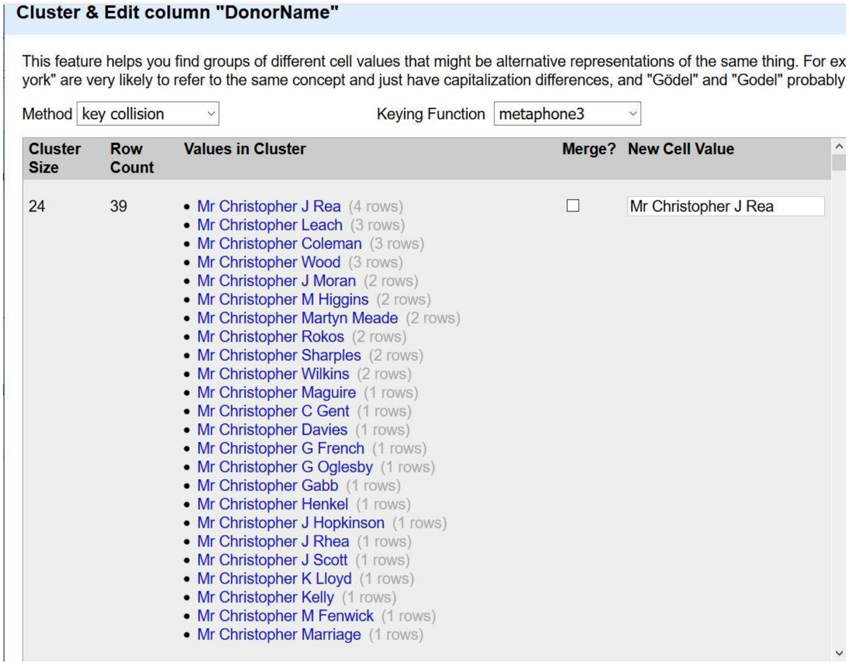
Cluster & Edit column "DonorName"

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 ngram-fingerprint Ngram Size 2 **55** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	56	• JC Bamford Excavators Ltd (46 rows) • J C Bamford Excavators Ltd (8 rows) • J. C. Bamford Excavators Ltd (1 rows) • J.C. Bamford Excavators Ltd (1 rows)	☐	JC Bamford Excavators Ltd
2	17	• UK-EU Open Policy Limited (13 rows) • UK-EU OPEN POLICY LIMITED (4 rows)	☐	UK-EU Open Policy Limited
2	23	• House of Lords Fees office (17 rows) • House of Lords Fees Office (6 rows)	☐	House of Lords Fees office
2	5	• BusinessFore (3 rows) • Business Fore (2 rows)	☐	BusinessFore

As you work through the different options, merging as you go, the number of options offered for merging in each cluster goes down – it’s worth doing the clustering in the order suggested by the drop down options. I often find that “metaphone3” is the least useful analysis – OpenRefine sees all entries beginning with the same words: “Mr Christopher”, for example, as being potentially the same as each other. Obviously this is nonsense. But it’s worth looking through all the results just in case there is something you wouldn’t otherwise have seen.



Cluster & Edit column "DonorName"

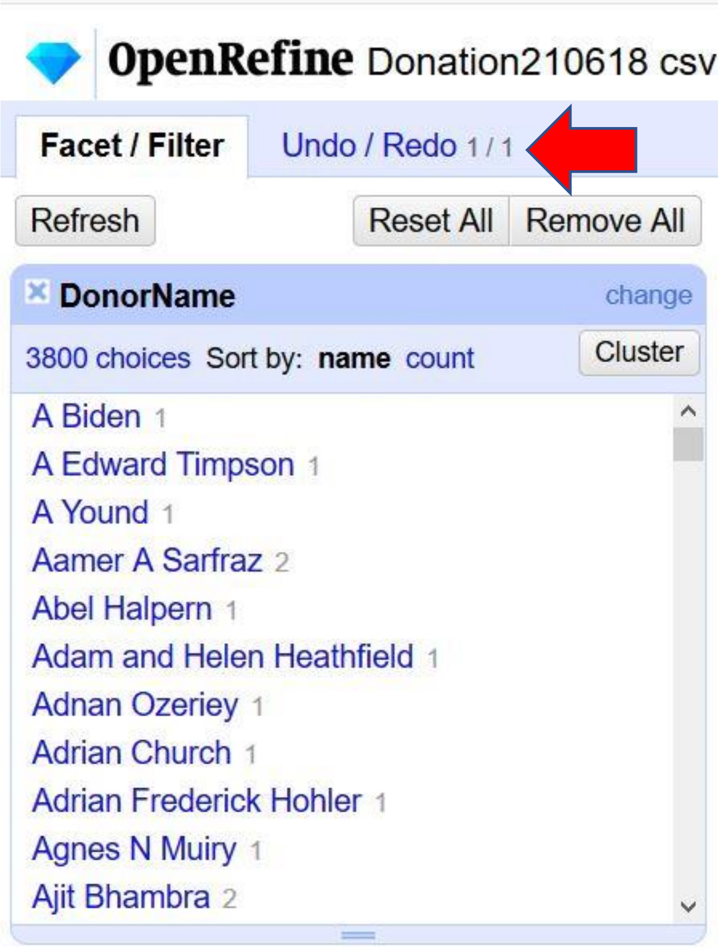
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For ex york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably

Method key collision Keying Function metaphone3

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
24	39	• Mr Christopher J Rea (4 rows) • Mr Christopher Leach (3 rows) • Mr Christopher Coleman (3 rows) • Mr Christopher Wood (3 rows) • Mr Christopher J Moran (2 rows) • Mr Christopher M Higgins (2 rows) • Mr Christopher Martyn Meade (2 rows) • Mr Christopher Rokos (2 rows) • Mr Christopher Sharples (2 rows) • Mr Christopher Wilkins (2 rows) • Mr Christopher Maguire (1 rows) • Mr Christopher C Gent (1 rows) • Mr Christopher Davies (1 rows) • Mr Christopher G French (1 rows) • Mr Christopher G Oglesby (1 rows) • Mr Christopher Gabb (1 rows) • Mr Christopher Henkel (1 rows) • Mr Christopher J Hopkinson (1 rows) • Mr Christopher J Rhea (1 rows) • Mr Christopher J Scott (1 rows) • Mr Christopher K Lloyd (1 rows) • Mr Christopher Kelly (1 rows) • Mr Christopher M Fenwick (1 rows) • Mr Christopher Marriage (1 rows)	☐	Mr Christopher J Rea

When you have done all the merging and clustering you can (or that you can stand – it can make your eyeballs burn after a while), close the clustering tool.

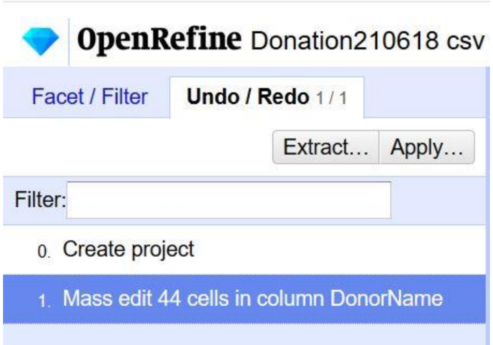
If you, or a colleague, wants to revise what you have done at any point, you can click on "undo/redo" at the top left of your screen



3800 choices Sort by: **name** count

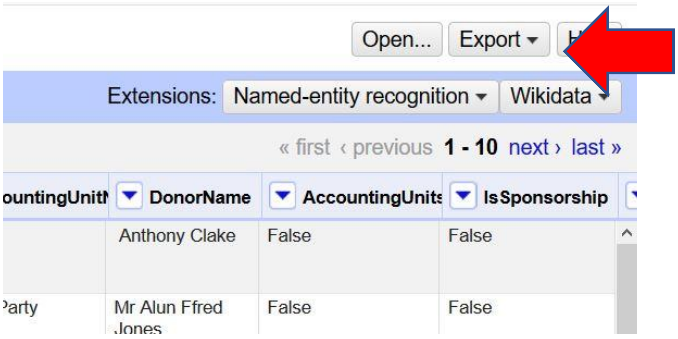
A Biden 1
A Edward Timpson 1
A Yound 1
Aamer A Sarfraz 2
Abel Halpern 1
Adam and Helen Heathfield 1
Adnan Ozeriey 1
Adrian Church 1
Adrian Frederick Hohler 1
Agnes N Muiry 1
Ajit Bhambra 2

This brings up a list in chronological order of every edit you have made – in this case, I have done two things – opened the project, and then a mass edit of 44 cells in one column. If I decide I did the wrong thing, or perhaps new information comes in, instead of starting all over again, I could just go back one step – in this case, if I click on "0. Create Project" I will go back to the way the dataset looked at that point, and either export it, or re-cluster the columns and edit it differently.

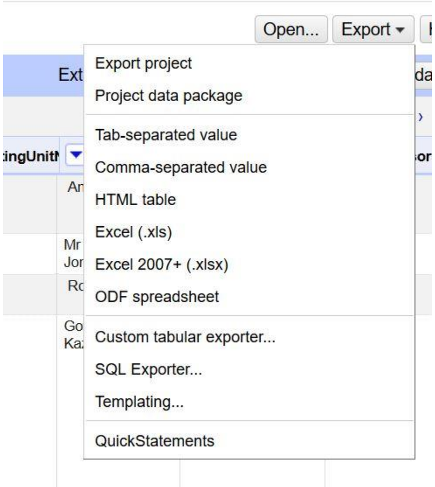


0. Create project
1. Mass edit 44 cells in column DonorName

(At that point, I can't unfortunately create two versions of history – my new set of edits will overwrite my previous set. If I want to preserve what I have done before I start re-editing it, then I can export the dataset – see all the various options.



ountingUnit	DonorName	AccountingUnits	IsSponsorship
	Anthony Clake	False	False
Party	Mr Alun Ffred Jones	False	False



Whatever I do at this point, one of the great strengths of OpenRefine is that what I have in my browser tab now will be there next time I reopen that project. It saves everything as I go along, and it records the edits I make forever.