An Introduction to data cleaning with Open Refine

A School of Data module created by Michael Bauer, Sam Leon, Anders Pedersen and the School of Data team.
Level: Intermediate

Introduction

Data, especially when collected by humans (or released by government agencies), tends to be messy. Things are not in the formats you need - often formatted for human readability, which can be a tough job to understand for computers. You can spend hours in spreadsheets trying to bring the data you have into the right form.

These inconsistencies and mistakes can wreak havoc when you come to analyse your data. Simple things like filtering, sorting and pivot tables will not work properly when you have multiple spellings of one name.

The process of correcting these mistakes or inconsistencies - a process known as data cleaning - can be very time consuming involving editing cell-by-cell. Fortunately there's a nice tool called Open Refine that can help by automating a lot of these mind numbing tasks so you can focus more of your energy on actually analysing your data.

The Web URL where the tool can be download for free is http://openrefine.org/

About Open Refine

OpenRefine is a program designed specifically to work with messy data and to clean nasty datasets.

For this module you will need:
OpenRefine: Download and Install from http://openrefine.org

Exercise data (some examples):

- The Dataset on Kenyan Tenders from https://scraperwiki.com/dataset/gjzj35i or
- Alternative: A list of campaign contributions from Nassau police force: https://docs.google.com/a/okfn.org/file/d/0B_dkMlz2NopELVRLcEp2ZDhLTlU/edit

Fields and columns where your data is often messy

First, have a look at the data you are working on - open it in a spreadsheet or use the datasets linked above.

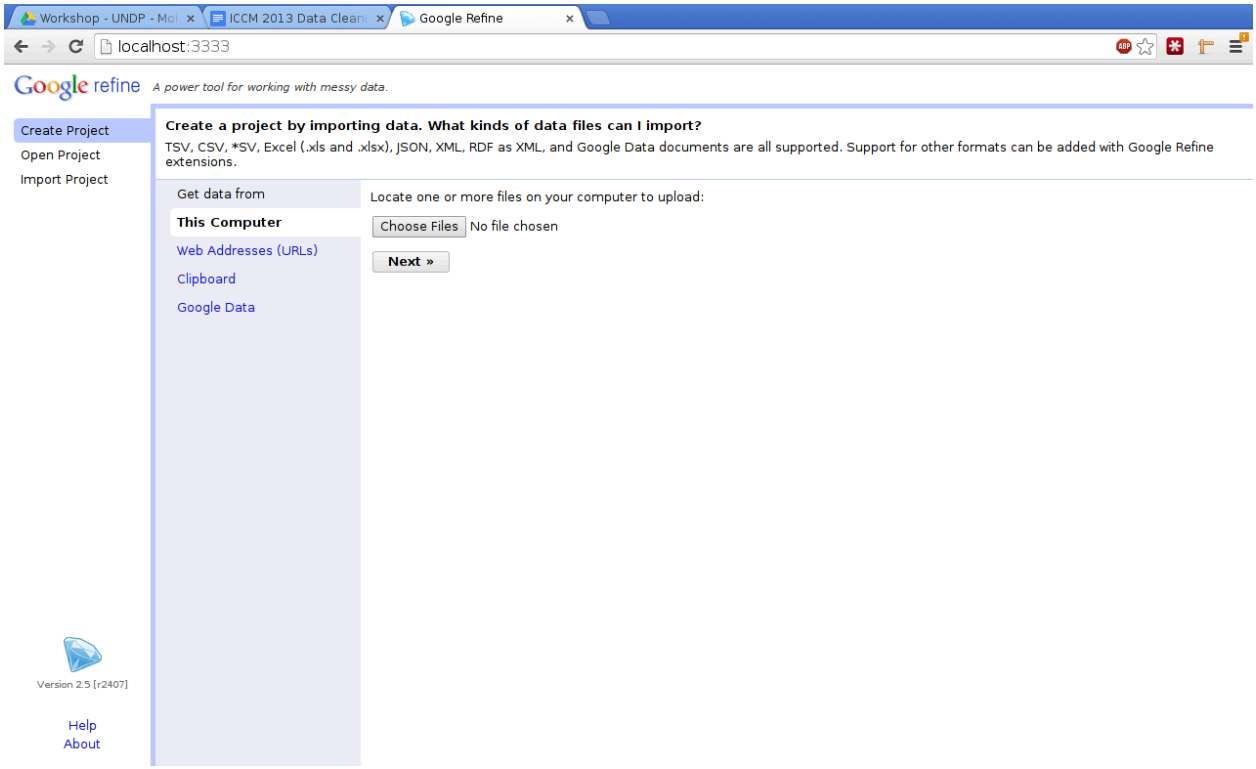
Datasets will often have certain very specific features, for example:

- *Amounts* might be written with the currency in front and formatted to be readable for humans - a spreadsheet will not be able to sum those up easily.
- *Dates* might be in special formats which are easy to read for humans but a hard task for a computer.
- *Names* are often spelled in various ways - even country or common company names can be misspelled. You will note with “human entered data” or surveys with open ended questions that some names are written differently, although they clearly refer to the same thing (eg. country, company, organisation)

So let’s go and clean this up!

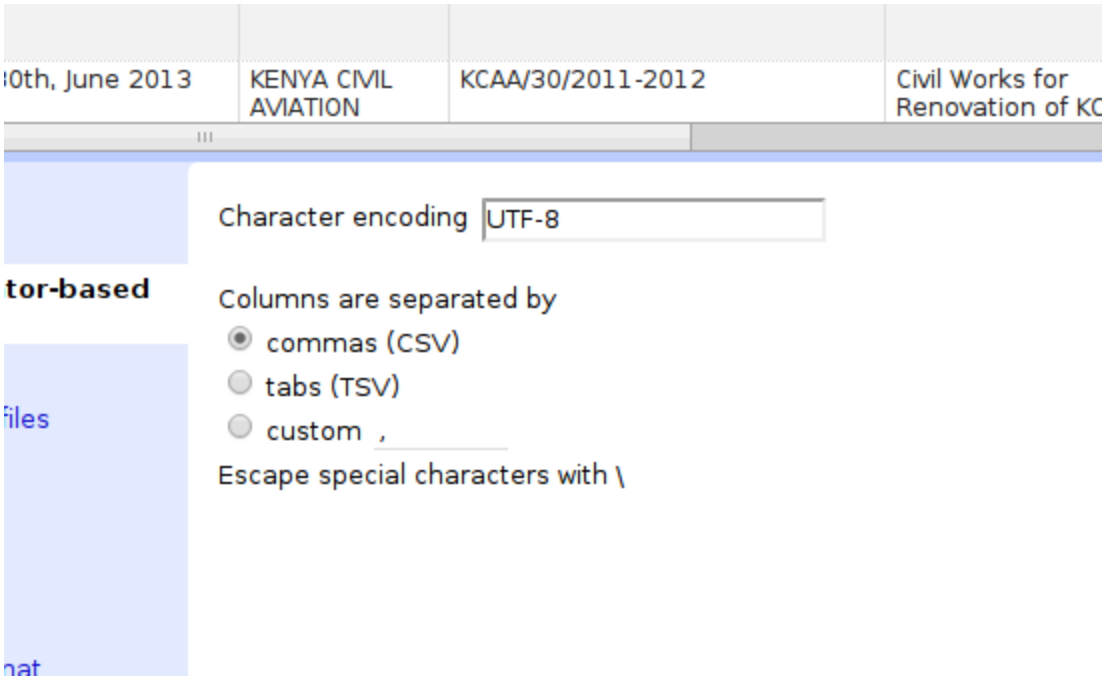
Starting a new project

Start google Refine by double clicking on the icon - after a short while a browser window will open that looks similar to this:



Next, let's create a new project. Click on the Create Project Tab on the left, select this computer and upload the file you have downloaded by clicking "Next"

This will open the Preview mode - in this mode Refine will show you how it would interpret the data you've uploaded. Make sure it's set to CSV. To include eventual special characters select UTF-8 as character encoding (this can vary on your datasource, try some of the popular ones, if UTF-8 does not work).



When you are done: Click on “Create Project” up right

You will now be in the project view - while you interact with Refine through your webbrowser the data stays on your computer - **so you can use Refine with sensitive data**.

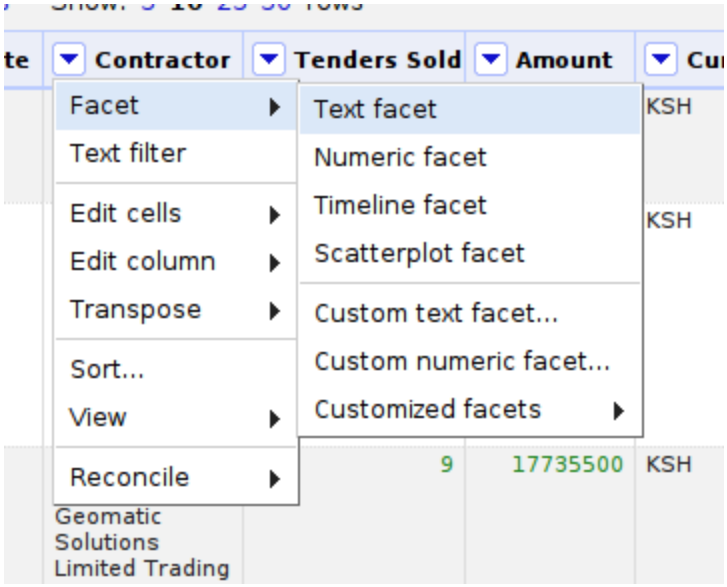
The project view looks quite similar to a spreadsheet - it has rows and columns. Because most of the changes you’ll have to do in datasets have to be applied to the whole column, this is the main mode you’ll interact with your data.

Facet: Remove multiple spellings of same name

One of the most commons problems with data is that the same entities - whether they’re persons, companies, or places - are referred to in many different ways. Addresses are common culprits as there is plenty of scope for inconsistently referring to place. Consider “1 Portsoken Street, Tower Hill, London, E1 8RB” or “ “Lloyds Chambers, 1 Portsoken Street, London, E1 8RB, UK” - same as far as you’re concerned, but very different for the computer.

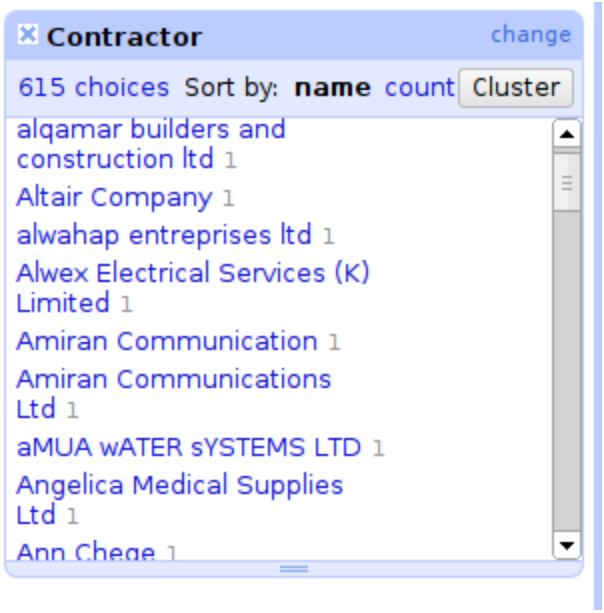
Refine helps you work with these common inconsistencies and has a function that will **cluster names** it thinks are similar. Facets are refines way of filtering data, but they are much more powerful. From the top of the column go to Facet → Text Facet. The left hand side of your screen will then give you a view of your data much like a pivot table. In the right corner a button “Cluster” will give you the opportunity to merge names that Refine thinks are similar.

To get started click on “Facet - text facet” in the column options for the column contractor

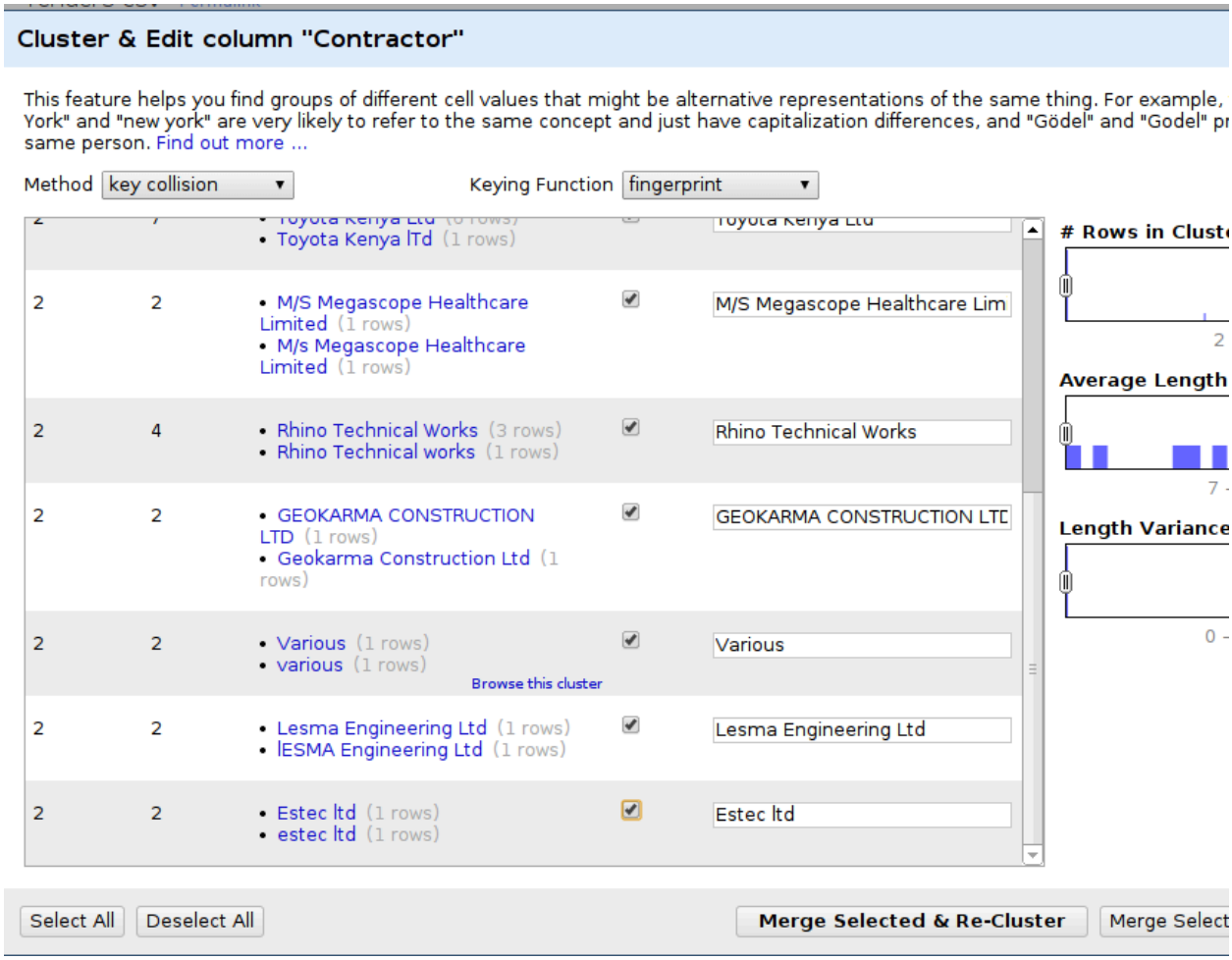


This will open a text facet on the left side, there you can go ahead and look at the different contractors.

There seem to be some companies which are similar but appear written differently - let's reconcile them. Refine allows us to do this with the "Cluster" button up right.



If you click on this button - refine will try to find matching pairs using different methods. It found some with the first method. If you agree with what Refine would match up (and how) you can click on the small checkbox.



If you're done click on Merge selected and re-cluster. Now go ahead and change the election methods and keying functions (drop downs on the top) - try to find all matches.

When you're done look at the list again. Sometimes there are names, that do not match up and you'll have to do this manually. See for example this list of countries below where "Republic of Armenia" and "Armenia" will need to be manually reconciled.

Tip: There are a number of algorithms (Keying Functions) for the original clustering, so have a crack at different ones to see what Refine will find:.

Remove weird white spacing

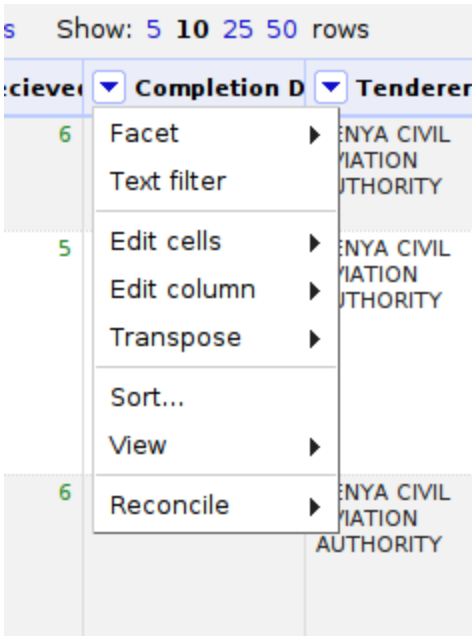
White spaces in odd places can be difficult to spot, but their existence means that your machine will treat names and numbers that look the same to you as completely different things. “Sam Leon” is some other person to “Sam Leon” in Excel. So doing basic analysis - including sorting, and filtering - on sets of names with inconsistent white spacing will be a nightmare.

Refine has some handy features for removing all kinds of of nasty spacing from all your records: trailing (removes spaces after the value), leading (removes spaces before the value), consecutive (removes double spacing). You can access these all from the button on the header of your column under Edit Cells → Common Transforms. Excel has some similar functionality, with it’s CLEAN and TRIM functions, but as Refine lets you transform whole columns at the click of a button, it’s often a better choice for this kind of task.

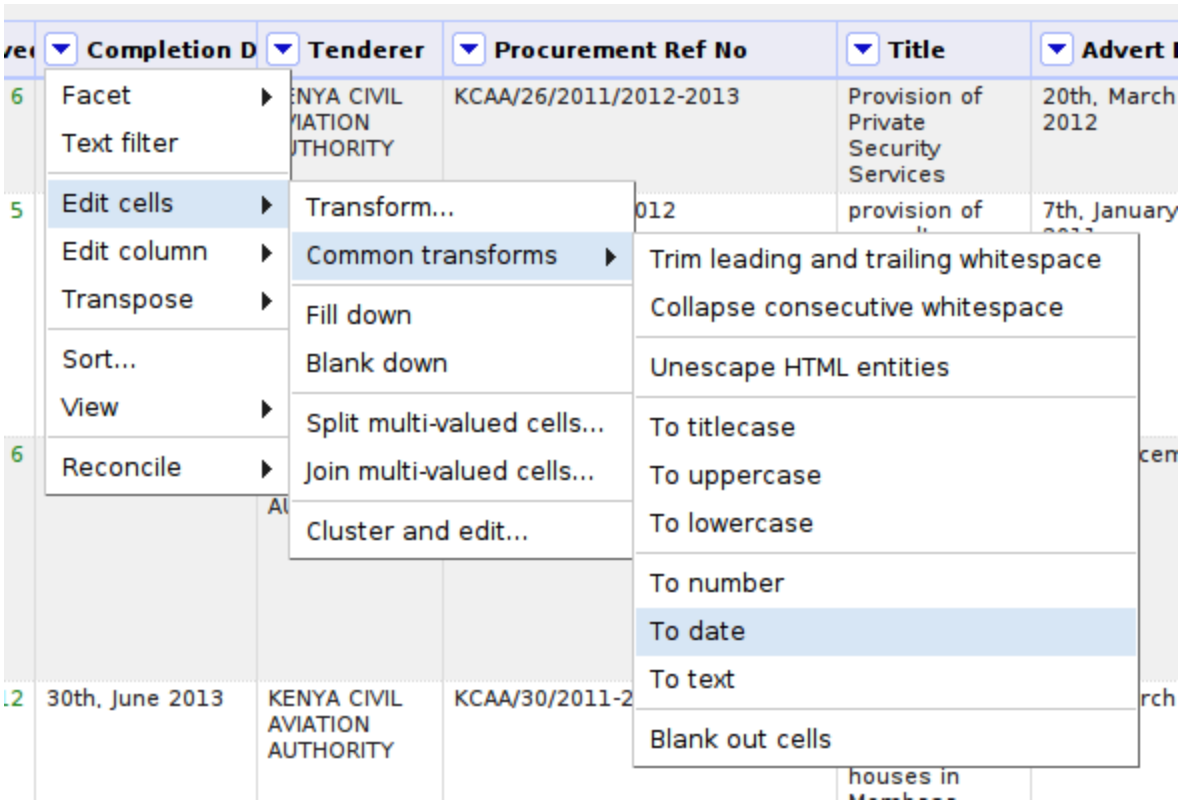
Formatting dates

Let’s try and see how we can fix dates with messy formats and have Refine convert them to something a computer will understand. This is a quite frequent problem (eg. for government entered data or in surveys with open ended questions), and thus it’s built into Refine.

Click on the blue triangle next to the column you wish to fix - this will open the column options.



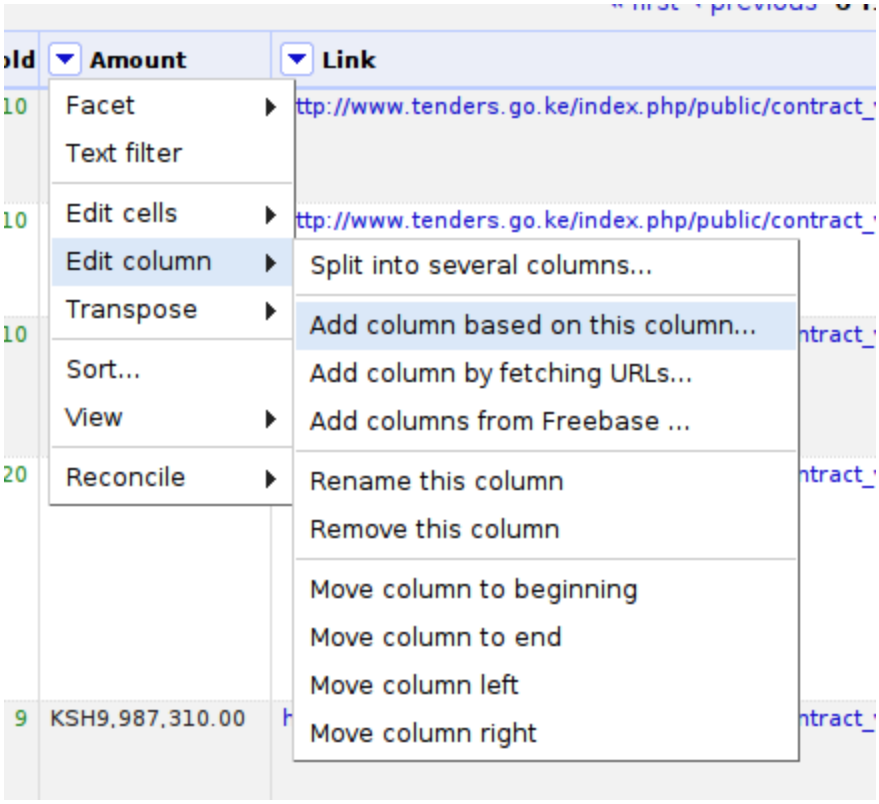
Then select “Edit Cells → Common Transforms → To date”



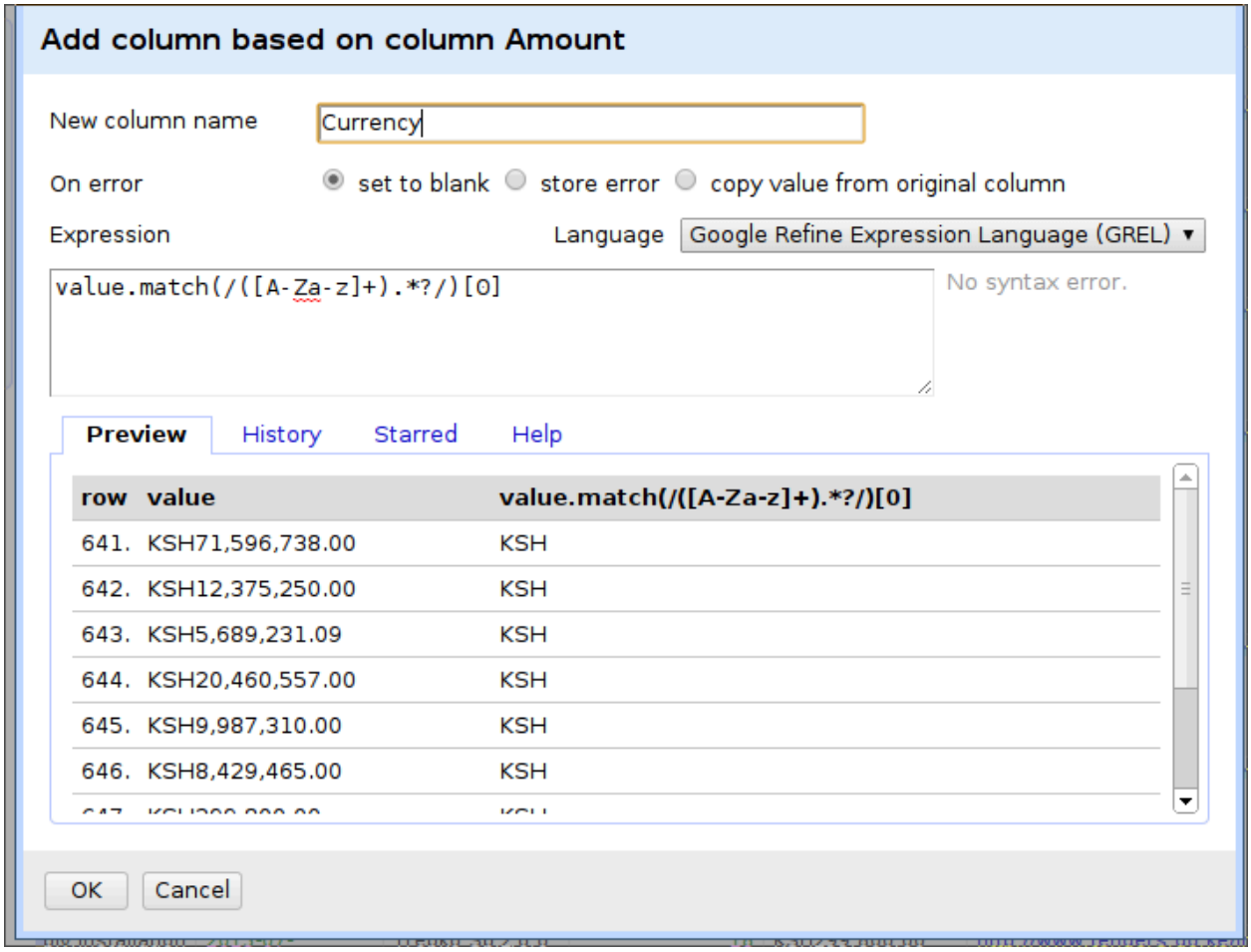
Great - this converted the dates into dates the computer can understand.

Next let's try and see how we can fix the Amounts - first:
We'll need to separate the currency out into a column on it's own. (The amounts here seem all to be in KSH, but you never know).

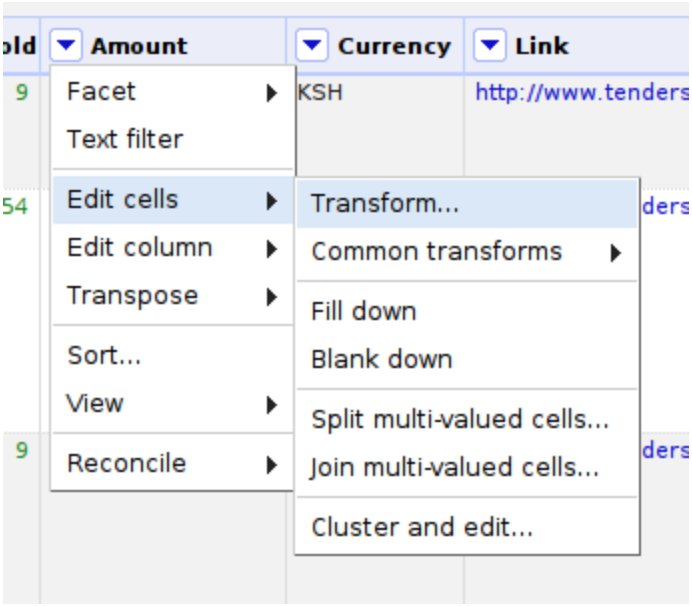
Click on the column options for the Amount column and select "edit column - add column based on this column"



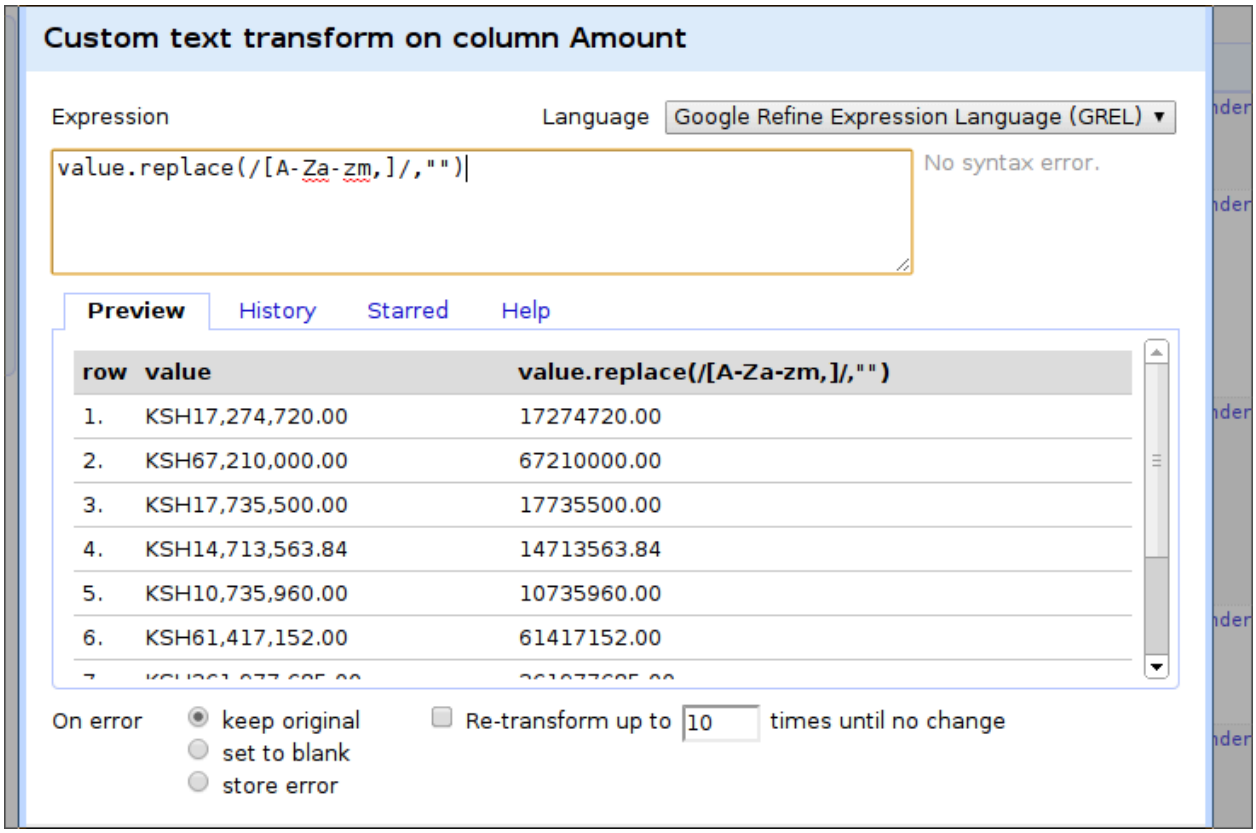
A new window will appear, asking you for an expression for the contents of the new column. We want any letters that are up front to go there. For this we'll use the expression 'value.match(/([A-Za-z]+).*?/)[0]' - also fill in a name for the column



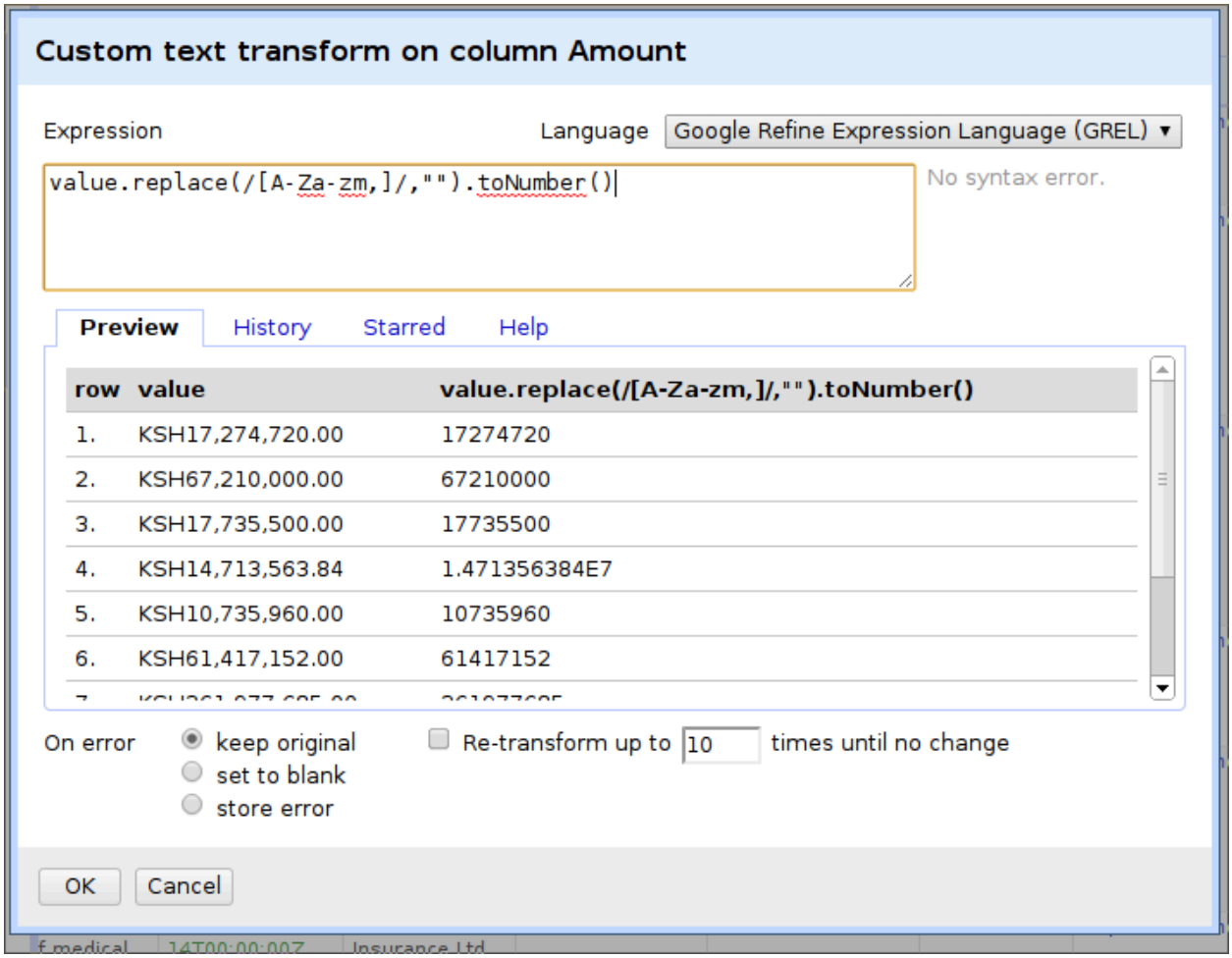
Great - now let's remove all letters and commas from the values. In the column options select: "edit cells - transform"



We need to enter an expression here as well. Using the expression: value.replace(/[A-Za-z,]/,"") we'll remove all the letters and commas.



We also want to convert the results to a number: we'll use .toNumber() to do so. - simply append it to the previous formula then click "OK"



Congratulations - you just cleaned your first Dataset using refine!

Example: Formatting years

When people conduct data entries or surveys they might end up allowing for non-formatted open-ended string responses for something which should actually be a numeric column (response).

See for example this dataset from a survey done with an educational programme:

126 rows

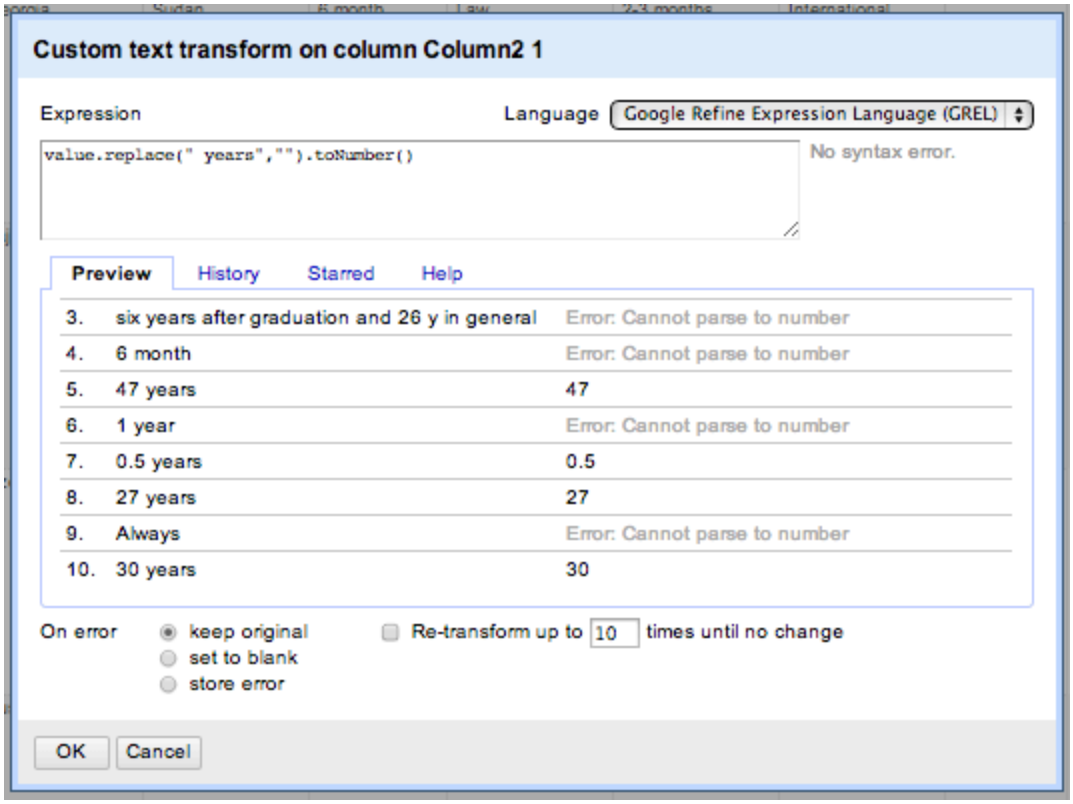
Show as: **rows** records Show: 5 **10** 25 50 rows

at year di	What is your g	What is your h	Current countr	Years	What was your	Upon your retu	What type of or	Column3
2002	Female	Georgia	Sudan	6 month	Law	2-3 months	International organization or donor	
1994	Male	Tajikistan	Tajikistan	47 years	Public Policy	2-3 months	Not applicable/not currently working	Currently I'm freelancing
2000	Male	Azerbaijan	Switzerland	1 year	Public Policy	2-3 months	International organization or donor	
1999	Female	Russia	Portugal	0.5 years	Public Health	0-1 month	International organization or donor	

Notice the *Years* column: It would clearly be useful to have this as numeric value, but currently letters are cluttering the column and making us unable to conduct any analytics on this.

In order to remove *Years* go to the Column tab → Edit Cells → Transform

Then insert: “value.replace(" years","").toNumber()”



I can now follow up with other adjustments to the same expression, for example if I wish to remove "year" from the column as well. Simply adjust the text inside the expression:

From:
"value.replace(" years","").toNumber()"

To:
"value.replace(" year","").toNumber()"

Annex: Reconciling company names to the OpenCorporates database

Not strictly data cleaning, but still hugely useful. Reconciling allows you to align the company names in any given dataset with those in the global open company database OpenCorporates (if they exist). Making this association then enables you to automatically import other information held by OpenCorporates about these companies into other columns of your dataset - giving you a richer dataset that would have taken hours to build manually. This can be very useful when working with lists of companies that would take hours to search for manually. For a full guide see this great tutorial.

Resources

- Full list of tutorials and resources maintained by the Open Refine community

- Reconciliation tutorial from OpenCorporates