## Introduction to Open Refine

### Goals

* Understand what is Open Refine
* Learn some basic functionalities
* Know where to look for help

### Links

Official screencast tutorials <https://github.com/OpenRefine/OpenRefine/wiki/Screencasts>

### Script

[Note: This tutorial runs on the alpha version of Open Refine 2.6. By the time of the course, stable version 2.6 should have been released]

[EXCEL ]

OpenRefine is a powerful tool to clean messy data and analyze it rapidly. Unlike a spreadsheet software, it is not made to input data manually. It is a tool that you use when you have pre existing data that you need to treat before use. Think of data that you obtained from a public organization or from a form on your website. This is the kind of data that a datajournalist works with, and that’s why Open Refine is especially well liked by the datajournalism community.

OpenRefine started as a tool under the name Freebase Gridworks, then Google Refine when Freebase was bought back by Google. The tool is now open source and spun out from Google, hence the new name. But the product remained mostly the same.

If you don’t have it already, you can download it from <http://openrefine.org/> under the download section. It can be run on Windows and Mac without admin priviledges, but you need to have Java installed. If Refine doesn’t start, check that you installed it into your Application folder on Mac and that you took it out from the zip archive on Windows.

Once it is running, Refine opens a new window in your browser. It’s important that you do not close the DOS-like window on Windows or the diamond icon on Mac. That’s actually the program working in the background, and you browser is just interacting with this program. One very important thing about Open Refine is that it runs only on your computer. You don’t need to have an internet access to make it run, you can use it offline. That’s very different from Google Spreadsheet, for instance, because it means that you can use it for data that might be too sensitive to be put online, provided you secure your computer well enough.

For this very simple tutorial, I’ll use a file containing the height of 40 people in CSV format. Open it on the main screen of Refine and click next. The preview screen you have here is very important. Refine actually shows you only the first 100 lines of the file, so as not to overwork your computer, in case you’re trying to open a file with one million rows. Here, we have 3 columns with the gender, the height and the place of birth of each person in the sample. There, you can select a few options to make sure that your data is understood properly. For instance, you can tell Refine to find which columns are numbers and which are texts. Numbers are then highlighted in green.

Once you are happy with your data, you can name and create the project. Only at this step is your project created and saved.

Now, with your project created, you notice that you are shown only 10 rows. Refine does that to save computer resources. Imagine if you had a million rows, your computer would suffer. If you prefer to see more rows, you can change the number of rows shown to 20 or 50. What’s really important though is the number just above, where Refine tells you how many records are in the table. In our case, we have the height of 40 people, so we need 40 rows. Always make sure that the number you have here matches with your data.

The most useful concept of Refine are facets. They are a simple version of a pivot table. Let’s do a facet on Gender, for instance, clicking on the arrow on the Gender column, then on Facet -> Text Facet. Refine lists all the different values in the Gender column and tells us how many rows of each we have. In this case, we have 20 males and 20 females in the sample. Let’s do the same on Place of birth. We see that we have 10 persons from Italy, 10 from Spain, 10 from Denmark and 10 from Sweden. What’s really great is that you can combine facets. Let’s select only males, and see how the numbers change in the Place of birth facet. That’s a very convenient way to mine your data very quickly.

You can also facet with numbers. Go to Facet -> Numeric facet on the Height column. What you see is a histogram of the height distribution. Each column in the histogram represents the number of rows for which the height is between 2 values. Try to select only the rows for the tallest people by moving the handles. Again, you can combine this with another facet, so that you can quickly spot where the tallest males were born, for instance.

Another feature of Refine is batch edit. Let’s say I’m just interested in 2 zones: Scandinavia and the Mediterranean. We can simply edit the contents of the cells directly in the facet and replace Sweden and Denmark by Scandinavia, Italy and Spain by Mediterranean.

If you’re not happy about the changes, you can undo them. Unlike other spreadsheet software, Refine stores all your modifications, so that you can always go back to the original version of the file. Here, I’ll go four steps back to get all my countries back.

The final step in this simple tutorial is to transform data points. Just like in spreadsheet softwares, you can modify the values of your data. However, the syntax is a bit different. Let’s try one operation. You might have noticed that the height is given in inches, let’s add the height in centimeters as well. To do that, we click on the height column and choose Edit Column -> Add column based on this column. Now we have an invite with just “value”. That represents the value in the column we selected in the first place. To transform inches into centimeters, we need to multiply this value by 2.54. The way it’s written is

value \* 2.54

And you see the outcome of the formula on a small subset below the formula. If it looks good, you can give a name to the new column and press OK.

Now, some values have quite a lot of decimals. The number of significant digits is out of proportion, compared to the original values. To fix this, we’re going to round the values. Instead of creating a new column this time, we go to Edit Cells → Transform. We have pretty much the same invite as before. To round a number, the function is

round(value)

Again, you can see the results in the preview before pressing OK.

To move on from this very brief overview of Open Refine, I suggest watching the official demo videos. You also have an extensive list of tutorials available on the Refine website. Refine can also be used with plugins or to access databases online. It’s a very powerful tool that has a gentler learning curve than Excel. I highly recommend it.

To go further, check the tutorials, which the community calls recipes.