**W45: OpenRefine**

**Exercise 1: Create a \*tidy\* spreadsheet/table listing the names of Danish monarchs with their birth- and death-date and duration of reign. They should be sortable by year of birth. Suitable source websites are**[**here**](https://kongehuset.dk/monarkiet-i-danmark/kongerakken)**and**[**here**](https://danmarkshistorien.dk/perioder/vikingetiden-ca-800-1050/)**, but you can also use another source, provided you reference it. (Collaboration is welcome. Remember to attach this spreadsheet to Brightspace submission)**

I will use the data available from kongehuset.dk ([here](https://kongehuset.dk/monarkiet-i-danmark/kongerakken)) to create the spreadsheet. I will not use any other sources. Here is an excerpt of the spreadsheet:

Et billede, der indeholder bord

Automatisk genereret beskrivelse

When creating the tidy spreadsheet, I made sure to:

* Not have multivalued attributes
* Be consistent with values, spelling and names
* Not color the cells
* Not have blank cells (instead I used NA for the data that was not available)
* Have one observation pr. row and cell
* Use \_ between words
* Be consistent with the dates (format: DD-MM-YYYY)

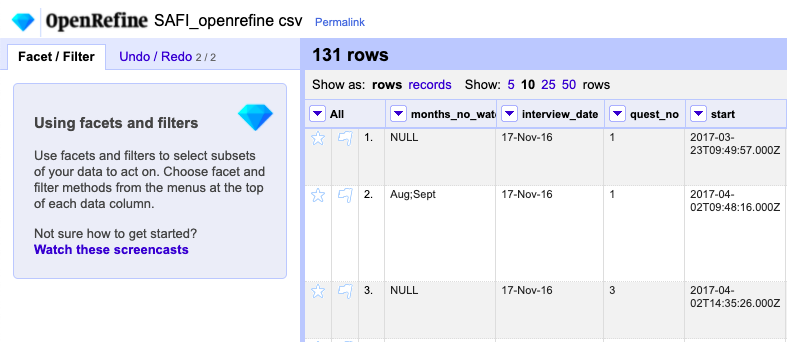
I have attached a CSV-file of the spreadsheet.

**Exercise 2: Does OpenRefine alter the raw data during sorting and filtering?**

OpenRefine does not alter the raw data during sorting and filtering. Therefore, OpenRefine is a good tool to use – it does not mess up or change your data. Moreover, OpenRefine is also free to use. One can sort and filter the data and explore it further. OpenRefine can be used to find errors such as values with spelling mistakes and it is always possible to undo any filters or manipulations done in the program. When the data is cleaned, it is possible to export it and share with other people so that they can reproduce it.

**Exercise 3: Fix the**[**interviews dataset**](https://ndownloader.figshare.com/files/11502815)**in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/driest by the interviewed farmer households?"**

Firstly, I open the dataset in OpenRefine and find the column called “months\_no\_water” because the months without water must be the driest. Next, I click the arrow in the column, and then I click “Edit column” and “Move column to the beginning”. I do this because it is easier to work with. Now my spreadsheet looks like this:



Then I click the arrow in the column “months\_no\_water” again and click “Facet” and “Text facet” to show which months are the driest. As you can see in the picture, it is very difficult to see the driest months when just looking at the data because the cells are multivalued:

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Therefore, I will split the data into the different months but before I do that, I must remove the square brackets, apostrophes, and spaces so I only get the name of the months. I do this by clicking the arrow in the column “months\_no\_water”, then “Edit cells” and “Transform…”. Then I type this code in the expression-line:

Et billede, der indeholder bord

Automatisk genereret beskrivelse

This replaces the characters mentioned above. The expression value.replace means that it replaces the first thing inside citation marks in the parentheses with the second thing inside citation marks. For example, value.replace(“[”,””) replaces the [ with nothing because there is nothing in the second citation marks. This is what I did with all the characters and instead of doing a new expression for each character I just kept adding .replace after the parentheses to minimize manual labor.

Now I split the data so I can sort it by the separate months instead of the multivalued cells and find out which two months are the driest. I do this by clicking the arrow in the column “months\_no\_water”, then “Facet” and “Custom text facet”. In the expression-line, I type in value.split(“;”) which splits the months in the multivalued cells that are separated by a semicolon:

Et billede, der indeholder bord

Automatisk genereret beskrivelse

Now I can see how many times the different months are mentioned as being without water. When sorting them by count in the facet, it is obvious that the two driest months are October (mentioned 74 times) and September (mentioned 70 times):

Et billede, der indeholder tekst

Automatisk genereret beskrivelse