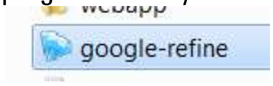
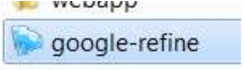
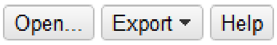
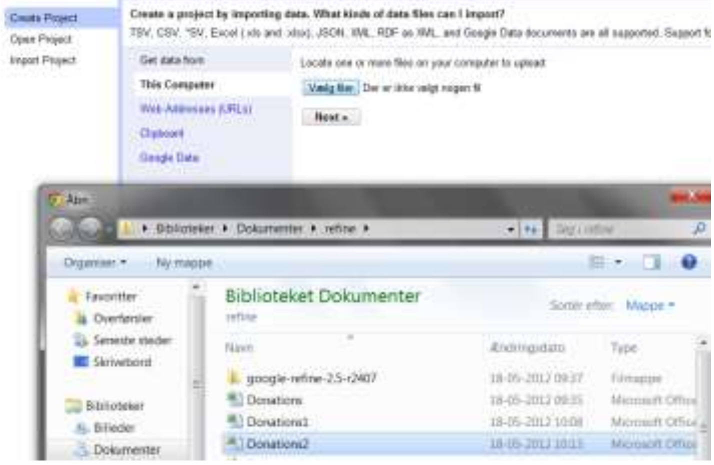
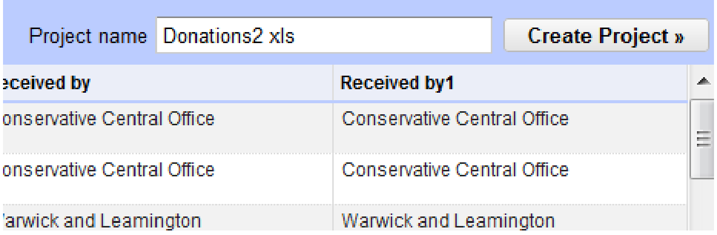
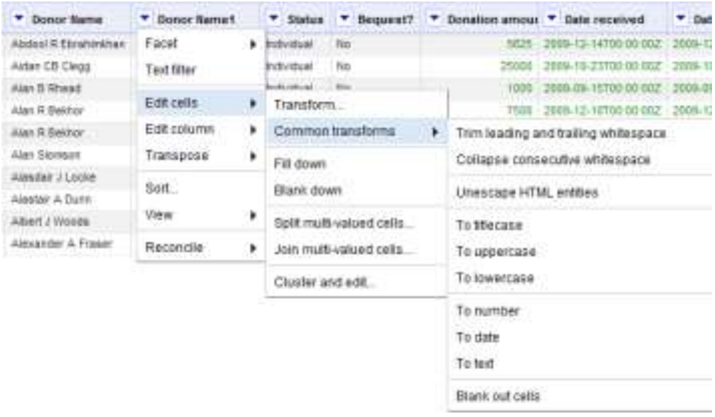
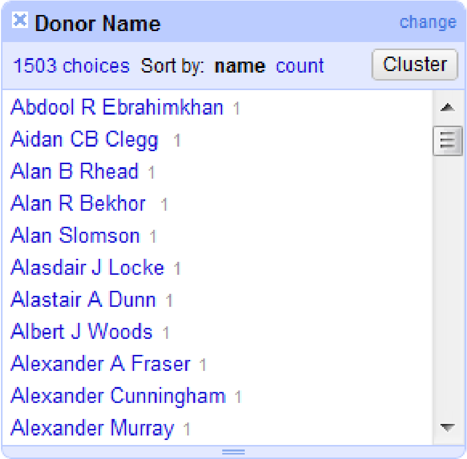
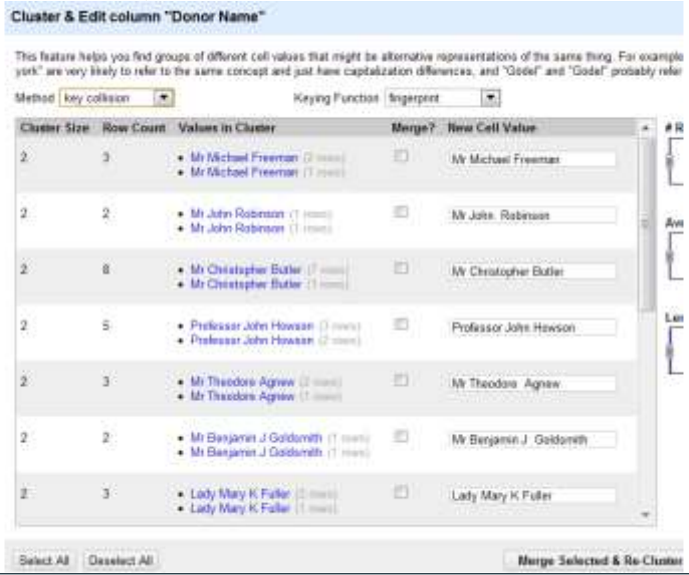
Kaas & Mulvad | 05.10.13 | OpenRefine | Download and install

MEMO

WHAT	WHY	HOW
	OpenRefine (ex-Google Refine) is a powerful tool for working with messy data, cleaning it, transforming it from one format into another, extending it with web services, and linking it to databases.	http://openrefine.org/
Who can use Refine?	Everybody can use Refine. It's free. The latest stable version is Google Refine 2.5. It's free. The next release, OpenRefine 2.6, will carry the new branding.	Download it to your PC, MAC or Linux.
Open Refine	OpenRefine opens in a browser in this url: http://127.0.0.1:3333/ It means it runs on your own machine. You are then ready to work on projects – either old or new after import of data.	Click on google-refine and run the program directly. 

WHAT	WHY	HOW
Prepare data	Prepare data in a spreadsheet. Use copy-paste and make sure that you have two versions of the the columns you will clean. This way it is possible to compare the original and cleaned.	http://code.google.com/p/google-refine/
Open Refine	Find the folder, where Google Refine is installed and start the program from here.	Click on google refine and run the program directly:  google-refine
Start project	If you should start a new project - click on the button "Open" in the upper right corner:  Open... Export ▾ Help	Click Open…
Choose file	 Create Project Open Project Import 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 all supported. Support fo Get data from This Computer Web Addresses (URLs) Clipboard Google Data Locate one or more files on your computer to upload Vælg filer Der er ikke valgt nogen fil Next » Åbn Biblioteker ▸ Dokumenter ▸ refine ▸ Organiser ▾ Ny mappe Favoritter Overførsler Seneste steder Skrivebord Biblioteker Billeder Dokumenter Biblioteket Dokumenter refine Sorter efter: Mappe ▾ Navn Ændringsdato Type google-refine-2.5-r2407 18-05-2012 09:37 Filmappe Donations 18-05-2012 09:35 Microsoft Office Donations1 18-05-2012 10:08 Microsoft Office Donations2 18-05-2012 10:13 Microsoft Office The project uses an excel file, Donations2, imported from the computer. Note that you can also choose data from URL's and Google and clipboard. URL's and Google provides unique opportunities to cleanse dirty data automatically and then send the cleaned data in to your own presentation. We choose the basic model of cleaning a spreadsheet.	Click on: Create Project Choose Files Higlight the relevant file Click Open Click Next >>
Create project	The Import Wizard allows you to change the settings. When importing from Excel it is usually not necessary.  Project name Donations2 xls Create Project » eceived by | Received by1 onservative Central Office | Conservative Central Office onservative Central Office | Conservative Central Office 'arwick and Leamington | Warwick and Leamington	Click on Create Project >>

WHAT	WHY	HOW
Trim cells	Refine contains a wealth of cleaning methods. We begin with removing spaces at the beginning and end of each cell:  Donor Name Donor Name1 Status Bequest? Donation amou Date received Facet Text filter Edit cells Edit column Transpose Sort... View Reconcile Transform... Common transforms Fill down Blank down Split multi-valued cells... Join multi-valued cells... Cluster and edit... Trim leading and trailing whitespace Collapse consecutive whitespace Unescape HTML entities To titlecase To uppercase To lowercase To number To date To text Blank out cells Note also that we are working with the column called Donor Name1, while Donor Name is unchanged throughout the process. Google Refine also contains a good log system, so the process can be rolled back, yet it is wise to preserve the original content in a column.	Click the dropdown ▾ next to the heading of the column you want to clean.
Facet	Facet is the name of the main cleaning method. Select Text Facet, which offers a variety of mathematical methods to compare text. When you click, you get the following image on the left side of the screen:  Donor Name change 1503 choices Sort by: name count Cluster Abdool R Ebrahimkhan 1 Aidan CB Clegg 1 Alan B Rhead 1 Alan R Bekhor 1 Alan Slomson 1 Alasdair J Locke 1 Alastair A Dunn 1 Albert J Woods 1 Alexander A Fraser 1 Alexander Cunningham 1 Alexander Murray 1 Out of 2,126 rows we have donor names in 1,503 different versions.	Select Facet - Text facet Click Cluster
Cluster	Under "Cluster" you are offered various methods to compare and link the content of the cells. You can choose everything at once and deselect one by one.	Select method and function Select dovetailing Click Merge Selected & Re -

<table>
<tr>
<td></td>
<td>Or you can select it one by one.
There are two basic methods :
Key collision
Nearest neighbor

<b>Cluster & Edit column "Donor Name"</b>
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example
york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer
Method: key collision ▾ Keying Function: fingerprint ▾
<table>
<tr><th>Cluster Size</th><th>Row Count</th><th>Values in Cluster</th><th>Merge?</th><th>New Cell Value</th></tr>
<tr><td>2</td><td>3</td><td>• Mr Michael Freeman (2 rows)
• Mr Michael Freeman (1 rows)</td><td>□</td><td>Mr Michael Freeman</td></tr>
<tr><td>2</td><td>2</td><td>• Mr John Robinson (1 rows)
• Mr John Robinson (1 rows)</td><td>□</td><td>Mr John Robinson</td></tr>
<tr><td>2</td><td>8</td><td>• Mr Christopher Butler (7 rows)
• Mr Christopher Butler (1 rows)</td><td>□</td><td>Mr Christopher Butler</td></tr>
<tr><td>2</td><td>5</td><td>• Professor John Howson (3 rows)
• Professor John Howson (2 rows)</td><td>□</td><td>Professor John Howson</td></tr>
<tr><td>2</td><td>3</td><td>• Mr Theodore Agnew (2 rows)
• Mr Theodore Agnew (1 rows)</td><td>□</td><td>Mr Theodore Agnew</td></tr>
<tr><td>2</td><td>2</td><td>• Mr Benjamin J Goldsmith (1 rows)
• Mr Benjamin J Goldsmith (1 rows)</td><td>□</td><td>Mr Benjamin J Goldsmith</td></tr>
<tr><td>2</td><td>3</td><td>• Lady Mary K Fuller (2 rows)
• Lady Mary K Fuller (1 rows)</td><td>□</td><td>Lady Mary K Fuller</td></tr>
</table>
# R
Ave
Len
Select All
Deselect All
Merge Selected & Re-Cluster

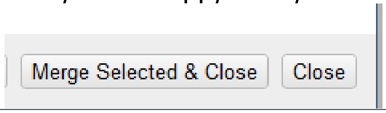
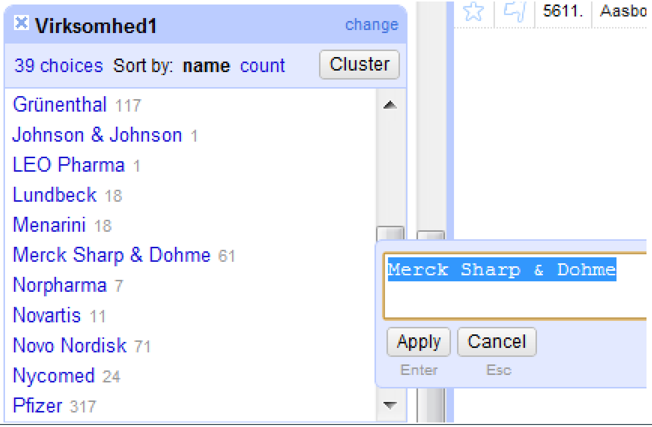
</td>
<td>Cluster</td>
</tr>
<tr>
<td><b>Key collision</b></td>
<td>Key collision looking for convergence of important words. It is based on the idea of bringing together the core content of a text string. The method is subdivided into these three :

<b>Fingerprint:</b> Looking for identical characters and match hereinafter "John Smith" to "Smith, John". This method removes all special characters, set everything to lowercase, splits in words separated by spaces and removes duplicates. It also converts special versions of letters to their ascii-based value, ie ö becomes o.
Fingerprint finds the various ways in which the significant words are put together.

<b>N- gram Fingerprint :</b> Using the same principles as fingerprint, but allows variation of the word. N stands for the number of variations, permitting. That is, the higher n is, the greater is the difference between the words in the match. Even high N gives relatively few false. Try to vary.
1 - gram allows, for example, match "Krzysztof”, "Kryzysztof" and "Krzystof" as identical.
N-grams find many typos.

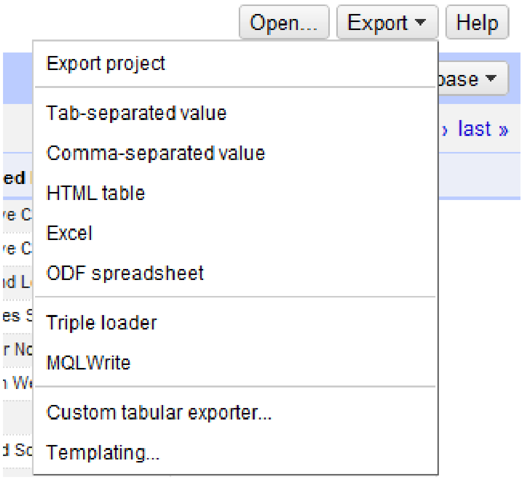
<b>Phonetic Fingerprint:</b> Looking for the same sounds and finds, for example, "Horowitz" and "Horowicz". <b>Metaphone3</b> used for English, while <b>Cologne-phonetic</b> is German sounds.

Normally, you start to cluster with key collision and takes these methods from the top down.</td>
<td><b>See more here:</b>
https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth</td>
</tr>
</table>

Nearest neighbor	Nearest neighbor is the second basic method. While key collisions is very fast and well defined, it can sometimes be either too strict or loose and end up giving you much manual work to clean. Here it is an idea to try to fine-tune the selection of nearest neighbor that makes it possible to provide the parameters (like you can in n- gram). **Levenshtein:** The method looks at how many edits that is needed to get two strings match. "Paris" and "paris" has an edit distance of 1 because P must be changed to p "New York" and "newyork" has an edit distance of 3 In Refine you can define both the radius (distance) and Block Chars. Distance is absolutely central to work with. The higher the more difference there may be in text blocks while they are still perceived as equal. **PPM (Prediction by Partial Matching):** The method examines how much difference there is in two strings. It looks like Levenshtein, but uses a different mathematical model. With Radius and Block Chars you can ask about how much difference you will accept and still consider it as a match. Matches typically "Johnson" and "Johnsons" .	
End Cluster	Once you are happy with your match, terminate the cluster.  	Click Close
Edit manually	Although there are many algorithms to find duplicate names, so they will not find everything. Merck Sharp & Dohme may for example be as abbreviated MSD. It must be edited manually by simply typing in the cell.  Virksomhed1 change 39 choices Sort by: name count Cluster 5611. Aasbo Apply Cancel Enter Esc Grünenthal 117 Johnson & Johnson 1 LEO Pharma 1 Lundbeck 18 Menarini 18 Merck Sharp & Dohme 61 Norpharma 7 Novartis 11 Novo Nordisk 71 Nycomed 24 Pfizer 317 Merck Sharp & Dohme	

Calculation	Refine has its own editing language. This can be used for calculations. Often it is easier to export to Excel and do the calculations here.	

MEMO

WHAT	WHY	HOW
Export	After cleaning of data, you would like to continue to analyze them in Excel.  Open... Export Help Export project Tab-separated value Comma-separated value HTML table Excel ODF spreadsheet Triple loader MQLWrite Custom tabular exporter... Templating... last » Choose Excel. Google Refine sends automatically an Excel-version to the browser download. Then you can work with it in Excel.	Click on Export Choose Excel