



# Handleiding OpenRefine

Shana Lardinois  
1706160

Pelle de Graaf  
1717399

Ranoy de Graaf  
1763026

Chris Smit  
1723196

Nina van der Werf  
1715528

Docent: Daniela van Geenen en Wouter van Dijke

Datum van inleveren: 28-10-2021

Versie: 1

Hogeschool Utrecht

Minor: Datavisualisatie en Infographics

## Inhoud

Introductie .....	2
1. Over de applicatie.....	3
1.1 Over OpenRefine.....	3
1.2 Geschiedenis & Ontwikkeling.....	3
2. Voor- en nadelen .....	4
2.1 Voordelen: .....	4
2.2 Nadelen:.....	4
3. Voorbereiding.....	5
4. Termologie & Functies.....	6
5. How to's .....	7
Stap 1: Aanmaken van het project.....	7
Stap 2: Data filteren.....	8
Stap 3: Data clusteren.....	8
Stap 4: Datums filteren.....	10
Stap 5: Nummers omzetten. ....	12
Stap 6: Uitschieters zoeken .....	14
Stap 7: Exporteren .....	15
6. Verder leren .....	17
7. Conclusie .....	18

## Introductie

In deze handleiding zal de applicatie OpenRefine stap voor stap worden uitgelegd. Aan de hand van afbeeldingen zal de flow door de applicatie worden gevisualiseerd. Tijdens deze handleiding hopen wij duidelijk te maken hoe OpenRefine in elkaar zit, en in welke situaties het handig is om te gebruiken.

OpenRefine gaat eigenlijk om het verwerken en opschonen van data. Tijdens deze handleiding van OpenRefine zullen de leerdoelen zijn om aan de hand van dit programma;

- Data op te schonen
- Data te transformeren
- Data te formateren
- Data te verbreden.

OpenRefine kan dus eigenlijk worden gezien als de eerste stap voor het opschonen van je data. Hierbij hebben we het dan vooral over grote aantallen data; bij kleine bestanden is het niet per se heel voordelig om met OpenRefine te werken omdat het hier vaak ook handmatig kan worden gedaan. Dus zodra grote aantallen waarden van een dataset moeten worden opgeschoond, gefilterd en getransformeerd is OpenRefine de perfecte tool om je data gestructureerd en overzichtelijk te krijgen.

Een voorbeeld van een loopbaan waar OpenRefine handig zou kunnen zijn, is tijdens onze minor Datavisualisatie & Infographics. Aangezien dit een journalistieke minor is, zal er veel worden gewerkt met het verzamelen en opschonen van data verkregen uit enquêtes. Hierna kan de data vaak rommelig zijn, doordat er veel verschillende antwoorden en waarden zijn gekregen van je deelnemers. OpenRefine zorgt ervoor dat deze waarden worden opgeschoond waarna er een duidelijker beeld krijgt van het onderzoek kan worden gevormd.

# 1. Over de applicatie

## 1.1 Over OpenRefine

OpenRefine is een open-source desktopapplicatie voor het opschonen en omzetten van gegevens, voorheen bekend onder de naam ‘Google Refine’.

OpenRefine verwerkt de gegevens van de gebruiker door een kleine lokale server op de computer van de gebruiker te plaatsen. De gebruiker communiceert met deze server via een webbrowser, maar de data blijft op de computer en kan alleen gedeeld worden door de gebruiker zelf. Dit maakt OpenRefine een veilige, privé-omgeving waarin de gebruiker onbezorgd en snel grote datasets kan opschonen en verwerken.

## 1.2 Geschiedenis & Ontwikkeling

Google Refine vond zijn oorsprong in het open-source project ‘Freebase Gridworks’ van [Metaweb Technologies](#). De applicatie was ontwikkeld door David Huynh om de Freebase-database en -gemeenschap te ondersteunen voor het opschonen, afstemmen en uploaden van gegevens.

In juli 2010 nam Google Metaweb Technologies over en daarmee ook Freebase Gridworks. Freebase bleef een open-source project voor het opschonen van gegevens, maar werd omgedoopt tot ‘Google Refine’ en de code en documentatie werden verplaatst naar een code.google.com-adres. Met de steun van Google-technici en -gemeenschap zijn in de periode van 2010 t/m 2012 meerdere upgrades van Google Refine uitgevoerd. De link met Freebase bleef desondanks aanwezig in Google Refine, omdat de tool afstemming met de Freebase-database ondersteunt.

Dankzij de gebruiksvriendelijke interface van Google Refine was vrijwel iedereen in staat de applicatie te gebruiken. Google Refine wekte interesse in veel verschillende vakgebieden en groeide dan ook hard. Nieuwe databases werden gebouwd en de code werd uitgebreid. Bibliothecarissen, journalisten en data-analisten gebruiken Google Refine om hun gegevens op te schonen en op elkaar af te stemmen. En een levendige gemeenschap ontstond die nieuwe mogelijkheden bood om de applicatie te verfijnen.

Refine is bij Google uitgegroeid tot een desktop-gebaseerde applicatie die zowel offline als online kan werken. Desondanks heeft Google gedurende deze tijd de applicatie niet ontwikkeld als een cloud gebaseerde tool en is de applicatie ook nooit gekoppeld aan andere Google-services (zoals bijv. Drive). Hierdoor is Refine altijd een veilige privé-omgeving gebleven voor de gegevens van de gebruiker.

In oktober 2012 heeft Google haar handen van Refine afgetrokken en staat de applicatie bekend onder de naam OpenRefine. De code en documentatie is sinds dien te vinden op [GitHub](#) waar tientallen bijdragers de applicatie onderhouden. De laatste update, OpenRefine 3.4, werd in September 2020 gelanceerd.

## 2. Voor- en nadelen

Om een beter beeld te krijgen van de situaties waarin je OpenRefine makkelijk zou kunnen gebruiken, zijn hieronder de voor- en nadelen genoemd.

### 2.1 Voordelen:

- De applicatie is gratis, wat het aantrekkelijker maakt om eens uit te proberen.
- Het is een lichte applicatie, waardoor het niet echt veel ruimte in beslag neemt op je pc.
- De applicatie kan snel grote datasets verwerken.
- De leercurve is groot. Je kan het werken met OpenRefine zo moeilijk maken als dat je zelf wil.
- De applicatie is geschikt voor zowel online- als offline gebruik.
- De stappengeschiedenis wordt geregistreerd waardoor de gebruiker op elk moment terug kan naar een eerdere versie van het project.
- De projecten worden automatisch opgeslagen.
- De applicatie biedt een combinatie van handmatige en geautomatiseerde controle over de stappen die je maakt in de applicatie. Dit zorgt ervoor dat je makkelijk die acties van je computer kan controleren. Hierdoor kun je dus computerfouten tijdens het opschonen van je data voorkomen.

### 2.2 Nadelen:

- De 'oude' look van de applicatie kan afschrikken.
- De applicatie is eenvoudig, maar vereist desondanks toch enige voorkennis.

### 3. Voorbereiding

In deze handleiding gaan we aan de slag met de applicatie OpenRefine, om deze handleiding optimaal te kunnen gebruiken heb je natuurlijk OpenRefine zelf nodig en een demo-dataset.

#### **Downloaden**

OpenRefine kan gratis gedownload worden van hun eigen website: <https://openrefine.org/download.html>. Op deze pagina kun je de versie voor Windows en Macintosh downloaden. Dit zijn respectievelijk de Windows kit of de Mac kit.

#### **Windows**

Op Windows computers is het een vereiste dat je Java geïnstalleerd hebt. Zodra je de Windows kit hebt gedownload, kun je het gedownloade bestand uitpakken. Vervolgens open je het bestand 'openrefine.exe' (of 'refine.bat' als het eerstgenoemde bestand niet werkt). Wanneer je een van deze bestanden hebt geopend komt je standaard internetbrowser geopend en wordt OpenRefine gestart. Een voordeel van OpenRefine is dat het volledig in je webbrowser gebruikt kan worden.

#### **Mac**

Voor Macintosh heb je wat meer handelingen nodig om OpenRefine te kunnen starten. Dit heeft te maken met het feit dat Apple zijn apparaten wat strenger heeft beveiligd dan Windows, maar dat is uiteraard geen probleem. In dit deel van de handleiding vermelden we stapsgewijs hoe je OpenRefine kunt openen.

Allereerst download je de Mac kit op de pagina die we hierboven aangeven.

Het bestand wordt gedownload in je downloads map. Zodra dit is voltooid open je de map en klik je met je rechtermuisknop op de applicatie. In het menu dat tevoorschijn komt kies je voor 'open'. Vervolgens krijg je een pop-up waarin je opnieuw kiest voor 'open' en daarna verschijnt er een venster met daarin de applicatie zelf. De applicatie kun je naar je map 'Apps' kopiëren, of rechtstreeks vanuit het venster openen. Hiervoor moet je opnieuw met de rechtermuisknop klikken op de applicatie, en vervolgens kiezen voor 'open'. Je krijgt opnieuw een waarschuwing waarin je kiest voor 'open'. Dit geldt ook voor wanneer je het bestand kopieert naar je 'apps' map. Nu opent de applicatie zich in je standaardbrowser en kun je aan de slag met OpenRefine!

#### **Demo-dataset**

In deze handleiding werken we met een demo-dataset. Deze demo-dataset gaat over het salaris van managers in o.a. verschillende beroepenvelden en landen over 2021.

Je kunt de dataset downloaden via de volgende link:

<https://docs.google.com/spreadsheets/d/1IPS5dBSGtwYVbjsfbaMCYIWnOuRmJcbequohNxCyGVw/edit?usp=sharing>

## 4. Termologie & Functies

Tijdens het werken met dit programma, zal snel duidelijk worden dat er een aantal termen zijn die je moet begrijpen om te kunnen werken met dit programma. Hieronder is een lijst te vinden met de termologie en de bijbehorende definities bij het gebruiken van OpenRefine.

### Algoritme

Een algoritme is een reeks instructies voor een computer en zorgt bijvoorbeeld voor het vinden van patronen in het bestand, zoals het zoeken van gelijknamige waarden.

### Cel

In OpenRefine worden alle waarden opgeslagen in een eigen cel. Een cel is niks meer of minder dan een hokje waar elke waarde in geschreven staat. Na elke bewerking wordt er weergegeven hoeveel cellen er bewerkt zijn, dit geeft een inzicht hoe groot de mutatie is die je doorgevoerd hebt.

### Clusteren

Een belangrijke functie van OpenRefine is de cluster tool. Deze tool kan de facet functie nog verder uitbreiden door waarden die veel op elkaar lijken samen te voegen. Zoals alle waarden voor “USA” en “usa” samen te voegen onder “USA”. Deze overeenkomsten kan OpenRefine zelf ontdekken en hiervoor hoeft de gebruiker alleen toestemming te geven.

### CSV-bestand

Een CSV, oftewel Comma Separated Value, -bestand is een bestand waar verschillende waardes opgeslagen kunnen worden, gescheiden door een komma of ander scheidingsteken. Een CSV-bestand kan hierdoor door veel programma's gebruikt worden omdat er een duidelijk en compact formaat gebruikt wordt.

### Export

Wanneer alle bewerkingen doorgevoerd zijn zal de gebruiker graag gebruik willen maken van de data die overgebleven is. In plaats van al deze data te kopiëren en plakken kan er een export gemaakt worden. Dit is een functie die alle waarden in het bestand automatisch op een bepaalde manier opslaat, bijvoorbeeld door het op te slaan in een Excel formaat zodat deze later gebruikt kan worden in Excel.

### Facet

Een Facet is een uitgebreide variant op de traditionele filter optie. Door een facet worden alle verschillende waardes die in het CSV-bestand staan gegroepeerd. Hiermee kan de gebruiker snel zien hoe vaak een bepaalde waarde, bijvoorbeeld “Nederland”, voor komt in het bestand.

### Filteren

In OpenRefine wordt er veel gebruik gemaakt van filteren. Heerbij wordt het bestand gefilterd op bepaalde zoekcriteria. Dit wil zeggen dat alleen waarden getoond worden die hetzelfde zijn als de criteria van het filter, waardoor een gebruiker bijvoorbeeld alleen maar waarden uit Nederland kan inzien.

### Formule

In OpenRefine worden formules gebruikt om, op een complexere manier, waarden te veranderen. Zo kan de gebruiker bijvoorbeeld een formule gebruiken om alle komma's uit een kolom te verwijderen.

## 5. How to's

### Stap 1: Aanmaken van het project

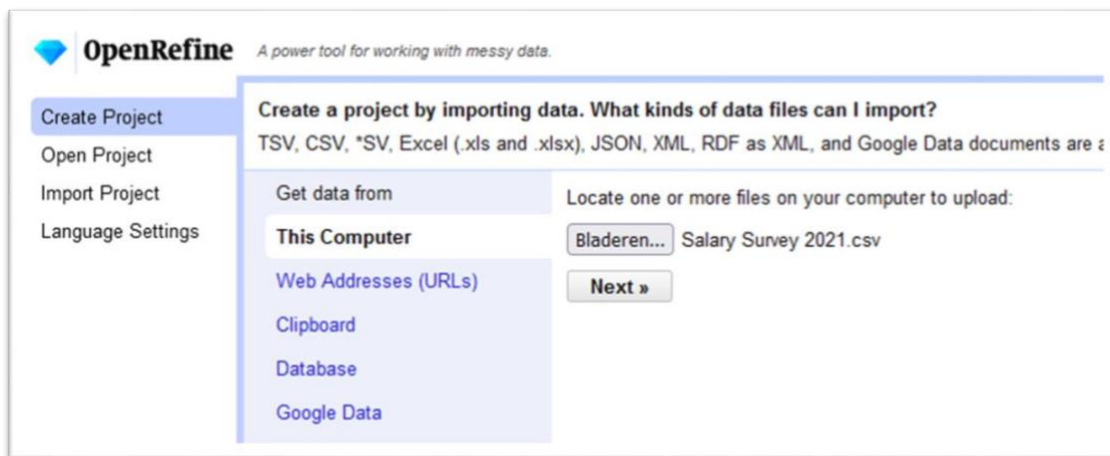


Figure 1 – Het creëren van een project

Tijdens het aanmaken van een project kies je welke data je wil gebruiken. Hierbij kun je voor bovenstaande opties kiezen. In deze demo importeren we een CSV-bestand maar als je op een andere manier data wil importeren kan dat ook.

Year	Month	Industry	Job title	Additional context on job title	Annual salary	Other monetary comp	Currency	Country	State	City	Overall years of professional experience	Years of experience in field	Highest level of education completed	Gender	Race
2014	01	Education	Research and Instruction Librarian		39,000	0	USD	United States	Massachusetts	Boston	3.7 years	3.7 years	Master's degree	Woman	White
2014	02	Education	Change Release Communications Manager		34,000	4000	USD	United Kingdom		Cambridge	9-10 years	5.7 years	College degree	Man	White
2014	02	Accounting, Billing & Finance	Working Specialist		34,000	0	USD	USA	Tennessee	Chattanooga	2-4 years	2-4 years	College degree	Woman	White
2014	02	Nonprofits	Program Manager		42,000	8000	USD	USA	Wisconsin	Milwaukee	9-10 years	5.7 years	College degree	Woman	White
2014	02	Accounting, Billing & Finance	Accounting Manager		60,000	7000	USD	USA	South Carolina	Columbia	16-18 years	5.7 years	College degree	Woman	White
2014	02	Education	Schools Planning Librarian		42,000	0	USD	USA	Iowa	Hammond	8-10 years	2-4 years	Master's degree	Man	White
2014	02	Education	Publications Assistant		35,000	2000	USD	USA	South Carolina	Columbia	2-4 years	2-4 years	College degree	Woman	White
2014	02	Education	Librarian	high school, FT	50,000	0	USD	United States	Alabama	Tonka	5.7 years	5.7 years	Master's degree	Man	White
2014	02	Education	Systems Analyst	State Developer/IT Developer	102,000	10000	USD	USA	Missouri	St Louis	25-30 years	21-26 years	College degree	Woman	White
2014	02	Accounting, Billing & Finance	Senior Accountant		45,000	0	USD	United States	Florida	Palm Coast	21-30 years	21-26 years	College degree	Woman	Hispanic, Latin or Spanish origin, White
2014	02	Nonprofits	Office Manager		47,000	0	USD	United States	Alabama	Bozale, AL	5.7 years	5.7 years	College degree	Woman	White
2014	02	Education	Health Care Support Worker	Health Care Support Worker	42,000	0	USD	USA	Pennsylvania	Scranton	11-20 years	5.7 years	PHD	Woman	Hispanic, Latin or Spanish origin, White
2014	02	Accounting, Billing & Finance	Manager of Information Services		100,000	0	USD	United States	Michigan	Detroit	11-20 years	11-20 years	College degree	Man	Asian or Indian origin, White
2014	02	Law	Legal and Ethical Advisor	non-profit law firm	52,000	0	USD	United States	Minnesota	Saint Paul	2-4 years	2-4 years	Master's degree	Woman	White
2014	02	Health care	Public care coordinator		32,000	0	CAD	Canada	Ontario	Hamilton	1 year or less	1 year or less	College degree	Woman	White
2014	02	Telecommunications	Quality Test Compliance Specialist		24,000	800	USD	United Kingdom		Lincoln	11-20 years	5.7 years	College degree	Man	White
2014	02	Business or Consulting	Graphic Designer		30,000	0	USD	USA	Illinois	Chicago	16-18 years	8-10 years	Some college	Woman	White
2014	02	Business or Consulting	Business or Financial Operations Manager		50,000	0	USD	USA	California	Fontana	21-30 years	21-26 years	College degree	Woman	White
2014	02	Business or Consulting	Business or Financial Operations Manager		30,000	1000	USD	USA	Georgia	Atlanta	10-20 years	2-4 years	Master's degree	Woman	White
2014	02	Education	Assistant Director of Healthcare Services		74,000	0	USD	United States	Florida	Boca Raton	11-20 years	11-20 years	Master's degree	Woman	White
2014	02	Health care	Data Management Specialist		74,000	0	USD	USA	Pennsylvania	Philadelphia	5.7 years	5.7 years	Master's degree	Woman	White
2014	02	Nonprofits	Program Coordinator & Assistant Center Staff Manager		50,000	0	USD	United States	Illinois	Aurora	2-4 years	2-4 years	PHD	Woman	White
2014	02	Nonprofits	Health Care Support Worker		43,000	0	CAD	Canada	Ontario	Toronto	11-20 years	8-10 years	Master's degree	Woman	White

Figure 2 – Overzicht van de dataset

Wanneer je je bestand(en) gekozen hebt kun je zien wat OpenRefine gevonden heeft in jouw gekozen dataset. In ons geval zien we een dataset van een vragenlijst over salarissen. OpenRefine heeft al gelijk het een en ander toegekend zoals titels en witregels. Dit is eventueel handmatig in te stellen als hier nog iets ontbreekt of verkeerd staat.



## Step 2: Data filteren

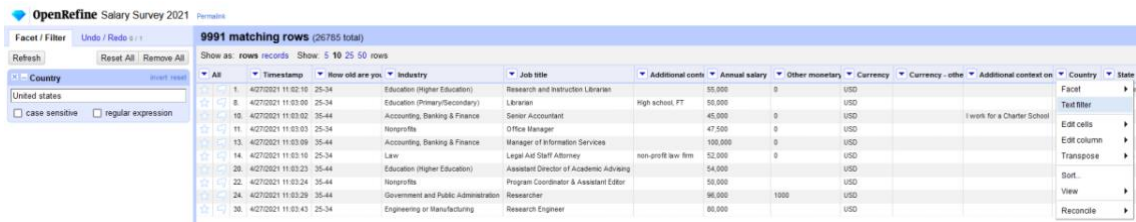


Figure 3 – Facet



Figure 4 – Facet > Tekst facet

Als het project aangemaakt is komen we in een scherm waar alle kolommen naast elkaar staan. Als eerst gaan we gebruik maken van een filterfunctie. Deze functie wordt, net als in Excel, gebruikt om de dataset te filteren op datgene wat je invoert.

OpenRefine heeft ook een andere manier om een dataset te filteren. Dit noemen ze een Facet, een facet geeft de mogelijkheden binnen een kolom. In dit voorbeeld alle verschillende tekst die ingevuld is bij het kopje “Country”.

## Step 3: Data clusteren

### Cluster & Edit column “Country”

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  Keying Function  35 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
9	3277	<ul style="list-style-type: none"> <li>US (2538 rows)</li> <li>U.S. (573 rows)</li> <li>Us (106 rows)</li> <li>us (33 rows)</li> <li>U S (20 rows)</li> <li>U.s. (3 rows)</li> <li>u.s. (2 rows)</li> <li>U.S&gt; (1 rows)</li> <li>uS (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	United States of America
9	8763	<ul style="list-style-type: none"> <li>USA (8097 rows)</li> <li>Usa (432 rows)</li> <li>usa (176 rows)</li> <li>U.S.A. (46 rows)</li> <li>U.S.A (8 rows)</li> <li>U.S.A (1 rows)</li> <li>U.s.a (1 rows)</li> <li>UsA (1 rows)</li> <li>uSA (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	United States of America
8	9525	<ul style="list-style-type: none"> <li>United States (9200 rows)</li> <li>United states (205 rows)</li> <li>united states (106 rows)</li> <li>UNITED STATES (7 rows)</li> </ul>	<input checked="" type="checkbox"/>	United States of America

# Choices in Cluster: 2 — 9

# Rows in Cluster: 0 — 9600

Average Length of Choices: 2 — 24

Length Variance of Choices: 0 — 1.25

Select All Unselect All

Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Figure 5 – Cluster & Edit column

Helaas vult niet iedereen deze kolom op dezelfde manier in. Sommige mensen kiezen voor “U.S.” om de verenigde staten aan te duiden. Andere gebruiken “USA” of “usa”, kortom er

zijn verschillende namen voor hetzelfde land. Dat is voor het verwerken van data niet handig. Daarom heeft OpenRefine een Cluster functie ingebouwd. Deze functie gaat opzoek naar overeenkomende waarden en maakt daar een waarde van.

**Cluster & Edit column "Country"**

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** | 3 clusters found

Cluster Size	Row Count	Values in Cluster	Row Cell Value
3	30	<ul style="list-style-type: none"> <li>United State (19 rows)</li> <li>United Stated (10 rows)</li> <li>United States (1 row)</li> </ul>	United States of America
2	8	<ul style="list-style-type: none"> <li>England, UK (7 rows)</li> <li>England/UK (1 row)</li> </ul>	United Kingdom
2	2	<ul style="list-style-type: none"> <li>United States (1 row)</li> <li>Uniteed States (1 row)</li> </ul>	United States of America

Keying Function dropdown menu options: fingerprint, ngram-fingerprint, metaphone3, cologne-phonetic, Daitch-Mokotoff, Beider-Morse

Visualizations on the right:

- # Choices in Cluster: 2 — 3
- # Rows in Cluster: 2 — 30
- Average Length of Choices: 10.5 — 14
- Length Variance of Choices: 0 — 0.8170000000000001

Buttons: Select All, Unselect All, Export Clusters, Merge Selected & Re-Cluster, Merge Selected & Close, Close

Figure 6 – Cluster & edit column – Ngram-fingerprint

Voor het Clusteren van waarden zijn er verschillende manieren. OpenRefine heeft verschillende algoritmen die voor een andere clustering zorgen. Grofweg kunnen deze van boven naar beneden ingedeeld worden van erg passend naar midden passend. Dus de eerste, Fingerprint, laat alleen waarden zien die uit dezelfde karakters bestaat en de laatste gaat dieper kijken naar verbanden om te achterhalen of twee waarden hetzelfde betekenen.

## Stap 4: Datums filteren

**OpenRefine** Salary Survey 2021 [Permalink](#)

Facet / Filter Undo / Redo 0 / 0

Refresh Reset All Remove All

26785 rows

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

Timestamp change reset

NaN-NaN-NaN NaN:NaN:NaN — 1970-01-01 01:00:00

Time 0  Non-Time 26785  Blank 0  Error 0

All	Timestamp	How old are you	Industry
1. Facet	Facet		
2. Text filter	Text filter		
3. Edit cells	Edit cells		
4. Edit column	Edit column		
5. Transpose	Transpose		
6. Sort...	Sort...		
7. View	View		
8. Reconcile	Reconcile		

- Facet
  - Text facet
  - Numeric facet
  - Timeline facet**
  - Scatterplot facet
  - Custom text facet...
  - Custom Numeric Facet...
  - Customized facets

Figure 7 – Datums filteren – Facet > timeline Facet

Zoals we eerder bij de tekst facet zagen is het ook mogelijk om datums te filteren. Een tekst facet is hier niet handig voor omdat er te veel unieke waarden zijn. Zoals je ziet kun je dit niet al gelijk toepassen. Wanneer je gelijk voor een “Timeline facet” kiest krijg je 26785 Non-Time waarden. Dit komt omdat OpenRefine nog niet weet dat deze waarden datums zijn.

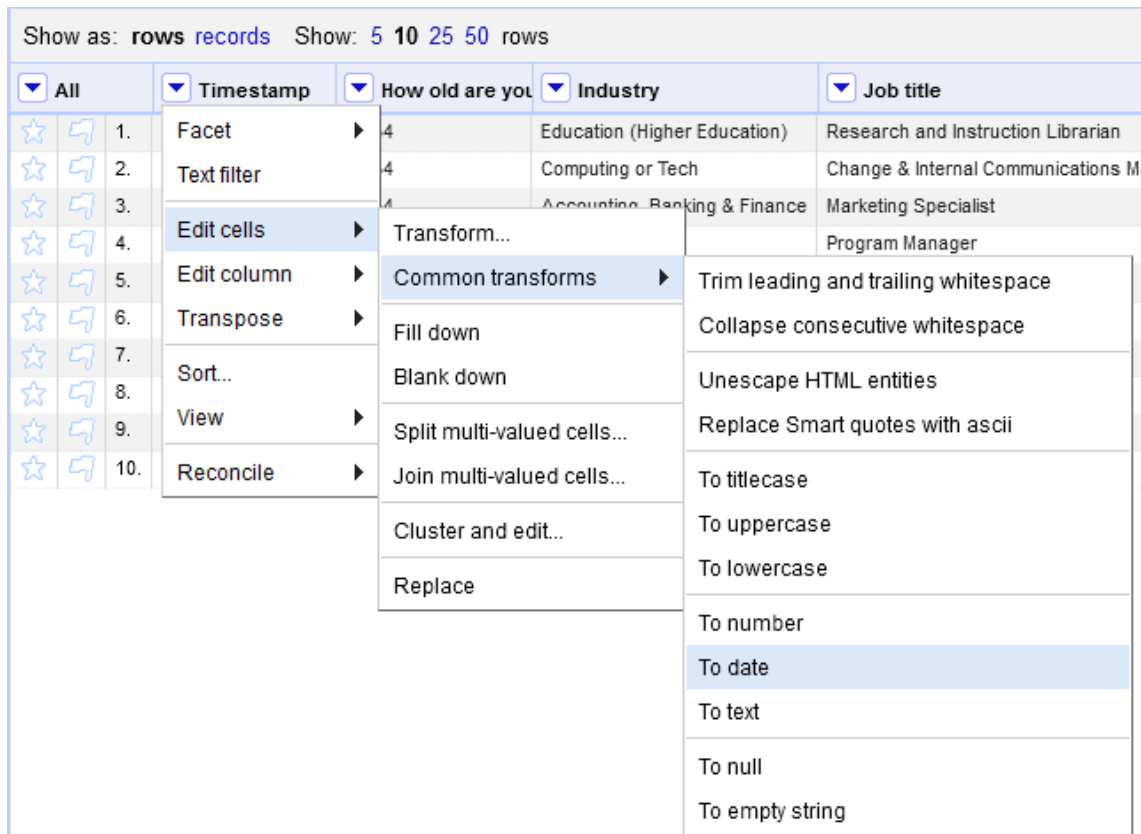


Figure 8 – Datums filteren – Edit cells > Common transforms > to date

Om OpenRefine te vertellen dat dit om datums gaat kun je via: **Edit cells > Common transformations > To date** de waarden automatisch naar een datum formaat omzetten die OpenRefine herkent. Hiermee verandert OpenRefine verschillende datum notaties allemaal te gelijk. Zoals **1-1-2021** of **1/1/2021** of **1 januari 2021**.

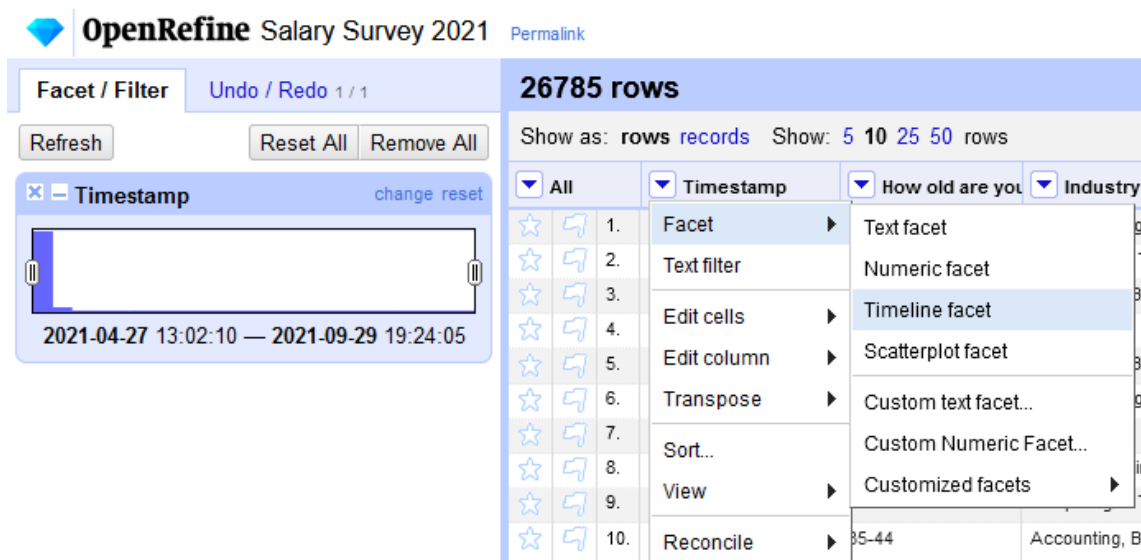


Figure 9 – Datums filteren – Facet > Timeline Facet

Nu OpenRefine weet dat dit datums zijn kunnen we de Timeline facet toepassen. Via **Timestamp > Facet > Timeline facet**

## Stap 5: Nummers omzetten.

OpenRefine Salary Survey 2021

26785 rows

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

All	Timestamp	How old are you	Industry	Job title	Additional context on job	Annual salary	Other monetary	Currency
1	2021-04-27T11:02:10Z	25-34	Education (Higher Education)	Research and Instruction Librarian				
2	2021-04-27T11:02:22Z	25-34	Computing or Tech	Change & Internal Communications Manager				
3	2021-04-27T11:02:30Z	25-34	Accounting, Banking & Finance	Marketing Specialist				
4	2021-04-27T11:02:41Z	25-34	Nonprofits	Program Manager				
5	2021-04-27T11:02:42Z	25-34	Accounting, Banking & Finance	Accounting Manager				
6	2021-04-27T11:02:46Z	25-34	Education (Higher Education)	Scholarly Publishing Librarian				
7	2021-04-27T11:02:51Z	25-34	Publishing	Publishing Assistant				
8	2021-04-27T11:03:00Z	25-34	Education (Primary/Secondary)	Librarian	High school, FT			
9	2021-04-27T11:03:01Z	45-54	Computing or Tech	Systems Analyst	Data developer/ETL Developer			
10	2021-04-27T11:03:02Z	35-44	Accounting, Banking & Finance	Senior Accountant				

Figure 10 – Nummers omzetten – Facet > Numeric Facet

Annual salary	Other monetary	Currency	Currency - othe	Additional context on
Facet		USD		Ur
Text filter	0	GBP		Ur
		USD		Ur
Edit cells	Transform...			Ur
Edit column	Common transforms	Trim leading and trailing whitespace		Ur
Transpose	Fill down	Collapse consecutive whitespace		Ur
Sort...	Blank down	Unescape HTML entities		Ur
View	Split multi-valued cells...	Replace Smart quotes with ascii		Ur
Reconcile	Join multi-valued cells...	To titlecase		Ur
	Cluster and edit...	To uppercase		
	Replace	To lowercase		
		To number		
		To date		
		To text		
		To null		
		To empty string		

Figure 11 – Edit cells > common transforms > to number

**Text transform on 6483 cells in column Annual salary:**  
**value.toNumber() Undo**

Figure 12 – Resultaat van de transformatie met undo optie

Het laatste facet dat we behandelen is de “Numeric Facet” deze facet laat de verschillende nummers in de kolom zien. Hiervoor moeten we net zoals bij de datum eerst OpenRefine vertellen dat het om cijfers gaat.

Hiervoor gebruiken we de “to number” transformatie. Deze kun je vinden onder:

**Annual salary > Edit cells > Common transforms > To number**

We zien bovenaan de pagina dat er 6483 cells aangepast zijn (dit is dus niet de hele dataset).

# OpenRefine Salary Survey 2021

Facet / Filter Undo / Redo 4 / 4

Refresh Reset All Remove All

Annual salary change reset

0.00 — 44,000,000.00

Numeric 6483  Non-numeric 20302  Blank 0  Error 0

Figure 13 – visueel overzicht van het resultaat

Om dit te controleren kijken we naar de Numeric facet. Hierin zien we net zoals bij de datum dat er 20302 Non-numeric waarden zijn gevonden.

Annual salary Other monetary Currency

Facet

Text filter

Edit cells Transform...

Edit column Common transforms

Transpose Fill down

Sort... Blank down

View Split multi-valued cells...

Reconcile Join multi-valued cells...

Cluster and edit...

Replace

Figure 14 – Cijfers veranderen

## Annual salary > Edit cells > Common transforms > To number

Custom text transform on column Annual salary

Expression: `value.replace(',',',')` Language: General Refine Expression Language (GREL)

Preview

row	value	value.replace(',',',')
1	55.000	55000
2	54.000	54000
3	34.000	34000
4	62.000	62000
5	60.000	60000
6	62.000	62000

On error:  keep original  set to blank  store error  Re-transform up to 10 times until no change

OK Cancel

Figure 15- Expression – waardes vervangen

OpenRefine Salary Survey 2021

Facet / Filter Undo / Redo 26785 rows

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

Annual salary change reset

All	Timestamp	Row old are you	Industry	Job title	Additional context on job	Annual salary	Other monetary	Currency
1	2021-04-27T11:02:16Z	25-34	Education (Higher Education)	Research and Instruction Librarian				
2	2021-04-27T11:02:22Z	25-34	Computing or Tech	Change & Internal Communications Manager				
3	2021-04-27T11:02:36Z	25-34	Accounting, Banking & Finance	Marketing Specialist				
4	2021-04-27T11:02:41Z	25-34	Nonprofits	Program Manager				
5	2021-04-27T11:02:42Z	25-34	Accounting, Banking & Finance	Accounting Manager				
6	2021-04-27T11:02:46Z	25-34	Education (Higher Education)	Scholarly Publishing Librarian				
7	2021-04-27T11:02:51Z	25-34	Publishing	Publishing Assistant				
8	2021-04-27T11:03:05Z	25-34	Education (Primary/Secondary)	Librarian	High school, FT			
9	2021-04-27T11:03:01Z	45-54	Computing or Tech	Systems Analyst	Data developer/ETL Developer			
10	2021-04-27T11:03:02Z	35-44	Accounting, Banking & Finance	Senior Accountant				

Figure 16 – Facet > Numeric Facet

Om deze waarden te kunnen gebruiken moeten ze zonder punten of komma's genoteerd worden. We zien in dit voorbeeld dat er voor 1000 tallen gebruik gemaakt is van een scheidingsteken. Deze moeten we weghalen door een transformatie uit te voeren. Hiervoor ga je naar: **Annual salary > Edit cells > Transform**

Om deze waarden aan te passen moeten we een syntax opgeven. Dit is net zoals in Excel een formule die de originele waarden van een cel aanpast naar een nieuwe waarde. In ons geval maken we gebruik van de "Replace" formule. Deze formule verandert een waarde door een andere. In ons geval veranderen we de komma met niks. Dit doen we met de formule: `value.replace(",","")`

## Stap 6: Uitschieters zoeken

All	Timestamp	Row old are you	Industry	Job title	Additional context on job title	Annual salary	Other monetary	Currency	Currency - other	Additional context on income
3606	2021-04-27T12:11:17Z	25-34	Utilities & Telecommunications	Operations Manager		102000000		USD		
10655	2021-04-27T21:44:57Z	25-34	Retail	Compliance Manager		115000000		JPY		
11235	2021-04-27T23:51:27Z	25-34	Education (Primary/Secondary)	Native English Teacher		27000000		Other	KRW (Korean Won)	*There is a huge benefit of having housing payed for/compensated
11455	2021-04-28T01:31:57Z	55-64	Education (Higher Education)	Regional Operations & Training Manager		870000000	120000000	Other	EUR	
18499	2021-04-29T00:33:04Z	35-44	Media & Digital	TOEIC Test Developer	Only for test-prep books, not the actual test	380000000		Other	KRW	
18985	2021-04-29T06:19:32Z	25-34	Education (Higher Education)	Researcher	Junior researcher	180000000		Other	EUR	
22872	2021-05-01T04:47:56Z	35-44	ESL Teacher	ESL English Teacher		36000000	0	Other	Korean Won	
23491	2021-05-03T04:17:56Z	35-44	Law	Legal Editor	I provide language assistance (proofreading, editing, translation) as most of our law firm's clients are international.	43000000	150000	Other	KRW	The bonuses I receive are in the form of gift certificates for local businesses.
24444	2021-05-05T15:31:16Z	25-34	Entertainment	Voice Actor	Voice acting for tv shows, movies, video games	208000000	1000000	JPY		I get paid per role, so my total amount fluctuates from year to year. Also, the additional compensation is in the form of 'gifts' into English (and do not know anything comparable in US or UK will give key people (including actors) gifts as a thank you, as working together in the future.

Figure 17 – Tekst zoeken

Door de numeric facet te filter op de hele hoge waarden zien we dat er 9 veel hoger zijn dan de andere salarissen. Dit kan verschillende oorzaken hebben, in ons geval heeft iemand een verkeerde munteenheid ingevoerd. Hierdoor lijkt het bedrag heel hoog maar de munt eenheden zijn niet evenveel waard en verschillen daarom van aantal.

## Stap 7: Exporteren

All	Timestamp	How old are you	Education
Transform	27T11:02:10Z	25-34	Educ
Facet	27T11:02:22Z	25-34	Com
Edit rows	27T11:02:38Z	25-34	Acc
Edit columns	27T11:02:41Z	25-34	Nonj
View			
8.	2021-04		
9.	2021-04-27T11:03:01Z	45-54	Com
10.	2021-04-27T11:03:02Z	35-44	Acc

Figure 18 – Edit columns > Re-order/remove columns

Nu we de dataset flink opgeschoond hebben is het tijd om de data te exporteren. Alleen is het niet altijd nodig dat alle data geëxporteerd wordt. Hiervoor is het mogelijk om de kolommen te verplaatsen en verwijderen. Dit doe je via: **All > Edit columns > Re-order / remove columns...**

### Re-order / Remove Columns

Drag columns to re-order

Timestamp
How old are you?
Industry
Job title
Annual salary
Currency
Currency - other
Country
Gender

Drop columns here to remove

Other monetary comp
Additional context on income
State
City
Overall years of professional experience
Years of experience in field
Highest level of education completed
Race
Additional context on job title

OK Cancel

Figure 19 – Re-order/remove columns keuzemenu – drag & drop

In dit menu kun je door de blokken te slepen bepalen welke volgorde je wil gebruiken en welke kolommen je weg wil gooien.

Nu je alle kolommen op de juiste volgorde hebt staan en alles wat niet nodig is weggehaald is kan de dataset geëxporteerd worden. Dit gaat via de exportknop rechts bovenin. Je kunt hier voor verschillende export mogelijk heden kiezen afhankelijk van je eigen behoefte.



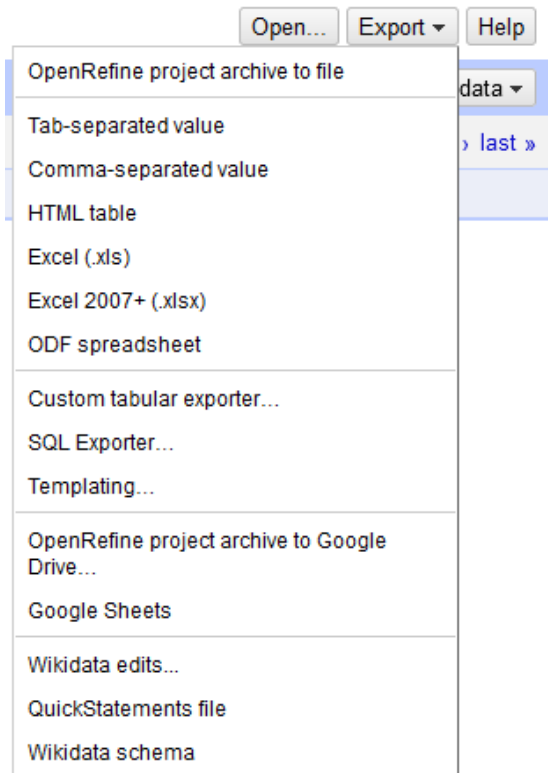


Figure 20 – Export opties

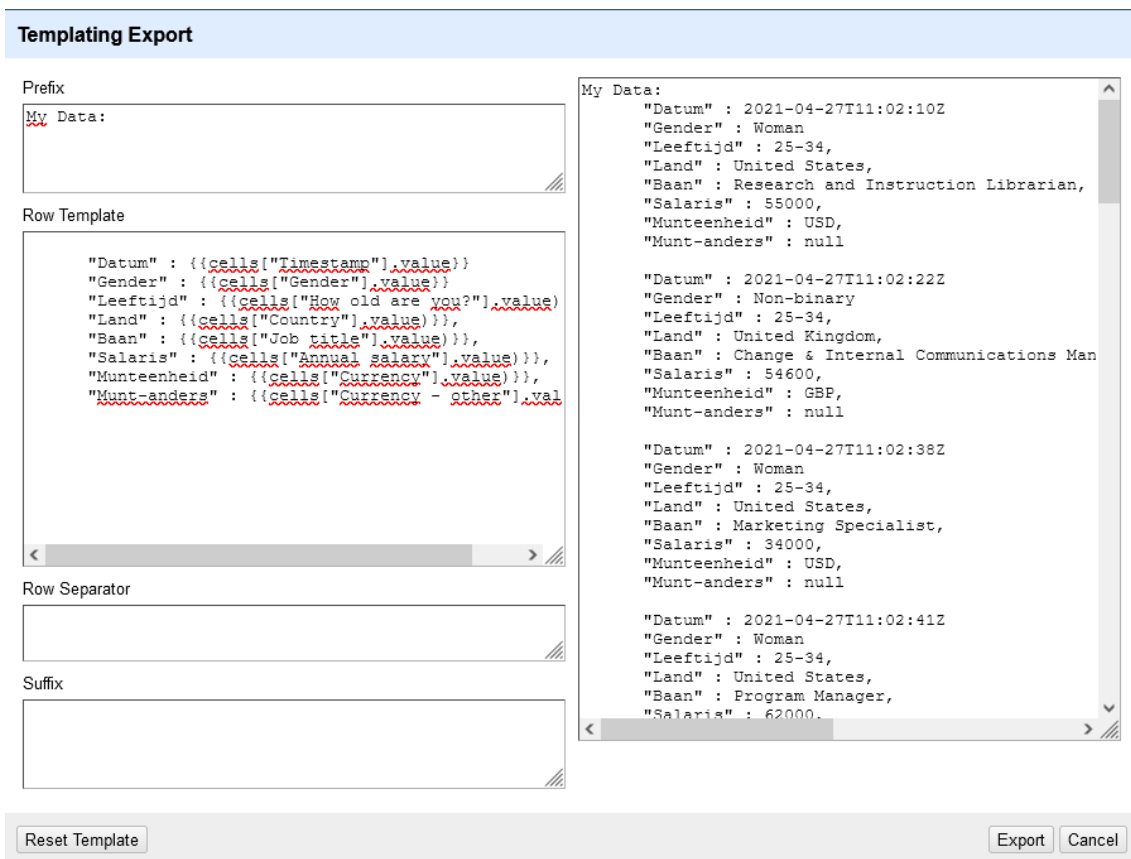


Figure 21 – Template export

Stel dat je een export precies zo wil maken als je zelf wil. Dat kan ook! Als je kiest voor de optie: **Export > Templating...** Dan kom je in het bovenstaande menu terecht. In dit menu kies je hoe je je dataset weer gegeven wilt hebben.

## 6. Verder leren

Wanneer je kennis hebt gemaakt met OpenRefine en enthousiast bent geworden, biedt de tool nog veel meer interessante mogelijkheden. Zoals al eerder aangegeven is de leercurve van deze applicatie erg hoog. Zo kun je bijvoorbeeld nog kolommen splitsen en samenvoegen of de tool gegevens laten aanvullen vanaf een externe website middels een soort webscraping.

## 7. Conclusie

OpenRefine is een zeer krachtig en goed te gebruiken tool voor gegevensanalyse en opschoning van datasets. Zelfs als de gegevens zeer rommelig zijn, een taak waar Excel bijvoorbeeld niet aan te pas komt, gaat dit wel gemakkelijk met OpenRefine.

De grootste uitblinkers van OpenRefine zijn:

- Met gemak in grote datasets werken
- Snel en relatief gemakkelijk rommelige datasets opschonen
- Lokaal werken (voor o.a. privacygevoelige informatie)
- Ondersteund meerdere bestandtypes voor importeren
- Gratis en open-source
- Unieke stappengeschiedenis
- Combineert human control en AI
- Interactieve en visueel aantrekkelijke UI

## Lijst van figuren

Figure 1 – Het creëren van een project.....	7
Figure 2 – Overzicht van de dataset .....	7
Figure 3 – Facet.....	8
Figure 4 – Facet > Tekst facet .....	8
Figure 5 – Cluster & Edit column .....	8
Figure 6 – Cluster & edit column – Ngram-fingerprint .....	9
Figure 7 – Datums filteren – Facet > timeline Facet.....	10
Figure 8 – Datums filteren – Edit cells > Common transforms > to date .....	11
Figure 9 – Datums filteren – Facet > Timeline Facet.....	11
Figure 10 – Nummers omzetten – Facet > Numeric Facet.....	12
Figure 11 – Edit cells > common transforms > to number.....	12
Figure 12 – Resultaat van de transformatie met undo optie.....	12
Figure 13 – visueel overzicht van het resultaat .....	13
Figure 14 – Cifers veranderen .....	13
Figure 15- Expression – waardes vervangen.....	13
Figure 16 – Facet > Numeric Facet.....	14
Figure 17 – Tekst zoeken .....	14
Figure 18 – Edit columns > Re-order/remove columns .....	15
Figure 19 – Re-order/remove columns keuzemenu – drag & drop.....	15
Figure 20 – Export opties .....	16
Figure 21 – Template export .....	16