



Handleiding OpenRefine

Deel 2 — Data aanpassen

Nynke Kuipers—Kubus erfgoed & digitalisering

Oktober 2024



KUBUS.fri
erfgoed & digitalisering

Beeld voorpagina: Pauw, Chara Koson, Publiek domein



Licentie

Op deze handleiding is een Creative Commons Naamsvermelding-licentie van toepassing. Deze publicatie mag worden bewerkt, doorgestuurd, geüpload, gedownload, gekopieerd en anderszins verspreid, zolang de naam van de maker vermeld wordt. De gehele licentietekst is te vinden op: <https://creativecommons.org/licenses/by/4.0/>

Als je deze handleiding wil hergebruiken kun je de volgende methode van naamsvermelding aanhouden:

Handleiding OpenRefine - deel 2 - data bewerken (Nynke Kuipers, Kubus erfgoed & digitalisering)
CCBY



Deze handleiding is tot stand gekomen binnen de Datawerkplaats van Landschap Erfgoed Utrecht.



Inhoudsopgave

Introductie	4
Vorbereiding: export en project maken	5
Edit	6
Cluster en edit	7
Common transforms	11
Spaties.....	11
Datum ISO-standaard	12
Cellen leegmaken.....	14
Common transforms over alle kolommen	16
Transform	17
Links en verwijzingen	18

Introductie

OpenRefine is een open source tool specifiek ontwikkeld om met ‘messy data’, ofwel: rommelige data, te werken. Naast een fijne analytische tool, is OpenRefine ook zeer geschikt om aanpassingen te doen in je data. Er zitten allemaal fijne tooltjes in die makkelijk in het gebruik zijn. Deze handleiding neemt je mee door een aantal basis functies in OpenRefine die je op weg helpen om met je eigen erfgoeddata aan de slag te gaan. Het leert je hoe je OpenRefine kunt inzetten in het **bewerken** van je eigen data.

Deze handleiding gaat verder met het [oefenbestand \(csv\)](#) die in [Handleiding OpenRefine deel 1 - het begin](#) is gebruikt.

Let vooraf goed op: *lang niet alle collectiebeheersystemen kunnen aangepaste data weer importeren. Laat je goed voorlichten door je leverancier over de (on)mogelijkheden.*

Let op: *het opschonen gebeurt buiten je systeem. Aanpassingen die in de tussentijd door collega's gedaan worden, worden bij het importeren van de schone set **overschreven**.*

Let op: *importeren is een onomkeerbare actie. Zorg dat je zeker weet dat wat je doet. Je kunt het proces ook eerst op drie records **uitproberen**: mocht er iets misgaan, kan je dit makkelijk terugdraaien / wijzigen.*

Tip vooraf: *Maak een **back-up** van je data voordat je ermee aan de slag gaat. Bekijk de officiële handleiding hoe dat aan te pakken. Mocht er dan toch iets fout gaan, kan je de data terugplaatsen. Raadpleeg de gebruikershandleiding van je collectiebeheersysteem of neem contact op met je leverancier.*

In deze handleiding is gewerkt met [OpenRefine Version 3.8.2](#). Bij nieuwere of oudere versies kan het er net iets anders uitzien dan getoond in de screenshots van deze handleiding.

Deze handleiding is ontwikkeld naar aanleiding van workshops die gegeven zijn in het kader van de Datawerkplaats Utrecht, een initiatief van Landschap Erfgoed Utrecht en Netwerk Digitaal Erfgoed.

Voor uitgebreide instructies over het gebruik van OpenRefine is er de [user manual](#) (engelstalig) van OpenRefine zelf.

Voorbereiding: export en project maken

Voordat je aan de slag kunt in OpenRefine ben je een export (ook wel rapport genoemd) nodig uit je collectiebeheersysteem. Kijk in de handleiding van je CBS hoe je dit moet doen.

Zoals al eerder genoemd: ga bij voorkeur per veld of veldgroep aan de slag. Hierdoor hou je de doorlooptijd van je OpenRefine werk zo overzichtelijk mogelijk.

Je komt in ieder geval de volgende stappen tegen:

1. Selecteer een bestandsformaat.

Let goed op als je met herhaalbare velden (occurrences) werkt: niet alle exportformaten nemen alle herhaalbare velden mee.

- a. CSV (Comma Seperated Values)
- b. TSV (Tab Seperated Values)
- c. XML-unstructured
- d. XML-grouped
- e. Tagged / adlib-tagged (alleen Adlib/Axiell collections)

2. Velden selecteren

- a. Neem altijd een unieke identificatie mee: recordnummer of
- b. Kies het veld (of de veldgroep) waar je wijzigingen op wilt toepassen

3. Exporteer de data naar een bestand

4. Maak een nieuw project aan in OpenRefine met je exportbestand. Zie [deel 1](#) van de handleidingenserie voor het uitgebreide stappenplan.

Edit

De simpelste manier om data aan te passen is met **edit**. Dit knopje verschijnt als je over de cellen in de tabel beweegt met je muis.

1. Beweeg met je muis over de cel die je wilt aanpassen > klik op **edit**

▼ Objectnummer	▼ Vervaardiger	▼ Soort object
ON0003.50001	Rembrandt edit	schilderij
ON0003.50002	Rembrandt	ets
ON0003.50003	Rembrandt	ets

2. Je kunt nu tekst typen, aanpassen of weghalen

Data type: text ▼

Rembrandt de oude **2**

Apply **Apply to all identical cells** Cancel

Enter **3** Ctrl-Enter **4** Esc

3. Klik op **Apply** op alleen deze cel aan te passen

▼ Objectnummer	▼ Vervaardiger	▼ Soort object
ON0003.50001	Rembrandt de oude	schilderij
ON0003.50002	Rembrandt	ets
ON0003.50003	Rembrandt	ets

4. Klik op **Apply to all identical cells** om alle cellen waar deze exacte waarde in staat aan te passen

▼ Objectnummer	▼ Vervaardiger	▼ Soort object
ON0003.50001	Rembrandt de oude	schilderij
ON0003.50002	Rembrandt de oude	ets
ON0003.50003	Rembrandt de oude	ets

Cluster en edit

Bij grote lijsten van bijvoorbeeld namen is er een handige tool die veel werk voor je kan doen: cluster en edit. Hiermee worden termen die erg op elkaar lijken samengevoegd. Je kan zelf aan de knoppen draaien, dus er is altijd een manier waarop je tot goede en correcte resultaten komt.

1. We gaan aan de slag met de vervaardigers. Dit zijn de waardes die voorkomen in deze kolom:

Vervaardiger

change

7 choices Sort by: name count

Cluster

Appel, Karel 13

Rembrandt 19

Rembrandt (etser) 13

Rembrandt Harmensz. van Rijn (1606-1669) 11

Rembrandt van Rijn 8

Rhijn, Rembrandt Harmensz. van 12

Rijn, Rembrandt van 13

(blank) 60

2. Klik op het pijltje voor de naam van de kolom waar je mee aan de slag wilt > **Edit cells > Cluster and edit...**

Vervaardiger	Soort object	Materiaal	
Facet	derij	doek olieverf	19-j
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...		

1

3. Klik op **Cluster** in het midden van je scherm

Cluster and edit column "Vervaardiger"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method Key collision

Keying function Fingerprint

☐ Auto-update

Click Cluster to find clusters on column "Vervaardiger" using the parameters above. Cluster

2

4. Je ziet nu welke clusters gevormd zijn

Cluster and edit column "Vervaardiger"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method Key collision

Keying function Fingerprint

☐ Auto-update

1 cluster found

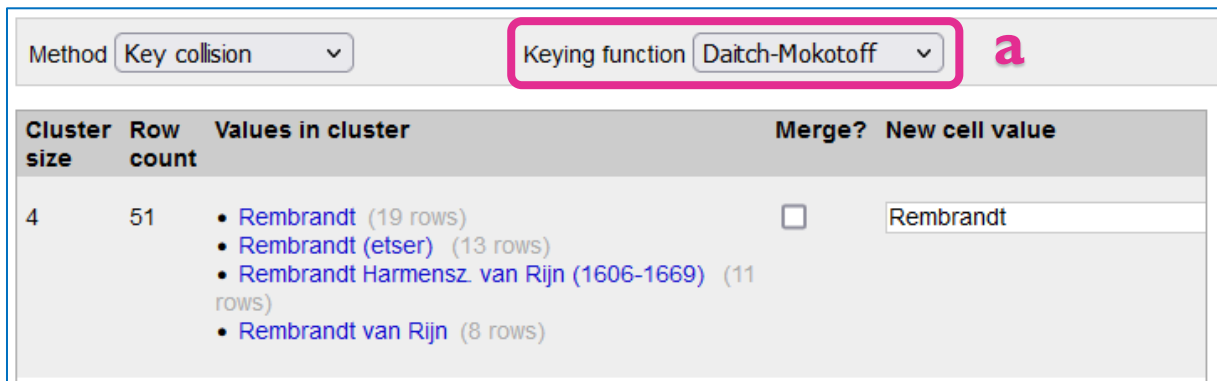
Cluster size	Row count	Values in cluster	Merge?	New cell value
2	21	- Rijn, Rembrandt van (13 rows) - Rembrandt van Rijn (8 rows)	☐	Rijn, Rembrandt van

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

3

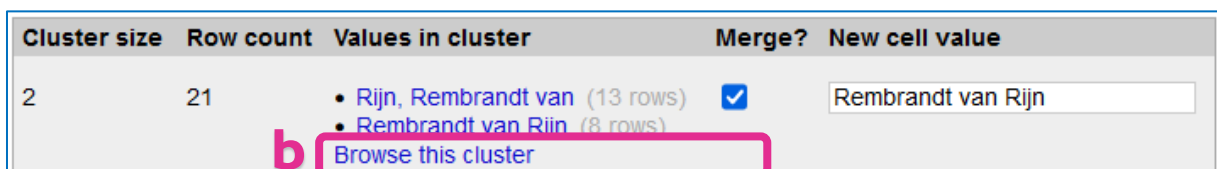
5. Aanpassingen doen en resultaten bekijken

- a. Je kunt experimenteren met welke **Keying function** je gebruikt



Cluster size	Row count	Values in cluster	Merge?	New cell value
4	51	- Rembrandt (19 rows) - Rembrandt (etser) (13 rows) - Rembrandt Harmensz. van Rijn (1606-1669) (11 rows) - Rembrandt van Rijn (8 rows)	☐	Rembrandt

- b. Met mouse-over verschijnt een optie om alle records die in dit cluster zijn samengebracht te bekijken > **Browse this cluster**

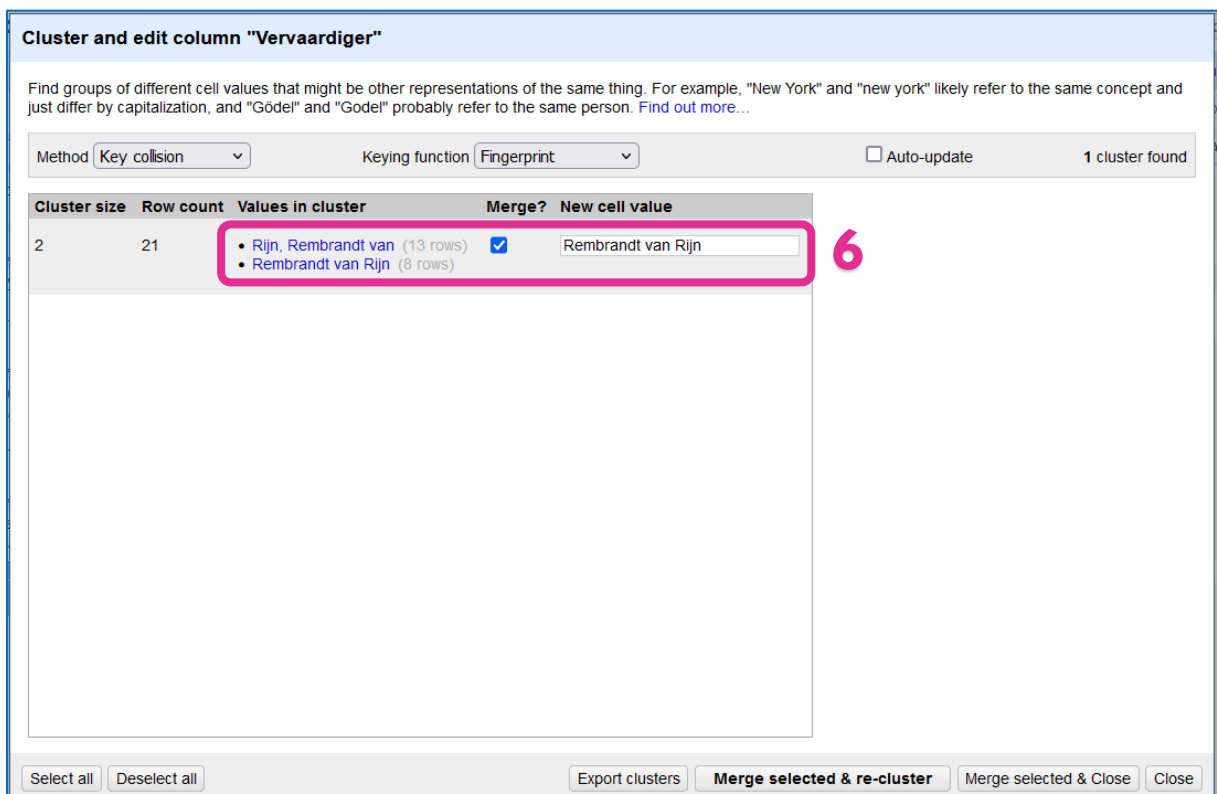


Cluster size	Row count	Values in cluster	Merge?	New cell value
2	21	- Rijn, Rembrandt van (13 rows) - Rembrandt van Rijn (8 rows)	☒	Rembrandt van Rijn

b [Browse this cluster](#)

6. Ben je tevreden met de resultaten?

Klik op de naam die je wilt gebruiken voor alle cellen waar deze naam voorkomt. Er wordt meteen een vinkje aangezet als je dit doet



Cluster and edit column "Vervaardiger"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

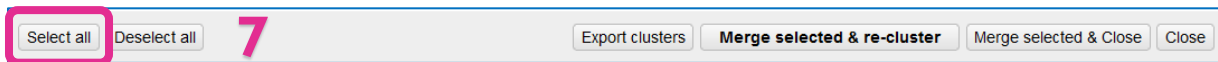
Method: Key collision Keying function: Fingerprint ☐ Auto-update 1 cluster found

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	21	- Rijn, Rembrandt van (13 rows) - Rembrandt van Rijn (8 rows)	☒	Rembrandt van Rijn

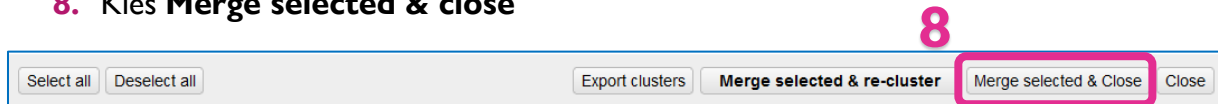
6

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

7. De clusters met een **vinkje** worden uitgevoerd. Je kunt ook alle clusters selecteren



8. Kies **Merge selected & close**



9. In de **facet** op vervaardiger zie je dat de aanpassing is doorgevoerd.



Common transforms

Er zijn meerdere transformaties uit te voeren met simpele functies. We lopen een aantal veelvoorkomende langs.

Spaties

Soort object

change

22 choices Sort by: name count

Cluster

lithografie 1

olieverfschilderij 1

penning 1

prent 2

prent 3

prent 1

prenten 4

schilderij 5

schildering 2

(blank) 103

Facet by choice counts

Soort object

change

18 choices Sort by: name count

Cluster

gravure|prent 1

huishoudschort 1

lithografie 1

olieverfschilderij 1

penning 1

prent 8

prenten 4

schilderij 5

schildering 2

(blank) 103

Facet by choice counts

Bij zijn meerdere spaties zichtbaar. Dit gaan we heel simpel, met één actie oplossen.

1. Klik op het pijltje voor de kolomnaam > **Edit cells** > **Common transforms** > **Trim leading and trailing whitespace**

Soort object	Materiaal	Datering	Afmeting type	Afmeting waarde	Afmeting e	
Facet	▶	olieverf	19-jun-04	hoogte	10.00	mm
Text filter			breedte	225.00	mm	
			diepte	525.00	mm	
Edit cells	▶	Transform...				
Edit column	▶	Common transforms		▶ Trim leading and trailing whitespace		
Transpose	▶	Fill down		Collapse consecutive whitespace		
Sort...		Blank down		Unescape HTML entities		
View	▶	Split multi-valued cells...		Replace smart quotes with ASCII		
Reconcile	▶	Join multi-valued cells...		To titlecase		
		Cluster and edit...		To uppercase		
		Replace...		To lowercase		
			breedte	To number		
			diepte	To date		
prent				To text		
prent		1620-1669	hoogte	To null		
				To empty string		

Datum ISO-standaard

De datumnotatie kan met één actie naar de ISO-standaard worden omgezet: jiji-mm-dd.

Datering		change
41 choices Sort by: name count		Cluster
20-7-1988	1	
21-10-1884	1	
28-5-2021	1	
28-sep-00	1	
31 april 1752	1	
5-apr-80	1	
7 augustus 1652	1	
7-aug-82	1	
dated 1631	1	
(blank)	106	
Facet by choice counts		

1. Klik op het pijltje voor de kolom met een datumnotatie > Edit cells > Common transforms > To date

Datering	Afmeting type	Afmeting waarde	Afmeting eenheid	Titel	Beschr
Facet		10.00	mm		
Text filter		225.00	mm		
		525.00	mm		
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...			
	breedte	7.50			
	diepte	8.80			
1620-1669	hoogte	17.90			

Trim leading and trailing whitespace

Collapse consecutive whitespace

Unescape HTML entities

Replace smart quotes with ASCII

To titlecase

To uppercase

To lowercase

To number

To date

To text

To null

To empty string

2. Je ziet nu dat de datums in de kolom groen worden en volgens de ISO-norm worden weergegeven. De tijd wordt als T00:00:00 weergegeven omdat in dit oefenbestand geen tijd is opgenomen. Ontbrekende maanden en dagen worden automatisch 01. Met andere woorden, als er enkel een jaartal is ingevuld wordt dit automatisch 1 januari van het betreffende jaar, wel een dingetje om rekening mee te houden. Sommige datums zijn niet groen geworden. OpenRefine heeft deze niet herkend als datum, bijvoorbeeld omdat er twee jaartallen zijn ingevoerd of de tekst 'circa' in het veld staat.

▼ All	▼ Objectnummer	▼ Vervaardiger	▼ Soort object	▼ Materiaal	▼ Datering	▼ Afmeting type
☆	1.	ON0003.50001	Rembrandt	schilderij	doek olieverf	2004-06-19T00:00:00Z hoogte
☆						breedte
☆						diepte
☆	2.	ON0003.50002	Rembrandt	ets	1632-01-01T00:00:00Z	
☆	3.	ON0003.50003	Rembrandt	ets	1831-01-01T00:00:00Z	breedte
☆	4.	ON0003.50004	Rembrandt	ets	1767-01-01T00:00:00Z	hoogte
☆						breedte
☆	5.	ON0003.50005	Rembrandt	schilderij	olieverf paneel	1870-11-08T00:00:00Z hoogte
☆						breedte
☆						diepte
☆	6.	ON0003.50006	Rembrandt	prent		
☆	7.	ON0003.50007	Rembrandt	prent	1620-1669	hoogte

Cellen leegmaken

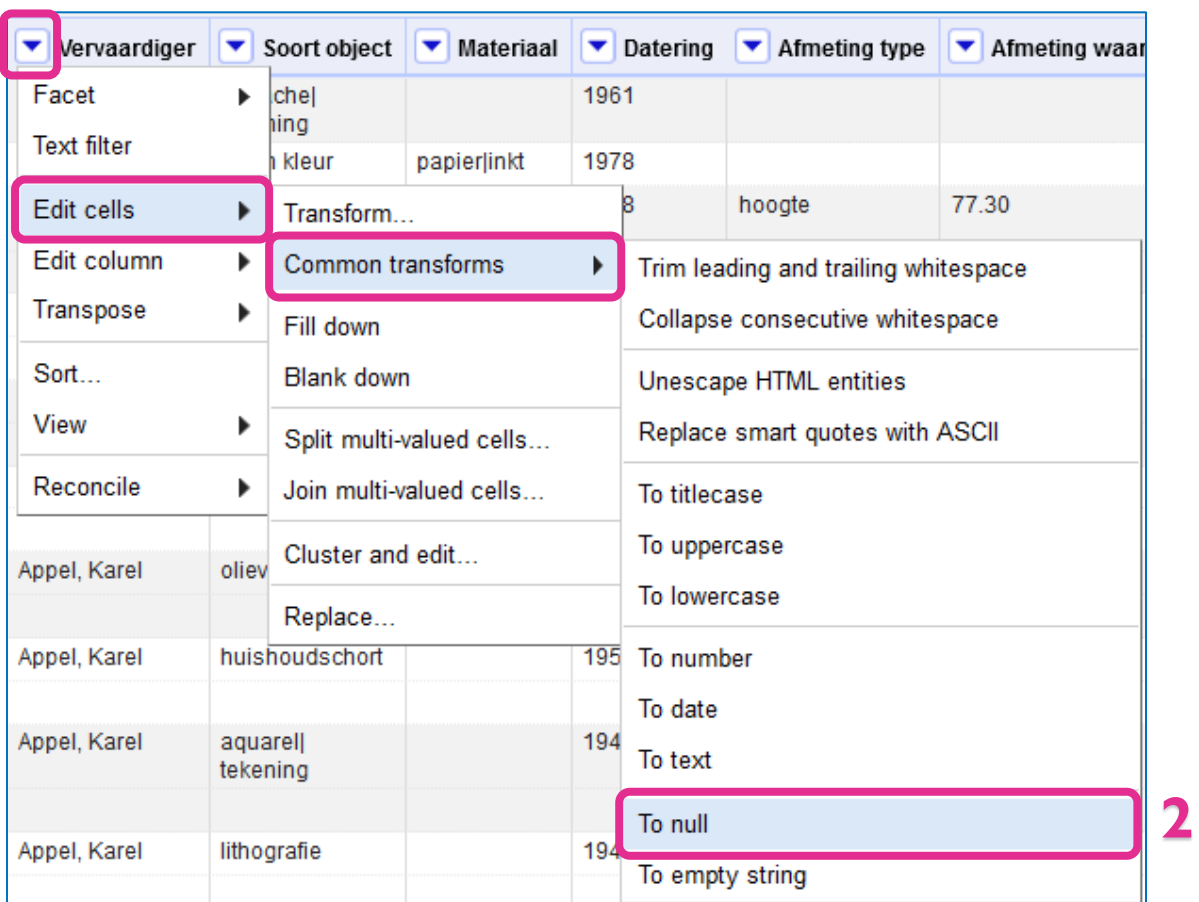
Soms wil je graag cellen leegmaken zonder dat je alle rijen langs hoeft. In dit voorbeeld gaan we alle cellen met **Karel Appel** leegmaken.

1. Zoek op **facet** en klik **Appel, Karel** aan



The screenshot shows a data application interface. On the left, there is a 'Facet / Filter' panel for 'Vervaardiger' (Artist). It lists 6 choices, sorted by name count. 'Appel, Karel' is selected and highlighted with a red box. A red number '1' is placed next to the selection. The main panel shows '13 matching records (98 total)'. The table has columns: Objectnummer, Vervaardiger, Soort object, Materiaal, and Datering. The first few rows show records for 'Appel, Karel'.

2. Klik op het pijltje voor de kolomnaam **Vervaardiger** > **Edit cells** > **Common transforms** > **To null**



The screenshot shows the 'Vervaardiger' column dropdown menu. The path 'Edit cells' > 'Common transforms' > 'To null' is highlighted with red boxes. A red number '2' is placed next to the 'To null' option. The menu also shows other options like 'Facet', 'Text filter', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', 'Cluster and edit...', and 'Replace...'.

3. Karel Appel komt nu niet meer in de lijst voor, terwijl alle vermeldingen van Rembrandt zijn blijven staan. Het aantal records met (blank) is groter geworden

Facet / Filter

Undo / Redo 10 / 10

98 records

Refresh

Reset all

Remove all

Vervaaardiger

change

5 choices Sort by: name count

Cluster

Rembrandt 19

Rembrandt (etser) 13

Rembrandt Harmensz. van Rijn (1606-1669) 11

Rembrandt van Rijn 21

Rhijn, Rembrandt Harmensz. van 12

(blank) 73

Facet by choice counts

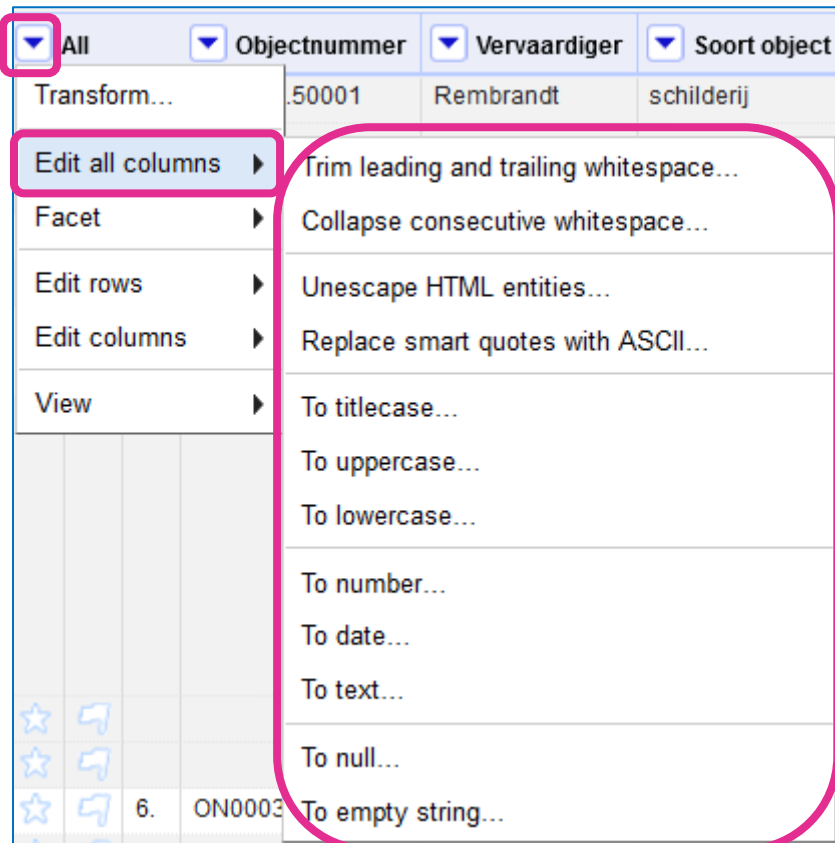
Show as: rows records

Show: 5 10 25 50 100 500 1000 records

All	Objectnummer	Vervaaardiger	Soort object	Materiaal	Datering	
☆	1.	ON0003.50001	Rembrandt	schilderij	doek olieverf	19-jun-04
☆						
☆						
☆	2.	ON0003.50002	Rembrandt	ets		1632
☆	3.	ON0003.50003	Rembrandt	ets		1831
☆	4.	ON0003.50004	Rembrandt	ets		1767
☆						
☆	5.	ON0003.50005	Rembrandt	schilderij	olieverf paneel	11/8/1870

Common transforms over alle kolommen

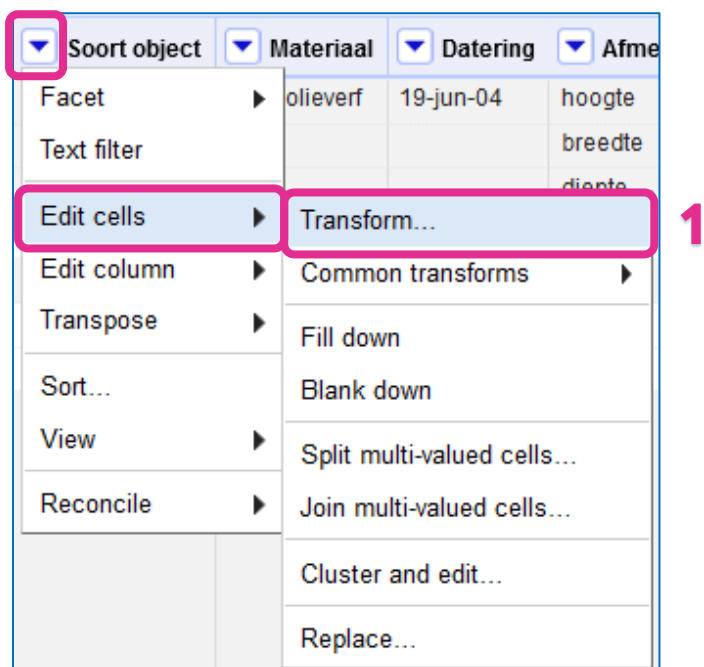
OpenRefine biedt ook de mogelijkheid deze Common transforms over meerdere kolommen tegelijk uit te voeren. **All > Edit all columns > ...**



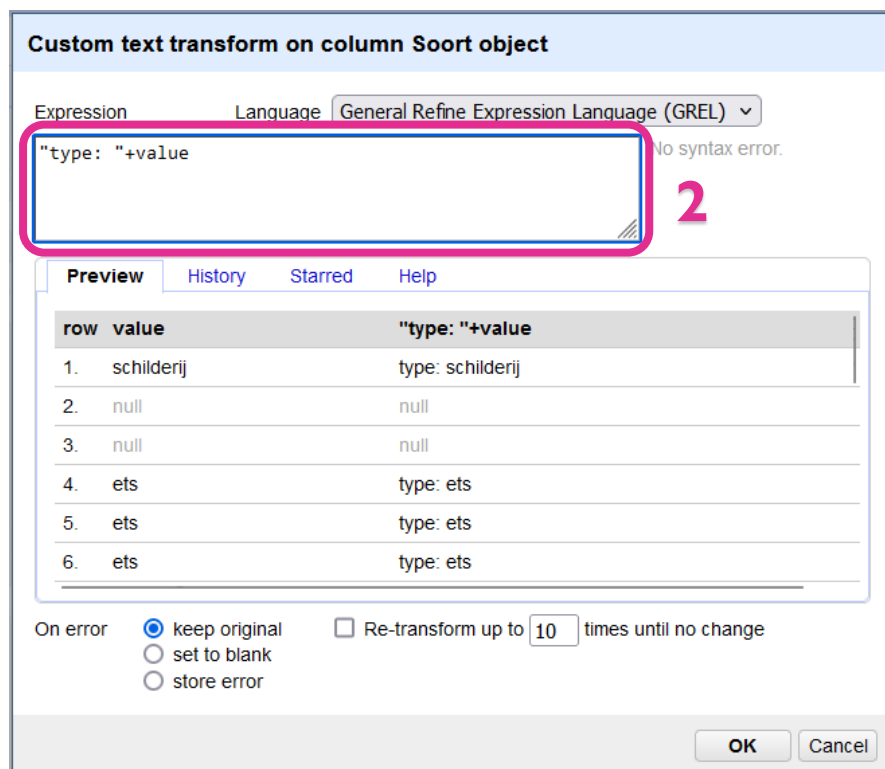
Transform

Met de functie **Transform** kan je bijvoorbeeld in één keer tekst toevoegen aan de cellen in een kolom. De bestaande data kan je hierbij behouden.

1. Klik op het pijltje voor de kolom **soort object** > **Edit cells** > **Transform...**



2. De tekst, inclusief leestekens en spaties, type je tussen dubbele aanhalingstekens gevold door een plusteken + om de data samen te voegen **"type: "+value**



Links en verwijzingen

- OpenRefine user manual: <https://openrefine.org/docs>
- OpenRefine forum: <https://forum.openrefine.org/>
- OpenRefine github: <https://github.com/OpenRefine/>
- Introductiecursus OpenRefine: <https://digitaalerfgoedcoach.online/online-leren-posts/openrefine/>
- Kennisbank Meemoo: <https://kennisbank.meemoo.be/toolbox/open-refine-handleiding-voor-cultureel-erfgoed-collecties>
- Netwerk Digitaal Erfgoed: <https://netwerkdigitaalerfgoed.nl/>
- Datawerkplaatsen: <https://netwerkdigitaalerfgoed.nl/activiteiten/datawerkplaatsen/>
- Landschap Erfgoed Utrecht: <https://www.landschaperfgoedutrecht.nl/>
- <https://www.landschaperfgoedutrecht.nl/erfgoed/digitalisering/>

Alle handleidingen zijn te vinden op: www.kubus.frl/openrefine

