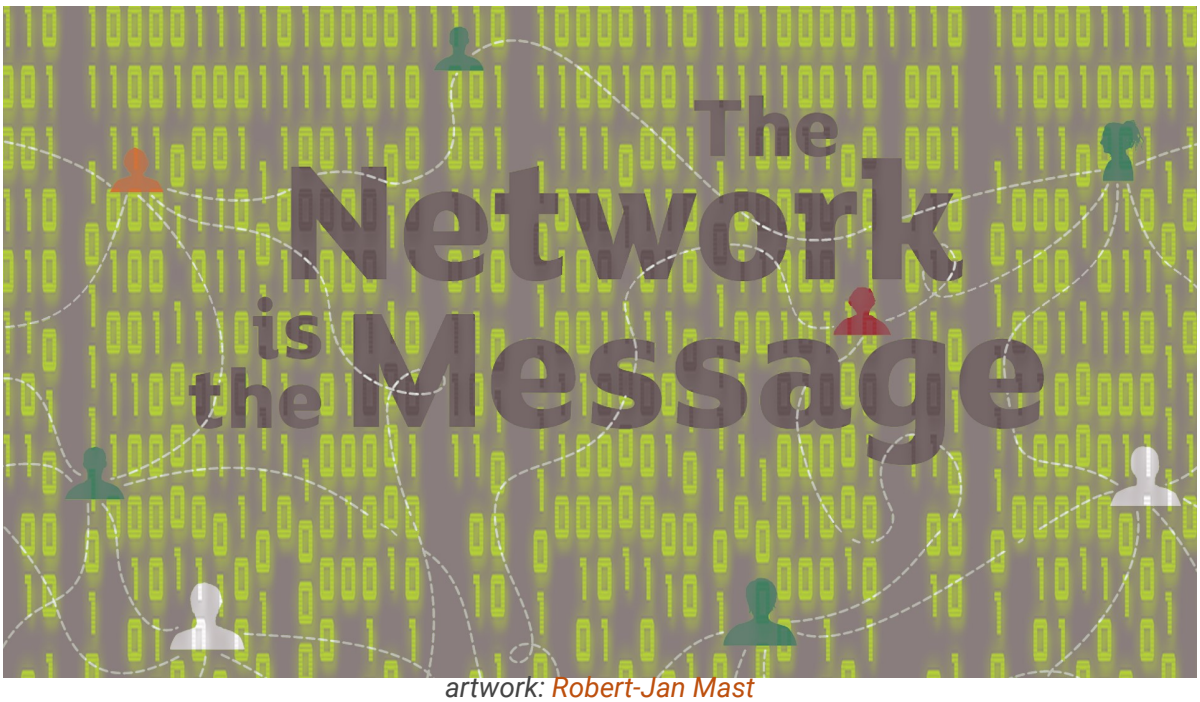
DOE-HET-ZELF: WEBDATA KOPIËREN, THE LAZY WAY

28 maart 2017 redactie No Comments



artwork: Robert-Jan Mast

Microsoft oprichter Bill Gates zou ooit eens het volgende gezegd hebben: "I choose a lazy person to do a hard job, because a lazy person will find an easy way to do it.". Hoewel het niet zeker is of hij dit ooit gezegd heeft, spreekt het velen wel tot de verbeelding. In deze tutorial leggen we uit hoe je -the lazy way- data van een webpagina kan kopiëren of scrapen, zoals ook wel eens genoemd wordt.

Dit blog is wat anders dan je gewend bent. Het is een zogenoemde DIY (Do it Yourself) of tutorial. Door middel van een stappenplan met afbeeldingen gaan we je proberen wegwijs te maken in het kopiëren van webdata, iets wat een belangrijke rol

speelt in het project. In het project The Network is the Message doen we onderzoek naar inhakers. Online zijn allerlei kalenders te vinden met momenten en onderwerpen waarop ingehaakt kan worden. Voor ons zijn die momenten een mooi uitgangspunt om onderzoek te doen. Wij willen die data dus graag hebben. Nu kan je dat doen door een kalender op internet te ‘copy-pasten’, maar het kost veel werk om het vervolgens te structureren (dat wil zeggen dat je een net lijstje met alle dagen van het jaar met daarachter de inhaakonderwerpen). Daarom volgt hier opgesteld door Thijs Waardenburg een tutorial voor de lazy way! In dit geval voor Inhaakkalender.com, maar dit principe is natuurlijk op andere websites toe te passen.

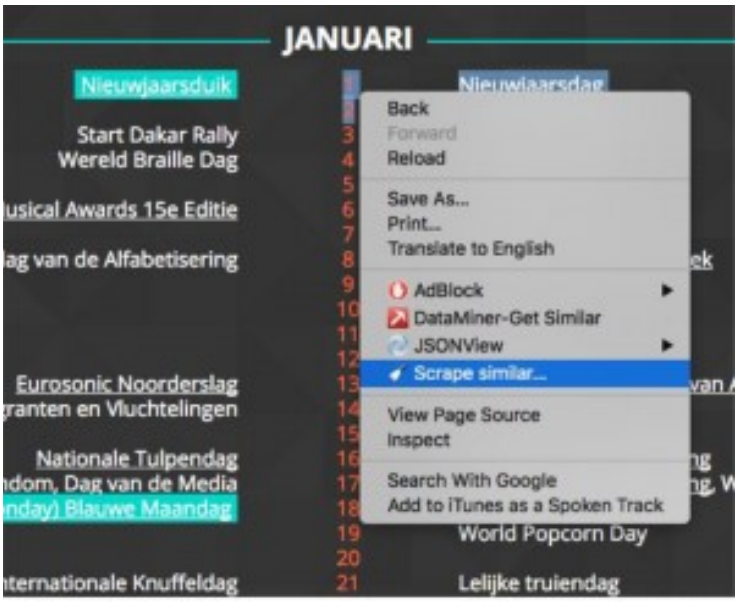
#HOEDAN?

HTML-webpagina’s zijn van nature al gestructureerd opgebouwd. Deze structuur is vastgelegd in het zogenaamde Document Object Model (DOM). Wanneer we bepaalde gegevens van een webpagina af willen halen kunnen we het in de meeste gevallen natuurlijk kopiëren en plakken in een ander programma (zoals Excel). Dat is geen probleem als het om één of een

paar items gaat. Het wordt vaak wel een probleem als het om, zeg, 50, 500 of 5000 items gaat. Hiervoor zijn een aantal handige software tools beschikbaar. Eén van de vele tools voor het scrapen van data is Scraper. Dit is een zogenaamde extension voor de Google Chrome browser. Met Scraper kan je stukje tekst/data in een webpagina selecteren. Vervolgens zal de software op zoek gaan in de pagina naar vergelijkbare data(structuren). Deze data kan vervolgens geëxporteerd of gekopieerd worden naar een ander programma. De data moet meestal daarna nog opgeschoond, aangevuld of van structuur veranderd worden. Dat kan onder meer worden gedaan met het gratis programma OpenRefine . Hieronder volgen de stappen die genomen kunnen worden om zogenaamde 'inhakers' te scrapen van de website inhaakkalender.com.

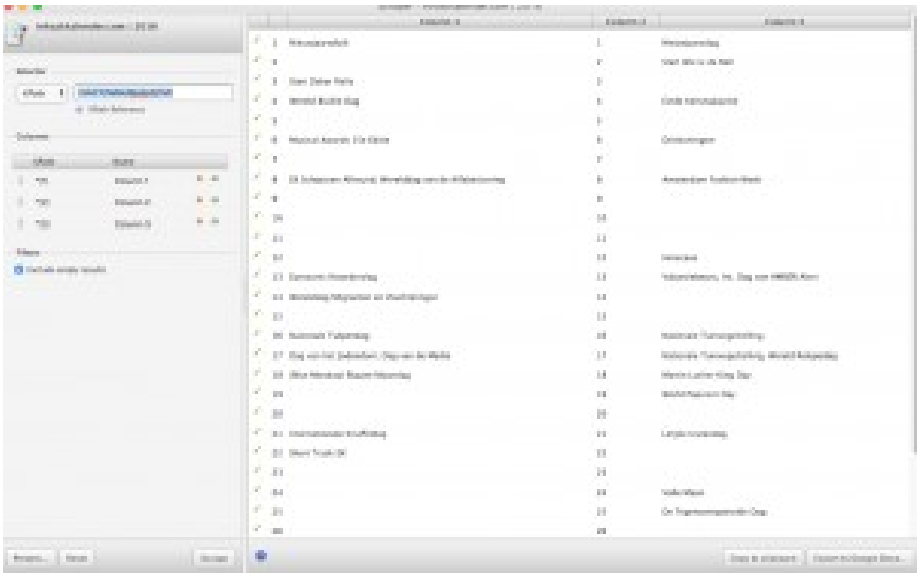
STAP 1

Selecteer de bovenste regel met inhaakonderwerpen (1 januari: Nieuwjaarsduik, Nieuwjaarsdag). Rechtermuisknop en selecteer 'Scrape similar'.



STAP 2

Hiermee selecteer je de data van 1 maand. Kopieer de data naar het clipboard (knop rechtsonder “Copy to clipboard”).



STAP 3

Plak de data in een Excel tabel.

	A	B	C
1	Column 1	Column 2	Column 3
2	Nieuwjaarsd	1	Nieuwjaarsdag
3		2	Start Wie is de Mol
4	Start Dakar R	3	
5	Wereld Brail	4	Einde Kerstvakantie
6		5	
7	Musical Awa	6	Driekoningen
8		7	
9	EK Schaatser	8	Amsterdam Fashion Week
10		9	
11		10	
12		11	
13		12	Horecava
14	Eurosonic Nc	13	Vakantiebeurs, Int. Dag van AMBER Alert
15	Werelddag N	14	
16		15	
17	Nationale Tu	16	Nationale Tuinvogeltelling
18	Dag van het .	17	Nationale Tuinvogeltelling, Wereld Religiedag
19	(Blue Monda	18	Martin Luther King Day
20		19	World Popcorn Day
21		20	
22	International	21	Lelijke truiendag
23	Short Track E	22	
24		23	
25		24	Volle Maan
26		25	De Tegenovergestelde Dag
27		26	

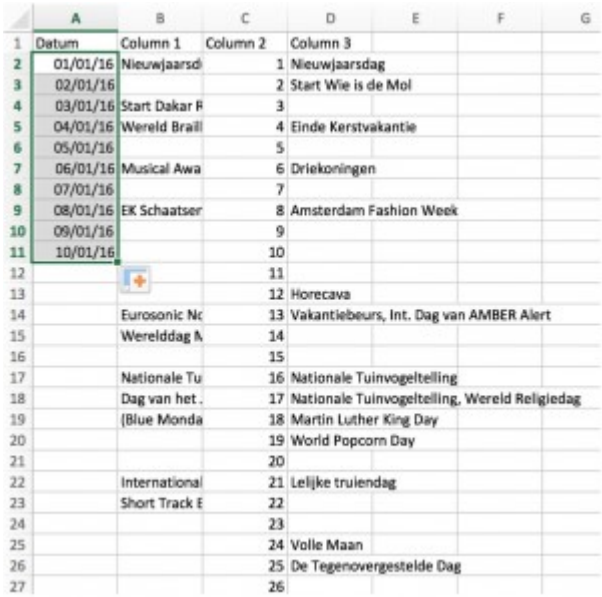
STAP 4

Herhaal stap 2 en 3 voor alle maanden; plak ze onder elkaar in dezelfde tabel. Verwijder de rijen met kolomnamen (in dit voorbeeld rij 33).

23	Short Track E	22	
24		23	
25		24	Volle Maan
26		25	De Tegenovergestelde Dag
27		26	
28	International	27	Hartjesdag, Nationale Voorleesdagen
29	Europese Da	28	Nationale Gedichtendag, Start Poëzieweek
30		29	Nationale Gedichtendag
31	rde Ant	30	Dutch Open Darts
32	eld Lepr	31	Verjaardag prinses Beatrix
33	Column 1	Column 2	Column 3
34	Collecte Her	1	
35		2	
36		3	Afsluiting Poëzieweek
37	Wereld Kank	4	International Friends Day
38		5	Warme Truiendag
39		6	
40	Carnaval	7	
41	Chinees Nie	8	Carnaval
42	Safer Interne	9	Carnaval, Gelukkige Dag van de Chocolade
43	Aswoensdag	10	
44	WK Afstande	11	
45	Charles Darw	12	
46	Wereld Radi	13	
47	Valentijn	14	
48	Week van de	15	President's Day (VS)
49	Int. Pizza dag	16	
50	Int. Doe Vrie	17	
51	Motorbeurs	18	
52		19	

STAP 5

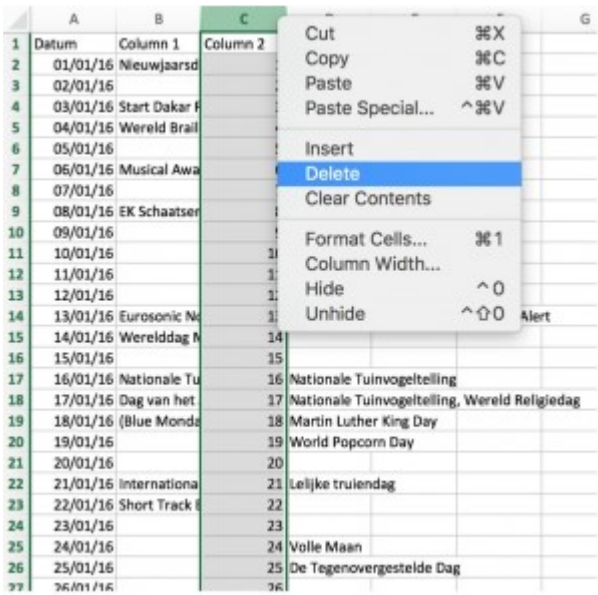
Voeg een ‘Datum’ kolom toe en vul deze vanaf 01-01-2016



	A	B	C	D
1	Datum	Column 1	Column 2	Column 3
2	01/01/16	Nieuwjaarsd	1	Nieuwjaarsdag
3	02/01/16		2	Start Wie is de Mol
4	03/01/16	Start Dakar F	3	
5	04/01/16	Wereld Brail	4	Einde Kerstvakantie
6	05/01/16		5	
7	06/01/16	Musical Awa	6	Driekoningen
8	07/01/16		7	
9	08/01/16	EK Schaatser	8	Amsterdam Fashion Week
10	09/01/16		9	
11	10/01/16		10	
12			11	
13			12	Horecava
14		Eurosonic Nc	13	Vakantiebeurs, Int. Dag van AMBER Alert
15		Werelddag N	14	
16			15	
17		Nationale Tu	16	Nationale Tuinvogeltelling
18		Dag van het .	17	Nationale Tuinvogeltelling, Wereld Religiedag
19		(Blue Monda	18	Martin Luther King Day
20			19	World Popcorn Day
21			20	
22		International	21	Lelijke truiendag
23		Short Track E	22	
24			23	
25			24	Volle Maan
26			25	De Tegenovergestelde Dag
27			26	

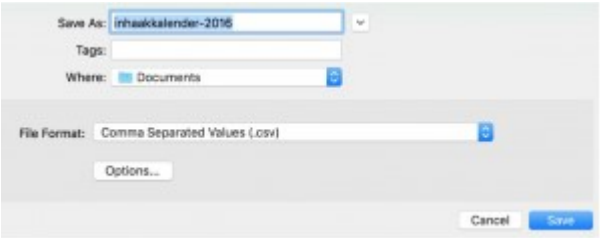
STAP 6

Verwijder ‘Column 2’ met de dagnummers (nadat je gecontroleerd hebt of de datum kolom juist is).



STAP 7

Sla het op als CSV-bestand (*comma seperated values*).

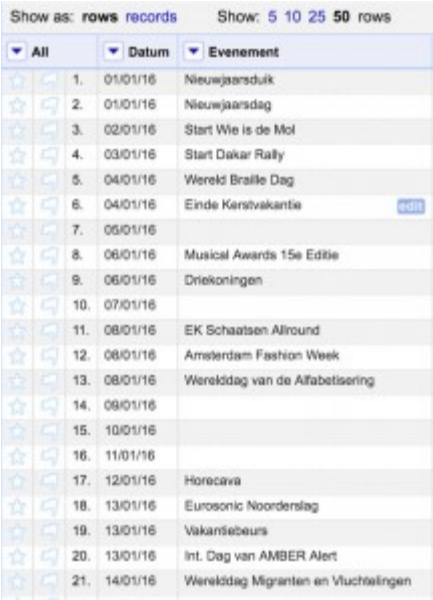


STAP 8

Importeer het bestand in OpenRefine.

STAP 9

Op sommige dagen staan meerdere *events*. Deze moeten gesplitst worden.

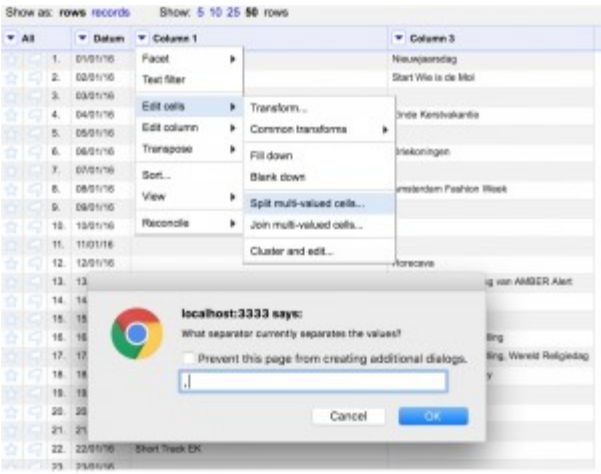


Show as: **rows** records Show: 5 10 25 **50** rows

All		Datum	Evenement
	1.	01/01/16	Nieuwjaarsduik
	2.	01/01/16	Nieuwjaarsdag
	3.	02/01/16	Start Wie is de Mol
	4.	03/01/16	Start Dakar Rally
	5.	04/01/16	Wereld Braille Dag
	6.	04/01/16	Einde Kerstvakantie edit
	7.	05/01/16	
	8.	06/01/16	Musical Awards 15e Editie
	9.	06/01/16	Driekoningen
	10.	07/01/16	
	11.	08/01/16	EK Schaatsen Allround
	12.	08/01/16	Amsterdam Fashion Week
	13.	08/01/16	Werelddag van de Alfabetisering
	14.	09/01/16	
	15.	10/01/16	
	16.	11/01/16	
	17.	12/01/16	Horecava
	18.	13/01/16	Eurosonic Noorderslag
	19.	13/01/16	Vakantiebeurs
	20.	13/01/16	Int. Dag van AMBER Alert
	21.	14/01/16	Werelddag Migranten en Vluchtelingen

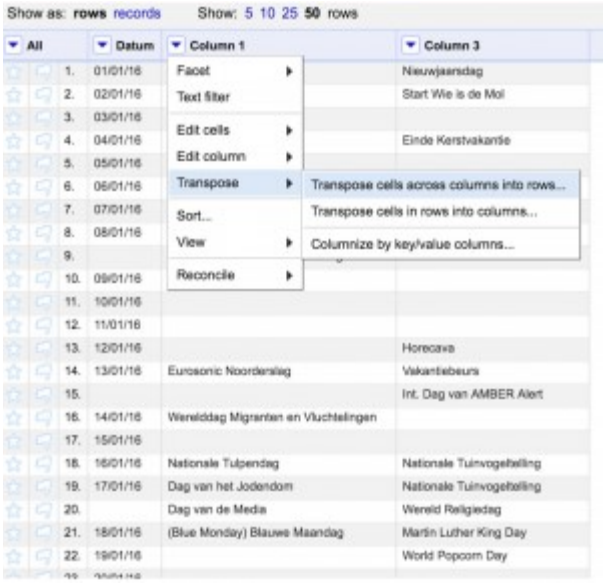
STAP 10

Kies het dropdownmenu van *Column 1 -> Edit cells -> Split multi-valued cells*. In dit geval zijn de events gescheiden (separator) met een komma + spatie. Herhaal deze stap voor Column 3.



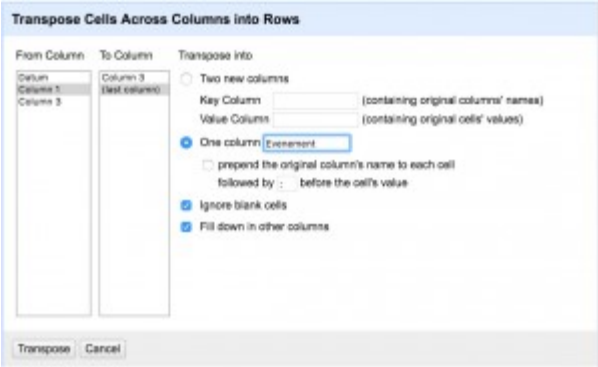
STAP 11

Zet vervolgens alle evenementen onder elkaar in één kolom *Transpose -> Transpose cells across columns into rows.*



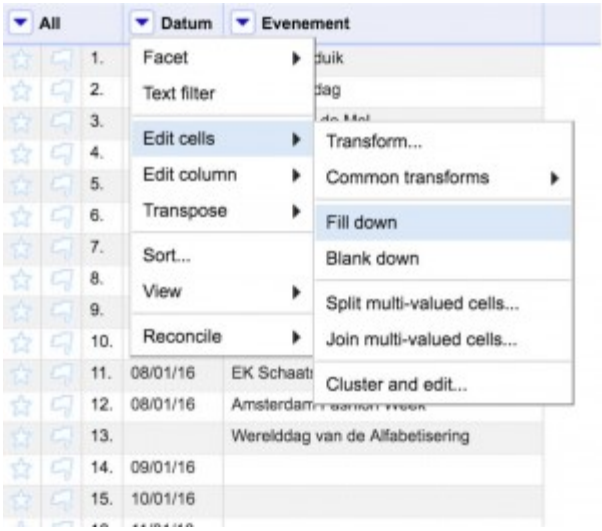
STAP 12

Neem voor het dialoogvenster van stap 11 de bovenstaande instellingen over om alles onder elkaar te zetten.



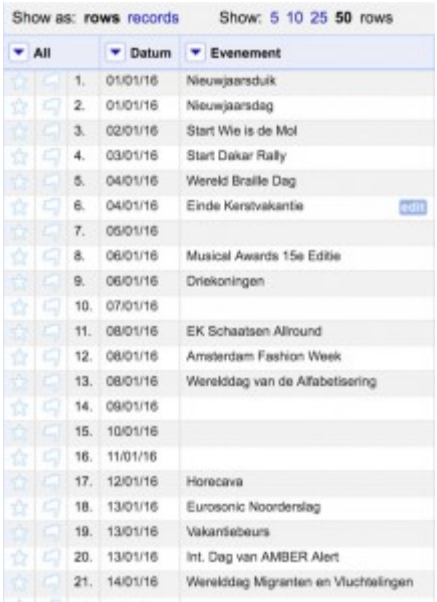
STAP 13

Vul de lege datumvelden in *Edit cells -> Fill Down.*



STAP 14

Klaar! Exporteer de data naar gewenst formaat.



Hoera! Je hebt zojuist data gekopieerd binnen 14 stappen! Wij zijn benieuwd of het gelukt is. Had je problemen, vragen of verliep alles vlekkeloos? Stuur ons dan een berichtje.

Your name *