

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

	A	B
1	Datum	Evenement
2	01/01/16	Nieuwjaarsdag
3	02/01/16	Nieuwjaarsdag
4	03/01/16	Start Nike in de Mol
5	04/01/16	Start Doping, Reky
6	05/01/16	Wereldleidschap
7	06/01/16	Ende Germaankanten
8	07/01/16	
9	08/01/16	Musical Awards De Edele
10	09/01/16	Orkestreringen
11	10/01/16	
12	11/01/16	Orkestreringen
13	12/01/16	Orkestreringen
14	13/01/16	Wereldleidschap van de Klukenkanten
15	14/01/16	Amsterdam Fashion Week
16	15/01/16	
17	16/01/16	
18	17/01/16	
19	18/01/16	
20	19/01/16	
21	20/01/16	
22	21/01/16	
23	22/01/16	
24	23/01/16	
25	24/01/16	
26	25/01/16	
27	26/01/16	
28	27/01/16	
29	28/01/16	
30	29/01/16	
31	30/01/16	
32	31/01/16	
33	01/02/16	
34	02/02/16	
35	03/02/16	
36	04/02/16	
37	05/02/16	
38	06/02/16	
39	07/02/16	
40	08/02/16	
41	09/02/16	
42	10/02/16	
43	11/02/16	
44	12/02/16	

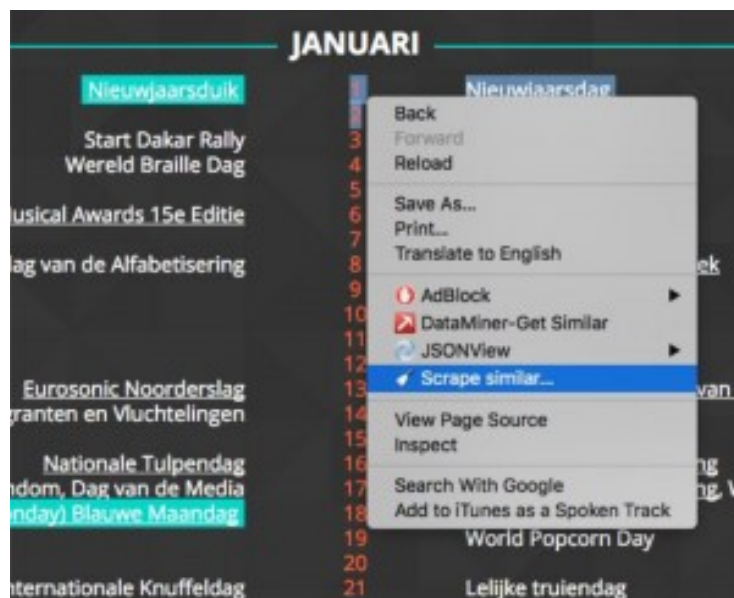
paar items gaat. Het wordt vaak wel een probleem als het om, zeg, 50, 500 of 5000 items gaat. Hiervoor zijn een aantal handige software tools beschikbaar. Eén van de vele tools voor het scrapen van data is **Scraper**.

45	28/01/18	Nationale Gedichtendag
46	28/01/18	Start Pilschowsk
47	28/01/18	Nationale Gedichtendag
48	28/01/18	Recurse Software Open Source Licentie
49	28/01/18	Dutch Open Beta
50	28/01/18	Wereld Leerdag
51	28/01/18	Dag van de Directeur
52	28/01/18	Verspreid prinses Beatrix
53	28/01/18	Cubede Herenafsluiting
54	28/01/18	

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.

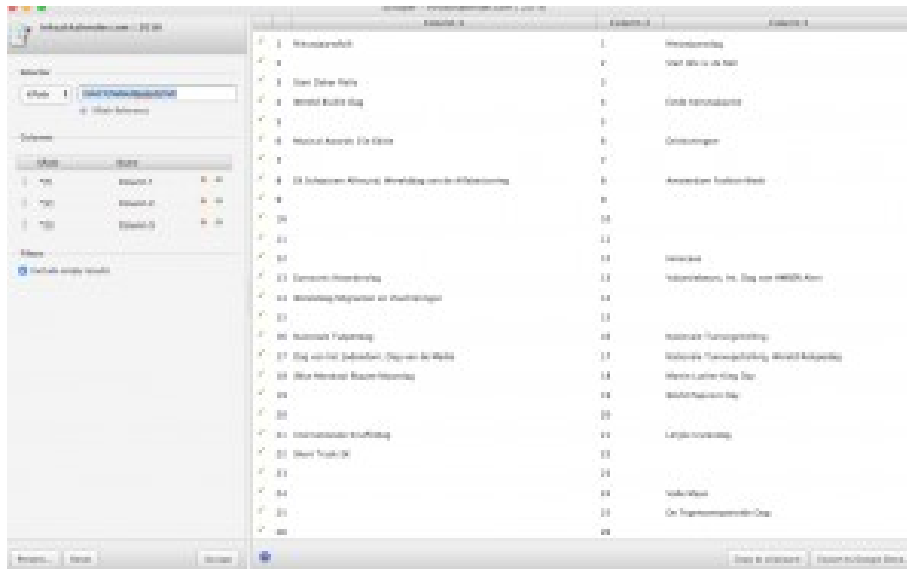
STAP 1

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	D	E	F
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 N	13	Vakantiebeurs, Int. Dag van AMBER Alert			
15	Werelddag N	14				
16		15				
17	Nationale Tu	16	Nationale Tuvogeltelling			
18	Dag van het	17	Nationale Tuvogeltelling, 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	Wereld Anti	30	Dutch Open Darts			
32	Wereld Lepri	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	Internationa Friends Day			
38		5	Warme Truendag			
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	E	F	G
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 Schaatsen		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 trulendag			
23		Short Track E		22			
24				23			
25				24 Volle Maan			
26				25 De Tegenovergestelde Dag			
27				26			

STAP 6

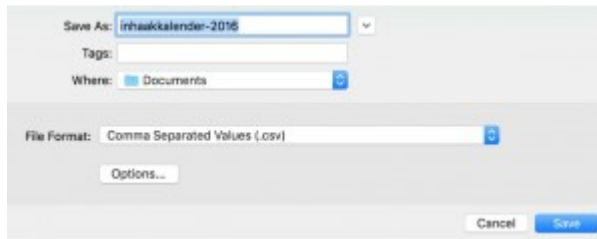
Verwijder 'Column 2' met de dagnummers (nadat je gecontroleerd hebt of de datum

	A	B	C		G
1	Datum	Column 1	Column 2	Cut	%X
2	01/01/16	Nieuwjaarsd		Copy	%C
3	02/01/16			Paste	%V
4	03/01/16	Start Dakar F		Paste Special...	^%V
5	04/01/16	Wereld Brail		Insert	
6	05/01/16			Delete	
7	06/01/16	Musical Awa		Clear Contents	
8	07/01/16			Format Cells...	%1
9	08/01/16	EK Schaatsen		Column Width...	
10	09/01/16			Hide	^0
11	10/01/16			Unhide	^0
12	11/01/16				Alert
13	12/01/16				
14	13/01/16	Eurosonic Nc			
15	14/01/16	Werelddag N			
16	15/01/16				
17	16/01/16	Nationale Tu			
18	17/01/16	Dag van het .			
19	18/01/16	(Blue Monda			
20	19/01/16				
21	20/01/16				
22	21/01/16	Internationa			
23	22/01/16	Short Track E			
24	23/01/16				
25	24/01/16				
26	25/01/16				
27	26/01/16				

kolom juist is).

STAP 7

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



STAP 8

Importeer het bestand in OpenRefine.



STAP 9

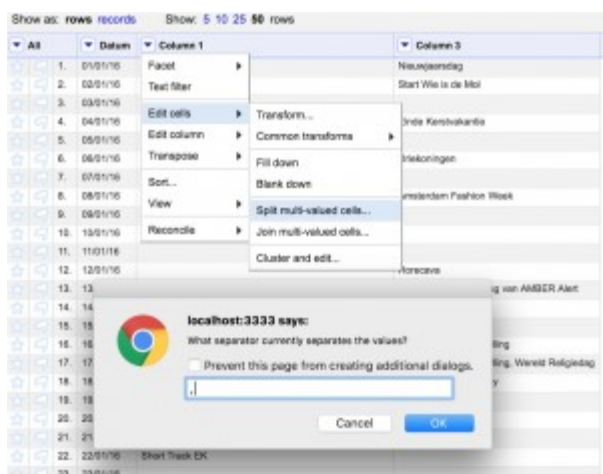
Op sommige dagen staan meerdere *events*. Deze moeten gesplitst worden.

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.		06/01/16	
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*.

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

		Datum	Column 1	Column 3
1.	01/01/16	Facet		Nieuwjaarsdag
2.	02/01/16	Text filter		Start Wie is de Mol
3.	03/01/16	Edit cells		
4.	04/01/16	Edit column		Einde Kerstvakantie
5.	05/01/16			
6.	06/01/16	Transpose		
7.	07/01/16	Sort...		
8.	08/01/16	View		
9.	09/01/16	Reconcile		
10.	10/01/16			
11.	11/01/16			
12.	12/01/16			
13.	13/01/16			Horecava
14.	13/01/16	Eurosonic Noordenlég		Vakantiebeurs
15.				Int. Dag van AMBER Alert
16.	14/01/16	Werelddag Migranten en Vluchtelingen		
17.	15/01/16			
18.	16/01/16	Nationale Tulpendag		Nationale Tuinvogelstelling
19.	17/01/16	Dag van het Jodendom		Nationale Tuinvogelstelling
20.		Dag van de Media		Wereld Religiedag
21.	18/01/16	(Blue Monday) Blauwe Maandag		Marin Luther King Day
22.	19/01/16			World Popcorn Day

STAP 12

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

Transpose Cells Across Columns into Rows

From Column	To Column	Transpose into
Column 1	Column 3 (last column)	<input type="radio"/> Two new columns Key Column: <input type="text"/> (containing original columns' names) Value Column: <input type="text"/> (containing original cells' values) <input checked="" type="radio"/> One column: <input type="text"/> Evenement <input type="checkbox"/> prepend the original column's name to each cell followed by : before the cell's value <input checked="" type="checkbox"/> Ignore blank cells <input checked="" type="checkbox"/> Fill down in other columns

Transpose Cancel

STAP 13

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

All	Datum	Evenement
1.	Facet	duik
2.	Text filter	dag
3.		de Mol
4.	Edit cells	Transform...
5.	Edit column	Common transforms
6.	Transpose	Fill down
7.	Sort...	Blank down
8.	View	Split multi-valued cells...
9.	Reconcile	Join multi-valued cells...
10.		Cluster and edit...
11.	08/01/16	EK Schaats
12.	08/01/16	Amsterdam Fashion Week
13.		Werelddag van de Alfabetisering
14.	09/01/16	
15.	10/01/16	

STAP 14

Klaar! Exporteer de data naar gewenst formaat.

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

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 *