

Inhoud

Over dit boek	9
1 Inleiding in de datajournalistiek	11
1.1 Grote hoeveelheden documenten	11
1.2 Grote hoeveelheden cijfers	15
1.3 Toegankelijk maken van data	16
1.4 Relevantie	17
1.5 Het verhaal	18
1.6 Andere definities	20
1.7 Competenties	22
1.8 Organisatievorm	22
1.9 Kansen	24
2 Vinden van data	27
2.1 Vragen naar data	27
2.2 Sociale media	44
2.3 Vragen om ongepubliceerde overheidsdata	59
2.4 Open data	67
2.5 Scrapen van data	94
2.6 Vrij zoeken	96
3 Verwerken van data	103
3.1 Voorbereidingen schoonmaak: data importeren	103
3.2 Vier populaire schoonmaakmiddelen	107
3.3 Een eerste analyse	112
3.4 De bezem erdoor met Google Refine	119
4 Presenteren van data	125
4.1 Wat is het verhaal?	125
4.2 Wie of wat vertelt het verhaal: de journalist of de data?	126

4.3	De narrativiteit van data	130
4.4	Hoe vertel ik mijn verhaal?	135
4.5	Hoe presenteer ik mijn verhaal? De kracht van datavisualisatie	139
4.6	Tabellen, grafieken en diagrammen	140
4.7	Kaarten	147
4.8	Woordenwolken	149
4.9	Tijdslijnen	151
5	Rokende pistolen	155
5.1	Data analyseren: twee strategieën	155
5.2	Case: de marathon in New York	157
5.3	Cijfers interpreteren	164
5.4	De tips van de pro's	173
6	Discussabele data	185
6.1	Onnozele verbanden	187
6.2	Uitspraken zonder waarde	190
6.3	Onzinnige conclusies	193
6.4	Foute data	195
6.5	Checklist voor betrouwbare data	198
6.6	Onzichtbare problemen	206
7	Dagelijkse data	209
7.1	Marije Rooze	209
7.2	Laura Norton Amico	210
7.3	Chrys Wu	212
7.4	Sarah Cohen en Anthony DeBarros	213
7.5	Paul Bradshaw	214
7.6	Geoff McGhee	216
7.7	Frédéric Ruys	217
7.8	Andrew Lehren	218
	Over de auteurs	221

Inleiding in de datajournalistiek

‘What I did was fairly straightforward.’

Het onderzoek van Philip Meyer naar de rellen in Detroit in 1967 vormt voor velen de bakermat van datajournalistiek of, zoals hij het zelf in zijn boeken beschrijft, *computer-assisted reporting*.

Meyer onderzocht als een van de eerste journalisten *met een computer* onderzoek naar de achterliggende oorzaak van een gebeurtenis. De wereld van de datajournalistiek is sinds 1967 flink veranderd. De rekenkracht, datasets en het aantal tools om data mee te bewerken groeiden. Maar de achterliggende gedachte? Die bleef hetzelfde:

Datajournalistiek is journalistiek waarbij grote hoeveelheden documenten of cijfers eerst toegankelijk moeten worden gemaakt voordat een relevant verhaal mogelijk is.

Onze definitie lichten we stap voor stap toe in deze inleiding in de datajournalistiek.

Hoe ga je om met grote hoeveelheden informatie? Wat is het journalistieke nut van ‘een bak met cijfers’? Welke kennis is nodig voor het toegankelijk maken van data? Waarom is datajournalistiek af en toe alleen relevant voor een klein publiek? En wat is eigenlijk een datajournalistiekproductie?

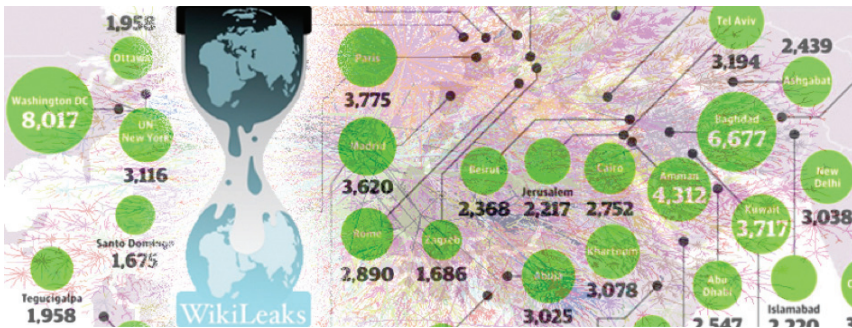
1.1 Grote hoeveelheden documenten

Sinds het bestaan van de mensheid worden data geproduceerd. De grottekening en kleitablet hebben met een boek of iPad gemeen dat ze informatie bevatten die door mensen begrepen kan worden.

Journalisten vertalen al honderden jaren informatie naar verhalen voor hun doelgroep. Waarom is datajournalistiek nodig?

Nieuw is dat de hoeveelheid data in korte tijd explosief is gestegen door de komst van computers en internet. Het is een kunst om uit de zee van informatie relevante en betrouwbare gegevens te vinden.

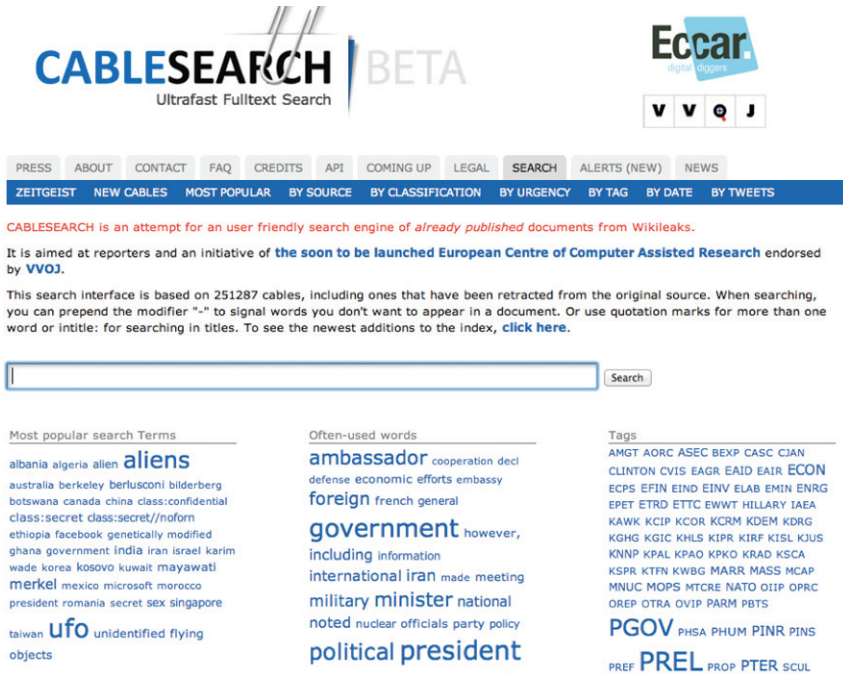
Nieuw is ook dat steeds meer data plompverloren worden gepubliceerd, zonder toelichting of context. Wikileaks is een voorbeeld van 'USB-stick-journalistiek'. Tienduizenden berichten van ambassadeurs uit heel de wereld werden zonder uitleg of context op het web gezet. Maar een handjevol journalisten kon de tijd nemen om al het materiaal te lezen en te duiden.



Figuur 1.1 *Grenzen aan begripelijkheid: wat is het verhaal achter honderdduizenden documenten?*

Bij Wikileaks en ander databases vallen journalisten steeds vaker terug op technische hulpmiddelen om de 'enorme databak' te doorgronden – een derde nieuwigheid. Zo maakte de Vereniging van Onderzoeksjournalisten een zoekmachine voor Wikileaks, bedoeld voor journalisten om gemakkelijker te kunnen zoeken.

Data worden meestal pas journalistiek relevant als het een nieuwswaarde heeft voor de ontvanger.



CABLESEARCH BETA
Ultrafast Fulltext Search

Eccar
Right Diggers

SEARCH

ALERTS (NEW) NEWS

ZEITGEIST NEW CABLES MOST POPULAR BY SOURCE BY CLASSIFICATION BY URGENCY BY TAG BY DATE BY TWEETS

CABLESEARCH is an attempt for an user friendly search engine of already published documents from Wikileaks.

It is aimed at reporters and an initiative of the soon to be launched European Centre of Computer Assisted Research endorsed by VVOJ.

This search interface is based on 251287 cables, including ones that have been retracted from the original source. When searching, you can prepend the modifier "-" to signal words you don't want to appear in a document. Or use quotation marks for more than one word or intitle: for searching in titles. To see the newest additions to the index, [click here](#).

Search

Most popular search Terms

albania algeria allen **aliens**
australia berkeley berlusconi bilderberg
botswana canada china class:confidential
class:secret class:secret//noform
ethiopia facebook genetically modified
ghana government india iran israel karim
wade korea kosovo kuwait mayawati
merkel mexico microsoft morocco
president romania secret sex singapore
taiwan **ufo** unidentified flying
objects

Often-used words

ambassador cooperation decl
defense economic efforts embassy
foreign french general
government however,
including information
international iran made meeting
military minister national
noted nuclear officials party policy
political president

Tags

AMGT AORC ASEC BEXP CASC CJAN
CLINTON CVIS EAGR EAD EAIR ECON
ECPS EFIN EIND EINV ELAB EMIN ENRG
EPET ETRD ETTT EWWT HILLARY IAEA
KAWK KCIP KCOR KCRM KDEM KDRG
KGHG KGIC KHLs KIPR KIRF KISL KJUS
KNPR KPAL KPAO KPPO KRAO KSCA
KSPR KTFN KWBG MARR MASS MCPA
MNUC MOPS MTCRE NATO OIIP OPRC
OREP OTRA OVIP PARM PBTS
PGOV PHSA PHUM PINR PINS
PREL PTER PTER SCUL

Figuur 1.2 De Wikileaks-zoekmachine van de VVOJ, <http://forreporters.com/wikileaks>

Nog een ander voorbeeld. De Volkskrant ontving duizenden pdf-bestanden met daarin belangwekkende informatie over verpleeg- en verzorgingstehuizen.



Figuur 1.3 Journalisten verdrinken in zee van informatie

Een zinnige, journalistieke conclusie kon pas worden getrokken nadat al deze gegevens werden overgezet naar een programma waarmee zoeken en sorteren mogelijk is, in dit geval Excel. Programmeurs, in samenwerking met journalisten, klaarden de klus:

	CONCERN NAAM	CONCERN PLAATS	OE PLAATS	CONCERN KV - Q1.1 SCORE	Q1.1 SCORE
2	De Borg Zorg			2079303	3.217600
3	Interkerkelijke Protestantse Stichting Woon-, Behandel- en Zorgcentra 'Sabina van Egmont'	OUDE-BEIJERLAND	GRONINGEN	24302083	3.091300
4	Zorggroep Fivelland	Appingedam	APPINGEDAM	2067888	3.211800
5	Stichting Zorggroep De Leiboorn	DEVENTER	DEVENTER	8089405	3.486500
6	Stichting Zorggroep De Leiboorn	DEVENTER	DEVENTER	8089405	3.026200
7	Stichting Zorgpartners Midden-Holland	REEUWIJK	REEUWIJK	24325621	
8	Stichting Zorgpartners Midden-Holland	REEUWIJK	REEUWIJK	24325621	
9	Stichting Viente	EDE GLD	WAGENINGEN	9148035	3.550000
10	Stichting Viente	EDE GLD	Oosterbeek	9148035	
11	Stichting Viente	EDE GLD	WAGENINGEN	9148035	
12	Stichting Verpleeghuis Het Parkhuis	DORDRECHT	DORDRECHT	41118833	
13	Stichting Christelijk Verpleeghuis Norschoten	BARNEVELD	BARNEVELD	41050547	
14	Stg. Verz. en Woonfaciliteit, Land van Horne	Weert	WEERT	41068374	
15	Stichting Zorggroep Ter Weel	GOES	GOES	41114350	
16	Stichting Zorggroep Ter Weel	GOES	KRABBENDIJK	41114350	3.417200
17	Stichting Berne Zorg	HEESWIJK DINTHER	HEESWIJK	41085597	
18	Stichting Sevagram	HEERLEN	VALKENBURG AAN DE GEUL	14082815	
19	Stichting Sevagram	HEERLEN	HEERLEN	14082815	2.582200
20	Stichting Sevagram	HEERLEN	HEERLEN	14082815	
21	Stichting Zorgcentra De Beluwe	CULEMBORG	CULEMBORG	11043745	2.845000
22	Stichting Zorgcentra De Beluwe	CULEMBORG	MAURIK	11043745	2.863100
23	Stichting Zorggroep Middenveld Drenthe	HOOGVEEN	HOOGVEEN	4079792	
24	Stichting De Vitaals Zorg Groep	EINDHOVEN	EINDHOVEN	41088225	
25	Stichting voor Regionale Zorgverlening	MIDDELBURG	MIDDELBURG	41113281	3.680300

Figuur 1.4 De vragenlijsten van verpleeghuizen nu niet meer in pdf, maar netjes in een enorme Excel-database

St Elisabeth

Adres:

Stie Brandaenstraat 4
3814 WZ Amersfoort

Beoordeling

Positie op de algemene ranglijst: 49

Positie volgens bewoners: 52

Positie volgens prestatiescore: 126

Prestatiescore

Huizen moeten ondermeer bijhouden hoe bewoners er lichamelijk aan toe zijn; hoeveel antidepressiva of sufmakende medicijnen ze gebruiken en hoe vaak ze worden gefixeerd. De resultaten worden gewogen en zijn gerangschikt op achttien thema's. Ze worden getoond als percentage van de best behaalde score in deze categorie. Hoe groener de balk, hoe beter.

Valincidenten:



Doorligwonden:



Onbedoeld afvallen:



Fouten met medicijnen:



Sufmakende medicatie:



Gebruik antidepressiva:



Vaccinatie bewoners:



Vaccinatiegraad medewerkers:



Incontinentie bewoners:



Diagnose van incontinentie:



Bewoners met katheter:



Probleemgedrag:



Fixatie van bewoners:



Depressieve bewoners:



Figuur 1.5 Het uiteindelijke resultaat is te vinden op <http://forreporters.com/volkskrant>

Dankzij de conversie door programmeurs konden journalisten opeens sorteren op verkeerd medicijngebruik, doorligwonden en andere ernstige zaken. Daardoor ontstonden allerlei nieuwe ideeën. *De Volkskrant* kon drie jaar lang belangwekkende verhalen publiceren en trok honderdduizenden bezoekers met een Verpleeghuiswijzer.

1.2 Grote hoeveelheden cijfers

De oorsprong van datajournalistiek is precisiejournalistiek: met hulp van computers en ponskaarten probeerden journalisten nieuws te ontdekken in cijfers. Wat betekent het als de sterftecijfers in een ziekenhuis gelijke tred houden met de stijging van foutief toegediende medicijnen? Is de oorzaak van de stijging van het aantal doden misschien dat verplegers pillen door elkaar halen? Datajournalistiek gebaseerd op cijferanalyse stelt journalisten in staat om met interessante hypotheses voor verhalen te komen.

NRC Handelsblad was niet tevreden met de persberichten van makelaars over de huizenbewegingen in Nederland. Die waren meestal rooskleurig opgesteld. De krant huurde een programmeur in om de voortdurend veranderende inhoud van huizenzoekmachine Funda.nl bij te houden. Het leidde tot een serie verhalen en af en toe een nieuwtje, zoals deze:

1. Makelaars manipuleren woningaanbod

ROTTERDAM, 3 JAN. NVM-makelaars en huizenverkopers manipuleren met te koop staande huizen, door ze een paar dagen van de markt te halen en daarna als nieuw aan te prijzen. De afgelopen drie maanden is dat over heel Nederland verspreid meer dan 2.700 keer gebeurd.

- Media proberen steeds vaker zelf databases te genereren. De Amerikaanse krant *Orange County Register* publiceerde een test over gezondheidsrisico's van snoep door 450 snoepsoorten chemisch te laten onderzoeken. Veel soorten hadden een te hoog loodgehalte.
- Het Britse *The Guardian* ontving ruim 450.000 bonnetjes en documenten met daarin de uitgaven van parlementariërs. Het publiek mocht aangeven welke bonnetjes het meest interessant zijn voor nader onderzoek.

Investigate your MP's expenses

Join us in digging through the documents of MPs' expenses to identify individual claims, or documents that you think merit further investigation. You can work through your own MP's expenses, or just hit the button below to start reviewing. (Update, Fri pm: we now have a virtually complete set of expenses documents so you should be able to find your MP's) Already created an account? [Log in here](#).

We have **458,832** pages of documents. **32,979** of you have reviewed **225,794** of them. Only **233,038** to go...

[Start reviewing](#)

Please read our [privacy policy](#) to find out how we use your data. You must also read our [terms of service](#). By reviewing pages, you are agreeing that you have read the terms of service, and that you agree to them.

Figuur 1.6 *Bonnetjes van Britse politici, openbaar gemaakt door journalisten van The Guardian, te vinden op <http://forreporters.com/guardian>*



Figuur 1.7 *Een Braziliaanse krant openbaarde de belastinggegevens van meer dan 6000 Braziliaanse politici*

1.3 Toegankelijk maken van data

Datajournalistiek vereist meestal hulpmiddelen om de grote hoeveelheden documenten of cijfers tot behapbare proporties terug te brengen. Voor het zoeken naar data is diepgaande kennis nodig van zoekmachines, interne procedures van dataleveranciers en enige kennis van de Wet openbaarheid van bestuur.

Voor het ontdekken van nieuws in data is het nodig om de gegevens te kunnen sorteren en te vergelijken – doorgaans via Excel. De meeste datajournalisten willen voor zichzelf de cijfers visualiseren.

NRC *Handelsblad* onderzocht presentielijsten van Europarlementariërs, gemakkelijk te downloaden via internet. Doel was te onderzoeken welke politici regelmatig afwezig waren bij stemmingen. De namen werden in Excel gezet, maar bij nadere bestudering bleek dat twee personen met op elkaar lijkende achternamen één en dezelfde persoon waren. De namen moesten handmatig worden gecontroleerd. Excel hielp bij het vinden van de verkeerde namen, niet bij het corrigeren ervan.

A	B	C	D	E	F
10036/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE JACKSON	Catherine F.
10037/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE JARZEMBOWSKI	Georg
10038/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE JEGGLE	Elisabeth
10039/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KARAS	Othmar
10040/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KAROUTCH	Roger
10041/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KAUPPI	Pia-Noora
10042/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KEPPELHOFF-WIECHERT	Hedwig
10043/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KHANBHAI	Bashir
10044/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KIRKHOPE	Timothy
10045/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KLANIT	Eva
10046/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KLASS	Christa
10047/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KNOLLE	Karsten
10048/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KOCH	Dietmar Lebecht
10049/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE KORHOLA	Eija-Riitta Avelin
10050/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LAHASSOURE	Alan
10051/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LANGEN	Vincent
10052/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LANGENHAGEN	Brigitte
10053/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LASCHET	Amin
10054/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LECHNER	Just
10055/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LEHNE	Klaus-Heiner
10056/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LISI	Giorgio
10057/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LOMBARDO	Raffaella
10058/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE LULLING	Astrid
10059/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MAAT	Albert Jan
10060/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MCCARTIN	John Joseph
10061/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MCMILLAN-SCOTT	Edward H. C.
10062/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MADELIN	Alan
10063/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MALWEGGEN	Hilja
10064/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MANN	Thomas
10065/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MANTOVANI	Mario
10066/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MARRIOS	Ismaïl
10067/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MARQUES	Sergio
10068/00195912	Rapport Valdelles de Cui A5-0001/1999	am 51,1ère partie	AM 51/1	PPE-DE MATHIAS	Georg

Figuur 1.8 Verkeerde namen

De presentielijst vormde de basis voor een interessante constatering: Nederland heeft in het Europees parlement meer macht dan op papier lijkt. Dat komt omdat de Nederlandse politici veel vaker aanwezig zijn bij stemmingen dan bijvoorbeeld Italiaanse collega's. Eigenlijk heeft Nederland op deze manier een paar zetels meer dan in werkelijkheid.

1.4 Relevantie

De uitkomsten van datajournalistiek zijn niet per definitie geschikt voor een breed publiek. Voor een fanatieke voetballiefhebber is een uitputtende analyse van spelersdata interessant, voor het grote publiek niet. Bepalend is de doelgroep. Een landelijke krant vindt verhalen interessant die de meeste mensen aanspreken (Hoe stemt Nederland?), een *special interest*-tijdschrift wil dat juist niet (Hoe stemmen autorijders?). De relevantie hangt ook af van het belang en de impact van het verhaal.

Veel datajournalisten stellen zich daarom de ‘Nou en?’-vraag: is wat ik heb ontdekt relevant voor mijn doelgroep; welk effect heeft het? Een Amerikaanse krant bestudeerde allerlei cijfers over strafbladen van chauffeurs in de transportsector, busvervoer en schoolbussen. Uiteindelijk werd voor het verhaal ‘Wie bestuurt de schoolbus?’ gekozen omdat daar de meeste chauffeurs met een overmatig alcoholverleden waren te vinden. Het Center for Public Integrity doorzocht rechterlijke uitspraken van de afgelopen dertig jaar op zoek naar namen van officieren van justitie die ooit eens in de fout zijn gegaan. Conclusie: honderden hadden een crimineel verleden.

Voorbeelden van relevante datajournalistiek staan op <http://forreporters.com/ire>.

1.5 Het verhaal

Datajournalistiek is meestal een ‘halffabriek’: de vermoedens die zijn ontstaan door data-analyse, moeten altijd worden getoetst bij deskundigen, betrokkenen, door eigen waarneming, door andere documenten en door sfeertekeningen. Datajournalisten schrijven niet altijd zelf het verhaal.

Datajournalist Christina Elmer van het Duitse weekblad *Stern*: ‘Ik moet er zelfs niet aan denken om het verhaal zelf te schrijven. Ik vind het veel te moeilijk om afstand te nemen van de cijfers.’ Andrew Lehren van *The New York Times* werkt meestal in duo’s.

Datajournalistiek heeft het gevaar in zich dat het zich te veel richt op het tonen van het probleem in plaats van het uitleggen van het probleem. Een verhalende dataproductie heeft verschillende lagen en stelt de ontvanger in staat om details te vinden die alleen voor hem belangrijk zijn, maar ook het algemene beeld geven.

In aanvulling op onze definitie is ook het effect van het werk belangrijk. Datajournalistiek stelt journalisten in staat om niet eerder vertelde verhalen te vinden, nieuwe invalshoeken te ontdekken of bestaande verhalen te complementeren met nieuwe feiten via het proces van vinden, verwerken en presenteren van een significante hoeveelheid data.

Datajournalistiek in de praktijk: Regiohack

Zet journalisten en programmeurs dertig uur bij elkaar, geef ze pizza en koffie, en er móét haast wel iets moois uitkomen.

Jerry Vermanen, werkzaam als dataredacteur bij Nu.nl, is een van de initiatiefnemers van Regiohack, een samenwerking tussen *de Stentor*, *Twentse Courant*/*Tubantia* en Hogeschool Saxion. Hij vertelt hoe het idee ontstond om dertig programmeurs en journalisten bij elkaar te zetten voor een marathonsessie van dertig uur, waarbij ze opgedeeld in groepjes allemaal een journalistiek interessante vraag zouden proberen te beantwoorden met behulp van computertechnieken. Beide beroepsgroepen werken nog te veel langs elkaar heen. 'Programmeurs vinden het fantastisch om iets nieuws te maken, maar ze missen dan het verhaal. Journalisten missen de technische kennis.'

De onderzoeksvraag van Regiohack: Wie zijn de machtigste personen in Twente? Regiohack is bezig met het opzetten van een database van belangrijke regionale mensen. De invloed wordt onder andere afgemeten aan bestuursfuncties bij bedrijven, organisaties en overheden, en aangezien er geen kant-en-klaar overzicht van die bestuurders bestaat, wordt die dataset van de grond af opgebouwd. 'Het mooie is dat je er vervolgens steeds weer nieuwe vragen aan kunt stellen en verhalen uit kunt halen. Stel dat ergens een bestuurder vertrekt, dan kun je in die database terugvinden met welke personen hij veel te maken heeft gehad en kun je die om duiding vragen', aldus Vermanen.

Wat voegen datajournalistiek en visualisaties nou toe aan de bestaande berichtgeving? Regionale kranten kunnen zich ermee onderscheiden, zegt Vermanen. 'Je haalt nieuws uit de regio dat je anders zou missen, je hebt nieuws dat anderen niet hebben. Daarmee kun je je onderscheiden en dat is, denk ik, belangrijk voor een krant, zeker als het financieel niet altijd even goed gaat.'



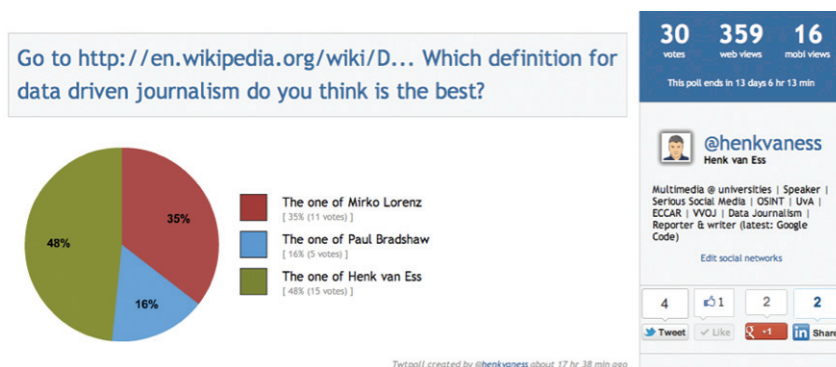
De gemeente Utrecht staat het schaamrood op de kaken. Jarenlang misbruikten enkele ambtenaren hun positie voor een winstgevende handel in onroerend goed. Waarschuwingen en meldingen werden tot op het hoogste niveau in de wind geslagen

Figuur 1.9 *Gedigitaliseerde aktes afkomstig uit het openbaar toegankelijke kadaster werden door journalisten Marcel van Sillfhout en Henk van Ess met elkaar vergeleken op de pc*

1.6 Andere definities

Volgens multimediajournalist Mirko Lorenz is door data gedreven journalistiek een *workflow* die bestaat uit de volgende elementen: *naar gegevens graven* door zelf een database samen te stellen via *scraping*, het schoonmaken van gegevens en structureren ervan, het met behulp van *filters specifieke informatie vinden*, *gegevens visualiseren en een verhaal vertellen*. Belangrijk in deze opsomming is dat de visualisatie van gegevens al plaats heeft *voordat* je als journalist publiceert. Je moet eerst zelf *zien* wat de samenhang tussen allerlei gegevens is met behulp van grafieken, draaitabellen, geplotte lijnen, kaarten, tijdslijnen en andere zaken die complexe cijfers visualiseren. De meeste datajournalisten die we spraken, doen dat altijd. Ze maken het kaartje of de grafiek eerst voor zichzelf, pas daarna voor de lezer.

Data-expert Paul Bradshaw beschrijft het proces vergelijkbaar: data moeten eerst worden gevonden. Dit kan speciale kennis vereisen van tools als MySQL of Python. Daarna moeten de data worden bevraagd, waarvoor kennis van statistiek nodig is. Tot slot moeten de data worden gevisualiseerd met behulp van open source tools.



Figuur 1.10 Aan collega-datajournalisten vroegen we welke definitie ze de beste vinden. Na 24 uur stemmen – hoe betrouwbaar zo’n poll is, lees je in hoofdstuk 6 over discutabele data – is er voor alle drie definities wat te zeggen. Ze liggen dan ook nauw in elkaars verlengde

Tim Berners-Lee, de uitvinder van het wereldwijde web, denkt dat het overleven van journalistiek afhangt van mensen die data kunnen analyseren. Uit de analyse van 86 internationale banen voor datajournalisten blijkt dat de verwachtingen over datajournalistiek bij hoofdredacties buitengewoon optimistisch zijn. Redacties zoeken nerds met vijf poten, die bovendien sociaal zijn, gevoel voor nieuws hebben, goed zijn in visualiseren en ook nog eens teamspelers zijn.

Who made the graphic? Henk Van Ess

Where can I find it? Interactive Timeline and YouTube

• Data journalism and data visualisations from the Guardian



Figuur 1.11 Voor dit boek maakten we een historische tijdlijn van alle Engelstalige boeken over datajournalistiek. Het Britse dagblad The Guardian publiceerde daar over. Het complete overzicht staat op www.forreporters.com/guardian2

1.7 Competenties

Maar wat die nerds met vijf poten dan precies moeten kunnen? De Universiteit van Tilburg, sinds 2012 actief met een master datajournalistiek, ziet zes competenties voor een datajournalist:

- data verzamelen en selecteren;
- data opschonen en analyseren;
- data doorzoeken en afwijkingen constateren;
- bekendheid met verschillende datastandaarden;
- maken van datajournalistieke producties;
- journalistieke verhalen extraheren en verspreiden.

1.8 Organisatievorm

Er zijn weinig mensen die alle disciplines beheersen. Samenwerking bij datajournalistieke projecten is daarom onontbeerlijk. Bij *The Guardian* (een van de grootste kranten in Groot-Brittannië) werken drie mensen in een team. Volgens datajournalist Simon Rogers werkt het team veel samen met andere departementen (ICT, gespecialiseerde journalisten, grafici) voor het verzamelen en analyseren van data. Datajournalistieke teams bevatten idealiter mensen met verschillende specialiteiten: online ontwikkelaars, data-analisten, grafici en statistici. Vaak is de aanvoerder van een team een journalist met data-analytische vaardigheden. De meeste mediabedrijven in Nederland hebben niet zo'n team.

Als niet alle kennis in huis is, huren mediabedrijven experts in. The Mechanical Turk en vWorker gelden als een veelgebruikte online marktplaats voor programmeurs. Hier worden opdrachten aanbesteed en kunnen programmeurs inschrijven op klussen zoals het opschonen van data of het omzetten en analyseren van grote databestanden. De naam Mechanical Turk is afkomstig van de mechanische Turk-schaakmachine. Deze werd gepresenteerd als een mechanische machine die kon schaken, terwijl er in werkelijkheid een goede schaker in de machine verborgen zat.

11/2/2007 1:42:58 PM	Twinklekumar	In Mountain View, United States (see local date/time)	9.8 (Excellent) out of 5 ratings.	Ranked #8,026 out of 194,697 (higher than 95.87% of their peers).	View 2 message(s) Reply	\$500.00
11/2/2007 1:43:46 PM	bluEyez	In Iasi, Romania (see local date/time)	0 (None) out of ratings.	Not yet ranked.	View 1 message(s) Reply	None yet
11/2/2007 1:49:01 PM	softpower	In Lahore, Pakistan (see local date/time)	7.58 (Very Good) out of 39 ratings.	Ranked #22,731 out of 194,697 (higher than 88.32% of their peers).	View 1 message(s) Reply	\$1,000.00
11/2/2007 1:49:17 PM	eshrak	In Cairo, Egypt (see local date/time)	10 (Excellent) out of 1 ratings.	Ranked #17,623 out of 194,697 (higher than 90.94% of their peers).	View 1 message(s) Reply	\$300.00

Figuur 1.12 *Programmeurs uit de hele wereld mochten meebieden om bestanden op te schonen voor een project bij een Nederlandse krant. Sommige journalisten gebruiken diensten als vWorker en Amazon Mechanical Turk om tegen weinig geld snel een klus te klaren*

Daarnaast is het bij datajournalistieke projecten heel gebruikelijk om samen te werken met vrijwilligers of universiteiten. Zo bouwden studenten informatica van de Radboud Universiteit Nijmegen in 2011 samen met onderzoeksjournalisten van het VPRO-programma *Argos* een nieuwe zoekmachine om drie grote bulken Wikileaks-documenten beter te kunnen bestuderen. Het Algemeen Nederlands Persbureau (ANP) werkt sinds 2011 nauw samen met nieuws- en opiniesite Sargasso om datajournalistieke producties te maken voor ANP-klanten. Sargasso werkt op haar beurt veel samen met opleidingen, ngo's, journalisten, bloggers en programmeurs. Hack de Overheid ontwikkelde met *NRC Handelsblad* een zoekmachine met de naam 'Toeval of niet' – een nog gebrekkige poging om dwarsverbanden te ontdekken in sociale netwerken.

Een andere vorm van samenwerking ontstaat bij het organiseren van hackatons, waarbij journalisten en programmeurs 'data martelen' en in een kort tijdsbestek data omzetten naar verhalen. Vaak worden bij deze dagen studenten en vrijwilligers uitgenodigd.

Hackatons kennen verschillende doelen. Ten eerste geven nieuwsorganisaties met een datadag een begin aan datajournalistieke projecten. Door verschillende groepen op verschillende manieren aan data te laten werken, kunnen ze verhalen verder uitdiepen. Daarnaast willen nieuwsorganisaties met dit soort

dagen nieuwe, innovatieve netwerken opbouwen, waarbij nieuwe vaardigheden in een organisatie kunnen worden getest. Tot slot zoeken nieuwsorganisaties tijdens hackdagen vaak naar nieuw talent.

1.9 Kansen

Het Finse onderzoeksbedrijf Next Media onderzocht in 2011 de kansen van datajournalistieke toepassingen. Deze kun je vinden op <http://forreporters.com/toekomst>.

De onderzoekers zien dat het aantal beschikbare databases groeit, deels als gevolg van technologische ontwikkelingen, deels door politieke besluiten, zoals de EU-richtlijn uit 2003 voor hergebruik van informatie uit de publieke sector. Ook maatschappelijke organisaties en bedrijven zijn een groeiende bron van data.

De kosten van data-analyse dalen door verbeterde analysetools, zodat datajournalistiek tot een journalistiek basisgereedschap kan uitgroeien. De journalistiek gaat samenwerken met andere data-organisaties. Samenwerkingen maken het mogelijk om innovatienetwerken aan te leggen, die binnen en buiten het bedrijf nieuwe technieken en applicaties kunnen opleveren.

Betaalde media zijn van oudsher gericht op het koppelen van publiek aan adverteerders. Ze zouden zich volgens de onderzoekers meer moeten ontwikkelen tot *datahubs*, die naast journalistieke inhoud, betaalde datadiensten leveren aan bedrijven en overheden. Belangrijkste verkoopargument daarbij is de betrouwbaarheid van de beheerde data en geleverde diensten. Onderzoeksjournalisten maken het regelmatig mee dat de overheid vraagt of zij de door media opgeschoonde gegevens mogen hebben.

Tot slot helpt datajournalistiek de basistaak van journalistiek beter te vervullen: het geeft duiding, verschaft context en levert soms ook hard nieuws op. Maar het kan nieuws ook juist persoonlijker maken. Als de NOS een lijst van ruim 200 gevaarlijke chemische bedrijven publiceert in één Excel-bestand, is dat wellicht handig, maar nog handiger is het als mensen hun eigen verhaal kunnen samenstellen: welke bedrijven in mijn buurt zijn gevaarlijk en waarom?

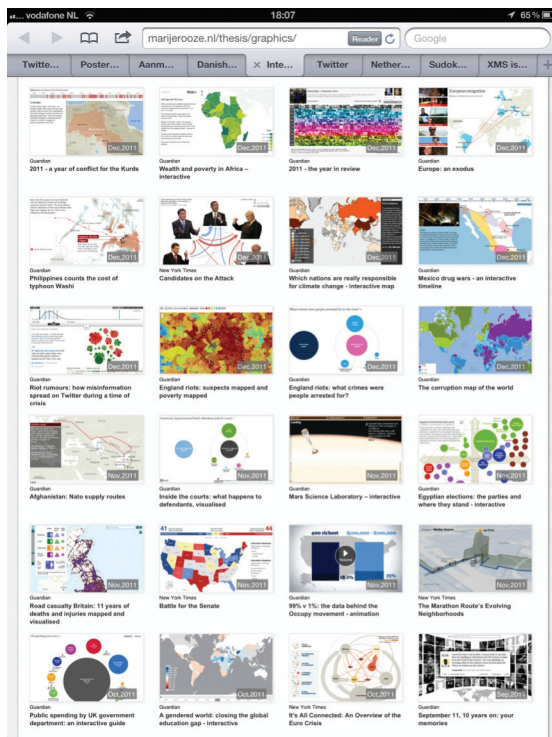
Dit boek helpt je op weg om de belangrijkste principes van datajournalistiek te doorgronden.

Oefening: Wat is goede datajournalistiek?

Ga naar <http://forreporters.com/cases>. Daar zie je meer dan 150 datajournalistische projecten.

Kies enkele projecten uit die je aanspreken en probeer te verwoorden waarom. Kijk daarbij nauwkeurig naar de navigatie, gelaagdheid, interactie en helderheid. Welke producties maken complexe problemen inzichtelijk? Welke gaan nog een stap verder en leggen uit wat de oorzaak is van de problemen?

Welke datajournalistische projecten vind je minder geslaagd? Waarom? Wat ontbreekt?



Figuur 1.13 Showcase van datajournalistische projecten op <http://forreporters.com/cases>

