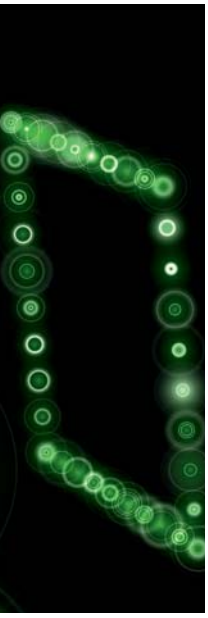


Publieke datasets worden inzichtelijk met infographics

Democratisering van data

Johan van der Kooij

Publieke datasets kunnen antwoorden geven op vragen die traditioneel voorbehouden zijn aan wetenschappers of medewerkers van overheidsinstanties. De infographic is daarbij een verhelderend middel.

Zomaar wat vragen, waarop het antwoord een interessant beeld kan werpen op ontwikkelingen in de wereld:

- Is er sprake van een dalende kindersterfte en welke landen blijven achter bij deze ontwikkeling; welke factoren zijn bepalend of afhankelijk van de kindersterfte;
- Wat is de ontwikkeling van de CO_2-emissie in de laatste 20 jaar;
- Wat is de verhouding per land tussen uitgaven aan het leger versus gezondheidszorg?

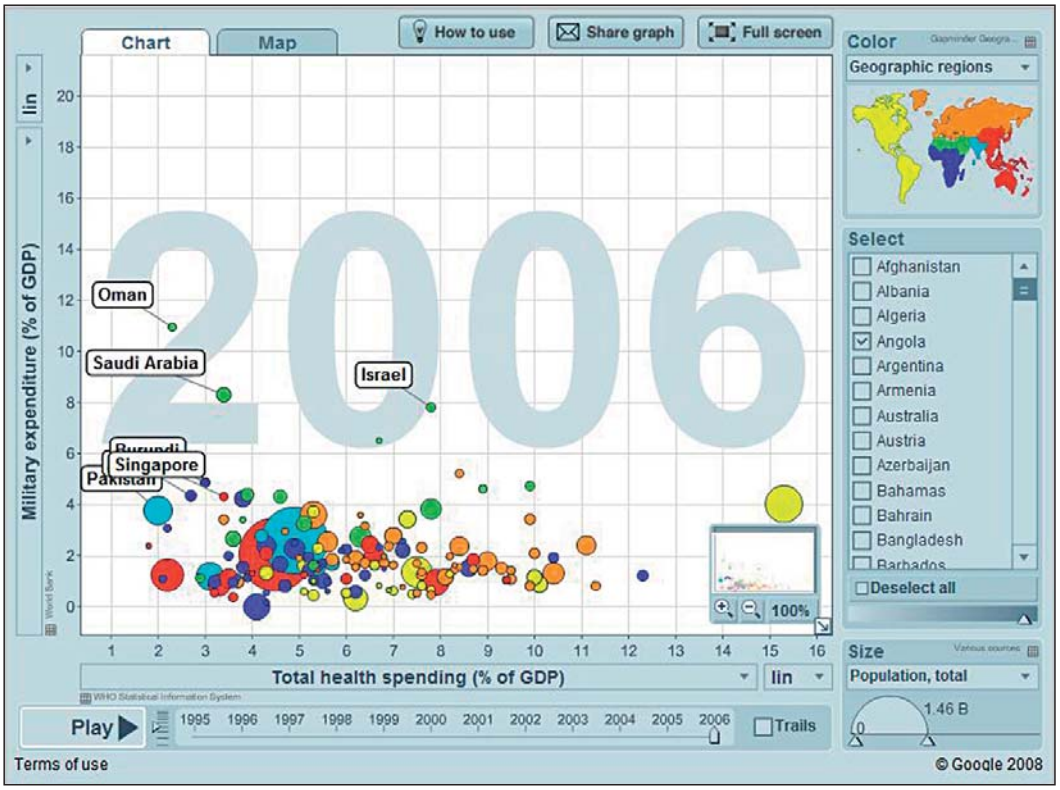
Het geven van antwoord op deze vragen zou traditioneel voorbehouden zijn aan wetenschappers of medewerkers van overheidsinstanties. Via dikke rapporten zou deze informatie beschikbaar gemaakt worden, waaraan de media – indien relevant of schokkend – in een kort nieuwsitem aandacht aan zouden besteden. In andere gevallen (uitgaven leger versus gezondheidszorg) zou de informatie waarschijnlijk nooit beschikbaar worden gesteld. De data die benodigd zijn voor het beantwoorden van deze vragen komen steeds meer via internet beschikbaar. In dit artikel wordt gekeken naar de achtergronden van deze ontwikkeling, en beschikbare databronnen en technieken om hier iets mee te doen. Ook worden enkele andere trends die hier mee te maken hebben toegelicht.

Maatschappelijke achtergronden

In 2006 werd Hans Rosling op slag een bekend persoon, door zijn presentatie op TED (Technology, Entertainment, Design). Tijdens deze presentatie gaf hij antwoord op de bovenvermelde maatschappelijk relevante vragen. Rosling, Professor of International Health, Karolinska Institutet in Stockholm, was in 2005 medeoprichter van de Gapminder stichting. Deze stichting ontwikkelt een softwareprogramma genaamd Trendalyzer. Het doel van Trendalyzer was "to unveil the beauty of statistical time series by converting boring numbers into enjoyable, animated and interactive graphics". De data die hij voor zijn presentaties gebruikt zijn via internet vrij verkrijgbaar, zie afbeelding 1.
Waar Rosling met zijn optreden forse publiciteit voor deze materie genereerde, liggen ook en vooral journalistieke en wetenschappelijke achtergronden ten grondslag aan de ontwikkeling. Enkele quotes volgen uit een artikel in 'De nieuwe reporter' uit 2007, waarin wordt beschreven hoe het tonen van onderzoeksresultaten alleen niet genoeg is, maar "de relevante achterliggende materialen en onderzoeksmethoden ook op tafel moeten. De lezer moet de conclusies kunnen controleren en reproduceren. Daarmee stelt de auteur van een wetenschappelijk artikel zich open voor kritiek, en discussie over het onderzoek. De wetenschappelijke publicatie is daarmee een tussenstop in het wetenschappelijke proces van waarheidsvinding … De wereld is opener geworden, journalistieke autoriteit is daarin niet meer vanzelfsprekend … Er wordt aan de journalistiek gesjord, gevraagd om meer openheid".
Een artikel op de website 'Open data speakers corner' refereert aan de oproep van Tim Berners-Lee: "raw data now". De stelling is dat "door middel van Open Data de wetenschappelijke innovatie versneld kan worden, door het combineren van databronnen … Bovendien kan het hogere aantal koppelingen helpen om een licht te doen schijnen op innovatieve toepassingen die mogelijk niet zichtbaar zijn als de data afzonderlijk worden onderzocht".
Naast wetenschappelijke achtergronden zijn er natuurlijk ook vele andere toepassingsgebieden denkbaar, zoals in verdere voorbeelden zal blijken.

Beschikbare datasets

Momenteel zijn datasets via vele verschillende kanalen te vinden. Data worden aangeboden door commerciële partijen en door overheden. Voorbeelden van 'datastores' zijn:



Afbeelding 1: Een voorbeeld van de interactieve 'grafiekenspeler' Trendalyzer. Hierbij de relatie tussen de uitgaven aan het leger en gezondheidszorg. De grootte van de cirkel wordt bepaald door het aantal inwoners van een land, de kleurgroepering is op wereldregio. (Bron: Google.)

- The London Datastore; "The London Datastore has been created by the Greater London Authority (GLA) as an innovation towards freeing London's data". Via de website kunnen verzoeken tot het beschikbaar stellen van data worden ingediend (bijvoorbeeld alle fietsverhuurlocaties of informatie over werkloosheid per postcode). Op de ingediende verzoeken kan gestemd worden: "406 people want this". Onder de sectie 'inspirational uses' is een voorbeeld opgenomen van real-time inzage in de fietsverhuurlocaties en aantallen beschikbare fietsen. In dit geval is een bijkomend voordeel voor Londen dat deze toepassing door een willekeurige persoon of organisatie wordt ontwikkeld, zonder verdere kosten voor de overheid, behalve dan het beschikbaar stellen van de data en een infrastructuur voor levering;

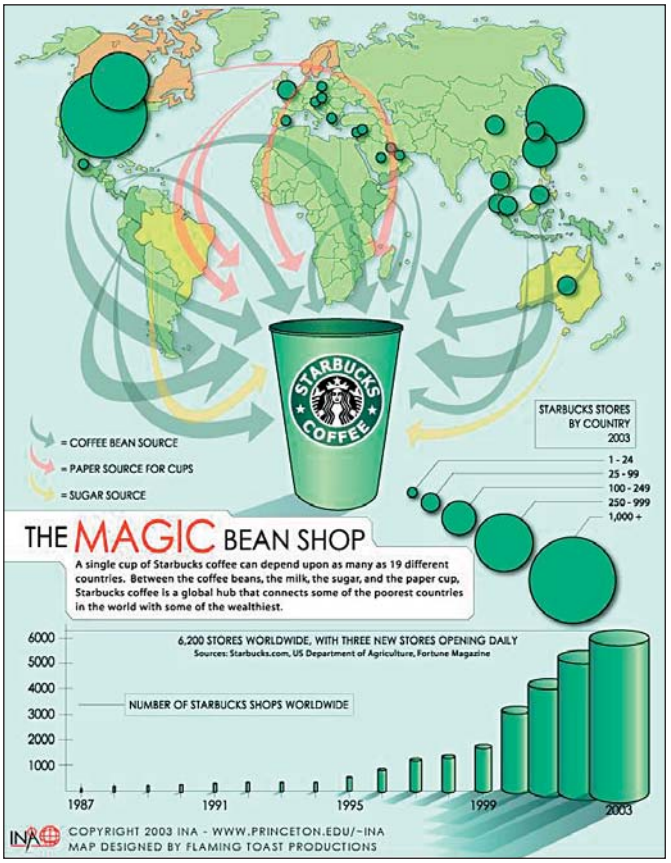
Er is een nieuwe categorie journalist ontstaan: de datajournalist

- Statline, de online database van CBS (Centraal Bureau voor Statistiek) bevat diverse datasets (ruim 2000), zoals de financiën van de waterschappen en het ziekteverzuim bij de rijksoverheid;
- The Guardian Datastore; een zeer grote verzameling van en verwijzing naar veelal overheidsdata. Voor innovatieve use cases is de 'MediaGuardian Innovation award' in het leven geroepen. Ook de WikiLeaks gegevens zijn door de Guardian Datastore door middel van vele interessante visualisaties gepresenteerd. Gebruikers van de datasets van de Guardian worden vervolgens weer verzocht om nieuwe visualisaties via een Flickr fotogroep beschikbaar te stellen;
- Datasets beschikbaar gesteld door Microsoft, Google en Amazon; Associated Press, Navteq, UNData, National Geographic, NASA, maar ook datasets met breedbandpenetratie, winkelverkopen in de USA. De rol van bijvoorbeeld Microsoft hierin is het beschikbaar stellen van de infrastructuurfaciliteiten, waarmee diverse databronnen en webservices worden gecentraliseerd, evenals een hieraan gekoppeld afrekenmechanisme. Data worden beschikbaar gesteld via het Odata protocol, een webprotocol voor het opvragen en bijwerken van data.

Mogelijkheden voor Business Intelligence

De beschikbaarheid van deze data biedt voor het BI-vakgebied diverse mogelijkheden, bijvoorbeeld op het vlak van datakwaliteit en data-analyse. Externe databronnen kunnen bijvoorbeeld geïntegreerd worden in het datawarehouse, maar ook in de rapportagetool, waarmee extra mogelijkheden worden geboden aan de eindgebruiker.

Masterdata en datakwaliteit; bij het laden van data in een datawarehouse, of zelfs het vullen van bronsystemen, kunnen publie-



Afbeelding 2: Infographic van de groei van Starbucks en de landen die profiteren van de Starbucks inkoop. (Bron: INA.)

ke datasets gebruikt worden als referentie. Zelfs het beheren van masterdata op basis van publieke datasets zou een toekomstig scenario kunnen zijn. Hierbij valt te denken aan postcodebestanden, branche-indelingen of artikelbestanden die door leveranciers of branche-organisaties beschikbaar worden gesteld via het web. Deze mogelijkheden bestaan in sommige gevallen nu ook al, echter er is daarbij veelal sprake van 'gesloten' datasets. Deze datasets kunnen mede zorgen voor een betere datakwaliteit. Bovendien kunnen resultaten door organisaties worden 'teruggegeven' aan de centrale dataset, waardoor ook de kwaliteit van de dataset weer omhoog gaat.
Benchmarkdata; voor het berekenen van marktaandeelcijfers kan de eigen omzet per regio afgezet worden tegen het totaal aantal inwoners van die regio, en kunnen ook specifieke demografische gegevens uit de dataset worden meegenomen. Hiermee kan bijvoorbeeld worden achterhaald welke producten beter verkopen in regio's met specifieke bevolkingsopbouw.

In algemene zin geldt dat de beschikbaarheid van deze data mogelijkheden biedt om de BI-analyses naar een hoger peil te trekken door de data te verrijken en de kwaliteit te verbeteren.

Tools

Voor het verwerken en presenteren van data zijn vele tools beschikbaar. Naast de traditionele BI- en ETL-tools die hiervoor ingezet kunnen worden, zijn er ook enkele nieuwe – online – tools beschikbaar gekomen. Deze tools zijn veelal in hoge mate geïntegreerd met de beschikbare datasets.Voorbeelden van deze online tools zijn:

- *Yahoo Pipes*, een online tool voor het verzamelen, bewerken en presenteren van data en webcontent. Veel databronnen zijn standaard in de repository opgenomen en daardoor eenvoudig te selecteren. Waar het bij traditionele BI-oplossingen draait om ETL (Extract, Transform, Load), voorziet Yahoo Pipes in aggregate, manipulate en mashup;
- *Google Data Explorer*; In maart 2007 nam Google de Trendalyzer technologie en medewerkers over van de Gapminder Foundation. Deze technologie is ingezet voor de Data Explorer, maar ook beschikbaar als gadget in Google Spreadsheets. Datasets kunnen eenvoudig geselecteerd worden, en met verschillende visualisaties gepresenteerd worden;
- *Many Eyes* van IBM. Deze website biedt diverse datasets inclusief visualisatie, zoals landkaarten, treemaps en bubble charts;
- Naar verwachting zal Microsoft nog dit jaar een visualisatiecomponent op het *Azure/Dallas* platform aankondigen, waarmee de beschikbare datasets online kunnen worden gepresenteerd.

Bij offline tools is het van belang dat ze voorzien in mogelijkheden om data rechtstreeks van het web (bijvoorbeeld uit HTML-tabellen), via API's of web services in te lezen. Ook van belang is dat de gebruikte tools geen beperkingen in de grootte van de in te lezen dataset kennen.

Datajournalisten

Met de beschikbaarheid van publieke datasets zijn – in navolging van Hans Rosling – diverse mensen toepassingen gaan bedenken, waarbij de data en tools op een interessante manier worden ingezet. Hierbij is een nieuwe categorie journalist ontstaan: de datajournalist. Enkele bekende zijn naast de journalisten van The Guardian ook David McCandless en Nicholas Felton. Een middel dat veel door deze journalisten wordt gebruikt is – naast de traditionele visualisaties en rapporten – de *infographic*. Een infographic is een creatieve en vrijere manier om gegevens te presenteren. Afbeelding 2 is een fraai voorbeeld. Voor het creëren van infographics wordt door de makers gebruik gemaakt van programma's zoals Photoshop en PowerPoint. Bij interactieve toepassingen kan ook Flash of Silverlight worden gebruikt. Het maken van deze infographics is veel arbeidsintensiever dan de traditionele BI-tools.

New Yorker Nicholas Felton publiceert vele infographics, deze worden door diverse media gepubliceerd. Hij geeft ook sinds 2005 jaarlijks een 'Annual Report' uit, waarin diverse infographics verzameld zijn. Deze methode van weergave onderscheidt zich van de traditionele BI-rapportages doordat het een maatwerk rapport is, dat specifiek voor de gerepresenteerde dataset van toepassing is.

Raakvlakken met andere trends

Self Service BI & Mashups – de mogelijkheden om datasets te verwerken, en dan niet door dataspecialisten maar door de 'gewone' gebruiker, wordt mogelijk gemaakt doordat desktoptechnologie beschikbaar is om grote datasets te verwerken (zoals QlikView en Microsoft Excel PowerPivot), waarbij het koppelen van tabellen tot een datamodel door de technische engine wordt verzorgd en de gebruiker zich ook niet hoeft te bekommeren om de sleutelvelden en tabelstructuren. Anderzijds komen steeds meer online en offline mashup tools beschikbaar, waarmee beschikbare datasets gepresenteerd kunnen worden op bijvoorbeeld een landkaart. Mashup tools zijn specifiek bedoeld voor eindgebruikers, om databronnen te combineren in een mooie visuele vorm.

Crowdsourcing kan vele nieuwe inzichten opleveren

Crowdsourcing – een dataset beschikbaar stellen aan de 'crowd' – de massa – kan vele nieuwe inzichten opleveren. Organisaties kunnen een dataset via bijvoorbeeld Amazon beschikbaar stellen, andere gebruikers kunnen deze verrijken met andere data, of door middel van (bijvoorbeeld statistische) analyses voorzien van feedback. Een voorbeeld van crowdsourcing is de wedstrijd van Netflix, waarbij ze 1 miljoen dollar uitloofden aan programmeurs, om op basis van de dataset van Netflix een beter aanbevelingsmechanisme te programmeren.

Conclusie

De markt voor publieke datasets zal binnen korte tijd gedomineerd worden door Google en Microsoft, gevolgd door Amazon. Niet geheel toevallig zijn dit ook de meest prominente Cloudleveranciers, waarmee deze partijen de data lijken aan te grijpen om hun Cloud-diensten meer waarde te geven. Naast deze commerciële partijen zijn er vele overheidsinstanties die data beschikbaar stellen, de rol van de Engelse krant The Guardian hierbij is het in kaart brengen van deze dataleveranciers, en vanuit de rol van datajournalist hier concrete nieuwsitems mee publiceren.

Geraadpleegde bronnen en achtergrondinformatie zijn verzameld op: www.delicious.com/johanvdk/publicdata.

Johan van der Kooij (johan.vanderkooij@vlc.nl) is managing consultant BI bij VLC.

Overheidsbestanden

Waar veel van de in dit artikel beschreven toepassingen gebaseerd zijn op eenvoudig verkrijgbare datasets, kan een andere ontwikkeling een zeer interessante toevoeging bieden: via de WOB (Wet Openbaarheid Bestuur) kunnen burgers recht op inzage krijgen in overheidsdocumenten. Een 'document' kan hierbij ruim worden geïnterpreteerd, en het kan bijvoorbeeld naast bijvoorbeeld correspondentie of vergunningen ook gaan om records uit een database. Hieraan zijn kosten verbonden, veelal worden de gegevens tegen kostprijs beschikbaar gesteld. Inmiddels zijn er (door nieuwe jurisprudentie) diverse nieuwe regels bijgekomen, maar concreet heeft dit al geleid tot diverse burgerinitiatieven.

Een voorbeeld van wat een initiatief binnen de WOB zou kunnen opleveren: een burger vindt dat de klachtafhandeling van zijn gemeente te wensen overlaat en besluit onderzoek te doen naar medewerkertevredenheid en het ziekteverzuim bij zijn gemeente. Hiertoe zal hij verzoeken moeten doen tot aanlevering van documenten (bijvoorbeeld functioneringsgesprekken) en gegevens (ziekteverzuimmeldingen) uit de omgeving van de gemeente. Bij dit verzoek zal vanzelfsprekend de privacy getoetst worden, maar aanlevering van deze gegevens zal zeer serieus beoordeeld worden. Na ontvangst kan de burger deze gegevens zelf combineren en vergelijken met andere databronnen (zoals het ziekteverzuim bij de overheid, te verkrijgen via Statline), om zodoende bijvoorbeeld met behulp van een statistisch pakket verbanden te kunnen ontdekken. Belangrijk aan de WOB-paraplu is dat in principe alle gegevens toegankelijk kunnen zijn voor geïnteresseerden.