

Eric Balster en Rob Grim

Vruchtbare ontmoeting statistiek en Linked Data

Een ontmoeting tussen enerzijds statistici en anderzijds Linked Data adepten die bij voorkeur alle mogelijke databronnen aan elkaar koppelen. Dat was het doel van de bijeenkomst op 7 en 8 juli in Tilburg van de Open Data Foundation (ODaF), een drijvende kracht achter de data-infrastructuur die door de OESO, WHO, IMF en Verenigde Naties wordt gebruikt.

De oproep van Tim Berners Lee, *'Bring your data to the web'* zal de lezer van dit blad niet ontgaan zijn. Sir Tim doelde daarbij op Linked Data, als opvolger van het Semantisch Web tijdperk waarin belofte en verwachting blijvend botsten. Het begrip Linked Data verwijst naar een verzameling van internettechnologieën waarmee op eenvoudige wijze relaties kunnen worden gelegd en beheerd tussen brokjes informatie. Linked Data technologie ligt bijvoorbeeld vaak aan de basis van *augmented reality* toepassingen. Een voorbeeld van zo'n toepassing is dat je als pizzaliefhebber in een onbekende stad op je mobieltje geattendeerd wordt op een sympathieke Italiaan in de buurt. Overtuigende toepassingen van Linked Data in de wetenschap zien we vooralsnog alleen in de bio-medische wetenschappen. Maar op de ODaF-bijeenkomst werd duidelijk dat ook toepassingen in de sociale wetenschappen voor de hand liggen.

Voordeel wordt nadeel

Ook voor Linked Data geldt de wijsheid van Johan Cruijff: 'Ieder voordeel heeft z'n nadeel'. Het voordeel van Linked Data is dat we een klein stukje informatie als een eenheid kunnen publiceren (nanopublicatie). Denk aan een eiwit met een streepjescode waarin is opgeslagen wat de bron is van het eiwit en hoe het eiwit in elkaar steekt. Dit voordeel, een microbestandje met contextinformatie, kent ook een nadeel. Het gevaar zit hem erin dat de brokjes makkelijk kunnen worden samengevoegd tot bijvoorbeeld een nieuw eiwit of bestand. Linked Data *kan* strenge regels opleggen voor de manier waarop de nanopublicaties opgaan in een groter geheel, maar is ook een vrijplaats voor het creëren van bestanden zonder regels. De computer ziet er bijvoorbeeld geen probleem in om een relatie aan te maken waarin mannen kinderen kunnen krijgen. Wat eerst een voordeel was – het behoud van context – is nu een nadeel geworden: de semantiek gaat verloren als het verband tussen de nanopublicaties onderling niet is vastgelegd. Hoe het automatisch koppelen mis kan gaan werd duidelijk tijdens de presentatie van Dan Gillman, waarvan de video-opname (net als die van de andere presentaties van de bijeenkomst) is te vinden op de website van de ODaF.

Controleerbaar en betrouwbaar

Open Data en Linked Data worden vaak in één adem genoemd. Een voorbeeld is het Open Data project dat vorige maand is ingezet door E-overheid voor Burgers, een initiatief om overheidsinformatie via webapplicaties te ontsluiten en vindbaar te maken. Toch is Open Data *niet* hetzelfde als Linked Data! Open Data omvat meer dan de geschetste technologie alleen en eist dat data ook controleerbaar en betrouwbaar zijn. In een tijd waarin databestanden op het Web

automatisch gekoppeld kunnen worden is het belangrijk dat de betrouwbaarheid van gegevens is gewaarborgd.

De ontmoeting tussen statistici en Linked Data adepten heeft de partijen ontegenzeggelijk dichterbij elkaar gebracht. De belangstelling voor Linked Data is groot, er is sterke behoefte aan kennisoverdracht en er zijn dan ook veel vervolgspraken gemaakt. Een van de afspraken is om de samenwerking tussen Nederland en het Verenigd Koninkrijk voor Linked Open Data toepassingen voor statistische gegevens te versterken. In december verschijnt het rapport waarin de case studies van de ODaF-bijeenkomst worden besproken. De bijeenkomst werd gesponsord door de Universiteit van Tilburg, DANS, SURF en de VU.

http://odaf.org/events/odaf_europe_2010.php