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 videoopname (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 dichter bij elkaar gebracht. De belangstelling voor Linked Data is groot, er is sterke behoefte aan kennisoverdracht en er zijn dan ook veel vervolgafspraken 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

Registers en sociaalwetenschappelijk onderzoek: een geslaagde combinatie?

BART F.M. BAKKER EN LÉANDER KUIJVENHOVEN

Het gebruik van registerdata in sociaalwetenschappelijk onderzoek levert veel nieuwe mogelijkheden op, maar ook nieuwe hindernissen. Het is de moeite waard om die laatste te slechten, bleek op een symposium eerder dit jaar, waarvan de bijdragen nu zijn gebundeld.

Steeds vaker worden administratieve gegevens gebruikt in sociaalwetenschappelijk onderzoek. Dat vraagt echter om nieuwe methoden en technieken en die zijn nog lang niet altijd voorhanden. Wel is er al vrij veel onderzoek gedaan dat kan helpen deze lacune te dichten. Een binnenkort te publiceren symposiumbundel* brengt daarvan een aantal voorbeelden samen. Nieuwe mogelijkheden zijn er zo veel, blijkt daaruit, dat het ontlopen van registerdata en de daarbij behorende methodologische hindernissen allang geen optie meer is.

...combineren van loopbaan- en ontslaggegevens...

Kwaliteit

Een van de factoren die voor het gebruik van registerdata pleit is dat de kwaliteit van informatie uit enquêtes is afgenomen, vooral door een steeds grotere non-response die ook steeds selectiever is geworden. Registerdata hebben daar geen last van. Een andere reden is dat enquêtes in toenemende mate duurder zijn dan het verwerven van registraties, die immers toch al om andere redenen worden bijgehouden. Ook komen er in dit e-tijdperk steeds meer registraties beschikbaar, door digitalisering of compleet nieuw ontworpen. In Nederland zorgt de steeds bredere toepassing van het BurgerService-Nummer voor meer mogelijkheden om onderling te koppelen.

Mogelijkheden

Een belangrijke winst ligt ook in de nieuwe onderzoeksmogelijkheden. Omdat registraties vaak de totale populatie beschrijven kunnen bijvoorbeeld heel kleine subpopulaties of zeldzame fenomenen worden onderzocht. Ook bieden registerdata meer mogelijkheden voor onderzoek tussen generaties en longitudinaal onderzoek. Bovendien kan de vertekening worden uitgeschakeld die enquêtes vaak opleveren door bijvoorbeeld sociaalwenselijk antwoordgedrag of geheugeneffecten – factoren die bij registraties een veel kleinere rol spelen. Natuurlijk kunnen niet altijd alle relevante vragen worden beantwoord met alleen registerdata, zodat het nodig kan zijn de ontbrekende variabelen met een enquête binnen te halen. Daarbij kunnen de registraties dan weer – heel efficiënt – als steekproefkader dienen. En ten slotte kunnen ad-hoc-vragen vaak eenvoudig worden beantwoord op basis van registraties, ook als ze gaan over variabelen die men niet snel samen in een enquête tegenkomt.

Obstakels

Er valt dus veel te winnen bij het gebruik van registerdata. Maar er is ook nog veel te *over*winnen. Beide aspecten kwamen aan de orde op het symposium waarvan de bijdragen nu zijn gebundeld. Er zijn bijdragen over de grote waarde van registergegevens voor criminologisch onderzoek, over de relatief lage kosten van longitudinaal onderzoek als registraties over loopbaan- en ontslaggegevens worden gecombineerd, en over de groei van het onderzoek over volksgezondheid door het koppelen van registraties. Het gebruik van btw-gegevens krijgt aandacht in een vergelijkende studie over Nederland en Groot-Brita n nië. Aan de obstakels wordt ook ruime aandacht besteed. Zo is er aandacht voor de validiteit van data uit enquêtes en registers (geven ze weer wat we willen meten?), voor aanpakken om die te vergelijken en voor het gebruik van registerdata als standaard om de kwaliteit van enquêtegegevens te beoordelen.

Oplossingen

Mogelijkheden om registerdata zo goed mogelijk te benutten en de beperkingen te overwinnen komen eveneens aan de orde. Manieren bijvoorbeeld om het ontbreken van variabelen in administratieve bronnen te overwinnen, uitvoeringsproblemen bij het gebruik van registerdata in de geestelijke gezondheidszorg, een raamwerk om de statistische bruikbaarheid van administratieve en andere secundaire bronnen in kaart te brengen, gebruik van schattingen om tegenstrijdigheden te overwinnen tussen registerdata en enquêtegegevens. Het CBS komt ten slotte met een onderzoekprogramma op het terrein van registers voor de toekomst . Dit programma levert niet alleen interessante onderzoekvragen op voor het CBS zelf, maar ook voor de sociale wetenschap.

Inzicht

Voor veel van de opdoemende problemen kunnen oplossingen worden gevonden, zo lijkt het. Eenvoudig zal het echter niet zijn. Er is nog veel meer inzicht nodig. Inzicht in de belangen van registratiehouders en geregistreerden, in manieren om dreigende vertekeningen en mogelijke lacunes in de gegevens te overwinnen, en om de omvang van zeer grote bestanden terug te brengen. Terwijl we de noodzaak van nieuwe methoden en technieken onderkennen, kan evenwel het huwelijk tussen registerdata en sociaalwetenschappelijk onderzoek alvast aangekondigd worden. De mogelijkheden die deze data bieden zijn te talrijk en te veelbelovend om daar geen gebruik van te maken.

* *Registers in sociaalwetenschappelijk onderzoek: Mogelijkheden en valkuilen*; informatie: www.cbs.nl/infoservice.