



RUBEN DOOD:
“BIJ ONS IS
EIGENLIJK
IEDEREEN DATA
OFFICER”

Data inzetten voor betere dienstverlening

Het Centraal Bureau voor de Statistiek werkt aan methoden om de grote hoeveelheid data nog effectiever in te zetten. Onder meer met inzet van AI en ‘synthetische data’. Bescherming van privacy blijft een kerntaak. “Hoe meer je data deelt, hoe kleiner de kans op fouten”, zegt chief data officer Ruben Dood.

“Bij ons is eigenlijk iedereen data officer.” Ruben Dood is chief data officer bij het Centraal Bureau voor de Statistiek. Een relatief nieuwe functie bij het CBS, die Dood sinds begin 2023 vervult. Hij is het contact met de buitenwereld als het over data gaat, en vooral de nieuwe ontwikkelingen op data-gebied.

Het CBS is spin in het web in het datastelsel van Nederland. Vandaar de opmerking van Dood: iedereen is er met data bezig. Talrijke overheidsinstanties verzamelen in Nederland gegevens. De basis wordt gevormd door de registers van onder meer gemeenten, Kadaster, Kamer van Koophandel, Belastingdienst en UWV. In totaal zijn er tien basisregistraties. Daarnaast zijn er nog eens tweehonderd sectorregistraties, waar het CBS uit put. De data komen samen bij het CBS, dat ook zelf data verzamelt.

“We willen in Nederland dat overheden makkelijk data kunnen delen en die vervolgens kunnen gebruiken voor beleidsdoeleinden. Wij brengen onze kennis en vaardigheden op statistisch gebied in. Dat delen is best ingewikkeld, want elke overheid heeft een grote mate van autonomie. Wij mogen gebruikmaken van de data van de Belastingdienst, kunnen ze zelfs opeisen, maar de data blijven van de Belastingdienst, niet van de overheid.”

**JE KUNT DATA ALTIJD ZO INZETTEN
DAT JE ALLEEN DE WENSELIJKE
CONCLUSIES ERUIT HAALT:
CHERRY PICKING**

Basis van de dienstverlening van het CBS is het Europese programma van verplichte statistieken. Alle EU-landen moeten die data verzamelen. Dan gaat het over zaken als werkloosheidscijfers en nationale rekeningen. In opdracht van de Nederlandse overheid verzamelt het CBS nationale gegevens, bijvoorbeeld over criminaliteit en onderwijs, maar ook gezondheidsdata. Overheidsinstanties kunnen daarnaast maatwerkopdrachten geven aan het CBS om bepaalde statistieken te maken. En dan zijn er nog een paar medewerkers die eigen onderzoek doen. “Dan gaat het doorgaans om ‘nice to know’-statistieken rond de actualiteit. Bijvoorbeeld bij de Brexit of rond 25 jaar homohuwelijk.”

De data van het CBS zijn ongelooflijk gedetailleerd. Elke Nederlander zit erin, vaak op veel verschillende manieren. Om te voorkomen dat de zeer persoonlijke gegevens herleidbaar zijn en kunnen worden

misbruikt, worden ze voor gebruik gepseudonimiseerd: BSN, naam, precieze geboortedatum en adresgegevens worden vervangen door een uniek nummer. Dat nummer is vervolgens wel hetzelfde voor dezelfde persoon in alle bestanden.

Deze zogenaamde microdata kunnen door het CBS en externe wetenschappers – onder strikte veiligheidscondities – worden gebruikt voor statistisch onderzoek. “In principe zou je elke Nederlander erin kunnen terugvinden op basis van een aantal unieke kenmerken. Ik ben zelf waarschijnlijk de enige 57-jarige afgestudeerde in geodesie die in Delft woont en in Den Haag werkt. Ik ben vermoedelijk makkelijk te vinden. Alleen is dat strikt verboden.”

Een stapje hoger zijn de geanonimiseerde statistische data. Die zijn voor iedereen toegankelijk, ook als het CBS de statistiek in opdracht maakt van derden. In deze open access data zijn alleen gegevens op groepsniveau te vinden.

Het CBS zit op een enorme hoeveelheid persoonlijke data over individuele Nederlanders. Hoe kwetsbaar is dat?

“Privacybescherming is onze kerntaak. Daarom worden individuele persoonsgegevens bij binnenkomst gelijk vervangen door een nummer. We zijn heel streng op hoe de microdata worden gebruikt. Wetenschappers mogen ze gebruiken voor onderzoek, maar op dat moment zijn ze al niet meer herleidbaar tot personen. Als wij geen goed beeld hebben van de bedoeling van het onderzoek en hoe er met de data wordt omgegaan, dan gaat het niet door. En de resultaten worden altijd door ons gecontroleerd. Is het veilige, niet-herleidbare statistiek?

Bij de statistische informatie ligt dat anders. Die mag iedereen gebruiken. De statistieken zijn betrouwbaar, maar partijen kunnen er op een oneigenlijke manier gebruik van maken. Je kunt data altijd zo inzetten dat je alleen de wenselijke conclusies eruit haalt: cherry picking. Op dat moment zullen we zeker aan de bel trekken. Bij die partij, bij de media. Maar veel meer kunnen we niet doen.”

Is het CBS met al die data geen buitengewoon interessant doelwit voor hackers?

“Daarom zijn onze data ongelooflijk goed beveiligd. Onze stelregel is dat het eenvoudiger moet zijn om bij de primaire bron de data te hacken dan bij ons. Veiligheid gaat boven alles. We werken met zwaarbeveiligde verbindingen. Maar dan nog: doordat we de data bij binnenkomst omzetten in nummers, zijn persoonlijke gegevens beschermd.”

Jullie gebruiken veel data die uit andere bestanden worden opgehaald, maar doen ook nog steeds enquêtes. Hoe betrouwbaar zijn die?

“Voor sommige statistieken zijn enquêtes onvermijdelijk. Als je wilt meten hoe tevreden mensen zijn over hun woonomgeving, of het producentenvertrouwen. Enquêtes doen we steeds minder. Omdat we steeds vaker bestaande data kunnen benutten. Dan heb je een veel preciezer en landsdekkend beeld, terwijl je bij enquêtes maar een beperkte groep bereikt.





We zijn in Nederland vrij ver in die omschakeling. Net als in de Scandinavische landen. Maar in Duitsland en België doen ze nog veel enquêtes. Daar houden ze ook nog fysieke volkstellingen waarbij enquêteurs van deur tot deur gaan. In Nederland doen we dat op basis van al geregistreerde gegevens. Een enorme kostenbesparing.”

Een groot deel van jullie dataverzameling is verplicht op basis van Europese afspraken. Toch zie je bij Europese statistieken vaak een waslijst aan voetnoten omdat cijfers niet vergelijkbaar zijn. Hoe kan dat?

“Dat wordt snel minder. Die harmonisatie gaat best snel. Europese cijfers zijn grosso modo goed vergelijkbaar. Zeker als het om de kerncijfers gaat. Werkloosheidscijfers waren lang verschillend per land, maar kennen nu dezelfde definitie. Maar er zijn nog altijd verschillen. Landen die veel data uit enquêtes halen, doen dat soms met net iets andere vragen, dus dat geeft verschillen. Veel Oost-Europese lidstaten hadden vroeger geen kadaster; vastleggen van bezit was niet nodig in communistische landen. Dat betekent wel een gat in de data.”

Vervuilde data zijn een groot probleem voor elke instantie. Hoe bewaken jullie de datakwaliteit?

“Omdat we steeds minder enquêteren, wordt de datakwaliteit hoger. De datakwaliteit wordt ook beter doordat we meer data delen. Bij de tien basisregistraties checken we bijvoorbeeld op congruentie. Als

iemand in het kadaster ergens bezit heeft, moet die persoon ook in het personenregister voorkomen. Op die manier kun je fouten eruit filteren. Die zijn beperkt. Een fractie van een promille. En dan zijn veel fouten vaak nog het gevolg van vertraging: gegevens die te laat worden doorgegeven. Ook bij de sectorregistraties die een wettelijke basis hebben, is de foutmarge erg klein.

We streven naar de overheid als ‘single point of truth’. Data die eenmaal zijn geregistreerd, hoeft je niet nogmaals op te geven. Vroeger moest je bij een verhuizing alle instanties op de hoogte brengen. Dat is nu niet meer nodig.”

Data zijn het nieuwe goud. Big Tech-ondernemingen worden er rijk van. Partijen als Google en Facebook kunnen heel veel informatie over mensen genereren. In hoeverre zijn dat concurrenten voor jullie?

“Wij zijn een gegarandeerd betrouwbare en onafhankelijke partij. Wij hebben data die mensen vaak verplicht moeten afstaan aan overheden. Data die controleerbaar zijn. Wij mogen die data ook combineren voor statistieken. Hoe we dat doen, is allemaal openbaar, dus transparant.

Dat ligt voor de Big Tech heel anders. Veel van hun data zijn niet betrouwbaar. Mensen geven op social media niet altijd een realistisch of eerlijk beeld van zichzelf. Waar dat soort partijen wel heel goed in is,

is ‘quick & dirty’-onderzoek. Als Google de vragen over verkoudheid ziet toenemen, kunnen ze een griep epidemie voorspellen. Dat doen wij niet.”

Maar zij hebben natuurlijk ook allerlei data waar jullie niet aan kunnen komen.

“Dat zijn grotendeels data die wij niet nodig hebben. Partijen als Google zie ik niet als bedreiging. Zij kunnen natuurlijk ook gebruikmaken van onze open data. Maar dat kan iedereen.”

In hoeverre gaat AI het werk van het CBS veranderen?

“Large Language Models zijn nu de rage. Die technieken kunnen wij ook inzetten, zoals we dat ook al doen met machine learning en zelflerende modellen. We zijn nu aan het testen met AI. Kunnen we AI inzetten bij het zoeken naar de goede dataset? Ik noem maar wat: je wilt een statistiek voor bosbranden. Dan kun je op bosbranden zoeken, maar dan mis je mogelijk data. Misschien moet je ook op woud zoeken, en op vuur. AI kan je helpen dat soort verbanden sneller te vinden.

Ik verwacht vooral een grotere rol van AI bij de interpretatie van data. Denk aan het uitleggen wat er in een bepaalde grafiek eigenlijk te zien is of welke conclusies je kan trekken op basis van de statistiek, bij het vertalen van data naar handelingsperspectief. Dat is iets wat wij niet doen, maar daar gebruiken overheden en bedrijven die data natuurlijk wel voor. AI kan suggesties aandragen, slimmer informatie uit datasets halen.”

IK VERWACHT VOORAL EEN GROTERE ROL VAN AI BIJ DE INTERPRETATIE VAN DATA

Overheden en bedrijven willen in toenemende mate hun beleid baseren op data. Een goede ontwikkeling?

“Ik ben buitengewoon enthousiast over databased beleid. Het is altijd goed als je je bij beslissingen op feiten baseert. Maar je kunt niet leven van feiten alleen. Je moet een visie hebben, ideeën waar je naartoe wilt. Die keuzes zijn aan de politiek.

Rond de verkiezingen ging het ineens bij alle partijen over bestaanszekerheid. Maar daar is geen duidelijke definitie van. Dus kan elke partij daar zijn eigen draai aan geven, zijn eigen cijfers gebruiken. We zijn nu bezig om tot een definitie te komen, gebaseerd op feiten. Dan weet je waarover je praat.”

De enorme hoeveelheid data die er inmiddels is over mensen, is ook beangstigend. Die data kunnen op allerlei manieren worden geïnterpreteerd en misbruikt. Hoe voorkomen jullie dat er misbruik wordt gemaakt van jullie data?

“Wij springen in Nederland en bij het CBS buitengewoon zorgvuldig om met data. Bij microdata worden alle onderzoeksvoorstellen van externe partijen voorgelegd aan een ethische commissie. Die kijkt naar de doelstellingen van het onderzoek, hoe betrouwbaar een organisatie is, welk onderzoeksplan er is en of de data daarvoor goed genoeg zijn. Jaarlijks worden er enkele verzoeken afgewezen.

Data hebben altijd een ethische component. Wij publiceren ook cijfers over migratie en criminaliteit. Daar is in toenemende mate discussie over. Ook bij ons. Je moet voorzichtig omgaan met data van kwetsbare groepen om discriminatie te voorkomen.”

Of een algoritme gaat ermee aan de haal.

“Dat kun je voorkomen door van tevoren goed te testen.”

Welke nieuwe ontwikkelingen spelen er op jullie terrein?

“De overheid is bezig met de vraag hoe je data in kunt zetten voor betere dienstverlening aan de burger. Als je ziet dat iemand zijn baan verliest, kun je hem erop wijzen dat hij recht heeft op een uitkering. Dergelijk proactief beleid kan ook op andere vlakken. Wij hebben nu data over energieverbruik op maandbasis, maar als je van energiebedrijven ook actuele data hebt, kunnen we de energietransitie

veel beter ondersteunen. En je kunt misschien voorkomen dat mensen te veel voor hun energie betalen.

Het niet betalen van de zorgverzekering is vaak een indicatie van financiële problemen. Met die betaalgegevens en data van BKR kun je wellicht voorkomen dat mensen in forse schulden terechtkomen. Statistieken over bij welke groepen mensen dit het vaakst voorkomt zijn belangrijke hulpmiddelen voor preventie. Dan moeten we wel over meer data beschikken dan we nu hebben. Ook data van bijvoorbeeld zorgverzekeraars. Bedrijven zien de voordelen wel van het delen van die data, maar ze willen nooit de enige zijn. Dus dat moet je op sectorniveau aanpakken.

Een andere ontwikkeling is het maken van synthetische data. Daarbij maak je data na op basis van bestaande data. Het zijn geen echte data, maar je kunt ze wel gebruiken om bijvoorbeeld algoritmes te testen of statistieken te maken waarvoor je te weinig echte data hebt. Het voordeel is ook dat ze veel minder privacygevoelig zijn.”

Het CBS wordt een nog grotere datafabriek.

“Dat hoeft niet. Wij zijn rupsje-nooit-genoege niet. Met de bestaande mankracht van 1800 mensen kunnen we door inzet van technologie steeds meer doen, en steeds sneller. Dat hebben we tijdens corona bewezen. We gingen van sterftecijfers op maandbasis naar sterftecijfers op weekbasis. Maar de basis blijft altijd betrouwbaarheid, onafhankelijkheid en bescherming van privacy.” ■



Ruben Dood (1967) studeerde geodesie in Delft. Daar werkte hij ook kort als onderzoeker om vervolgens over te stappen naar Rijkswaterstaat. Sinds 2008 werkt hij voor het CBS. Hij is directeur beleidsstatistiek en sinds 2023 ook Chief Data Officer.