

Datasets allemaal en voor altijd bewaren?

Expertmeeting

Bewaarbelangen Datasets

Hans Laagland ■

Moeten we datasets permanent bewaren, nu het nog ontbreekt aan richtlijnen voor het bewaren van datasets? Dat was een van de vragen op de Expertmeeting Bewaarbelangen Datasets die op 13 april plaatsvond. In een stampvolle zaal in de Rijksuniversiteit Groningen bogen drie sprekers en tegen de veertig deelnemers zich over dit onderwerp. Eén ding stond voorop na ruim drie uur spreken, luisteren en discussiëren: we zitten midden in een fase van pionieren. Lees een introductie op het onderwerp en het verslag van de Expertmeeting.

De Expertmeeting was een initiatief van gemeentearchivaris Rienk Jonker (Historisch Centrum Leeuwarden) en universitair docent geschiedenis Erwin Karel (Rijksuniversiteit Groningen). Dat het belangrijk is om bewaarbelangen van datasets te bespreken, blijkt uit de actualiteit. Dagvoorzitter Marcel Broersma (hoogleraar Journalistieke Cultuur en Media, RuG) schudde bij de opening al twee voorbeelden uit zijn mouw: Cambridge Analytica (hergebruik van datasets van Facebook voor *profiling*) en de implementatie van de Algemene verordening gegevensbescherming (AVG).

Datasets: overheidsdata en onderzoeksdata

Eerst even wat achtergrondinformatie. De Expertmeeting richtte zich op een nieuw thema, namelijk het vraagstuk rond het bewaren van datasets. En voor het goede begrip: we hebben het hier over data in digitale verschijningsvorm. Datasets worden veelal gegenereerd uit bestaande digitale systemen die worden gebruikt bij de ondersteuning van de dagelijkse werkzaamheden van organisaties. Denk aan het beschrijven van een foto in een beeldbank of het verlenen van een vergunning. De volledige beeldbank, dat wil zeggen, de beschrijvingen van alle foto's uit die beeldbank, is de dataset 'beeldbank'. Een subcategorie van deze overheidsdata betreft data die worden gecreëerd en gebruikt en het resultaat zijn van een wetenschappelijk onderzoek (onderzoeksdata).

De scope van de Expertmeeting betrof de data van overheden en de wetenschappelijke onderzoeksdata. Vanuit de provinciale en gemeentelijke overheden zien steeds meer dataportalen het levenslicht. Zo hebben in Friesland de provincie Friesland, de



Een van de conclusies van de Expertmeeting: we zitten in een fase van pionieren. Foto: Rienk Jonker.

gemeente Leeuwarden en Tresoar het initiatief genomen tot het Fries Open Data Platform (<https://friesland.dataplatform.nl>). Een dergelijk dataportaal biedt een overzicht van de beschikbare open data via datasets. Aanjagers van deze dataportalen zijn vooral afkomstig uit de opendatabeweging. Daarnaast gebruiken overheden intern in bijvoorbeeld *datawarehouses* of *datalakes* natuurlijk ook datasets voor de interne processen, voor toetsing en vorming van beleid, of om analyses te maken waaruit nieuwe producten kunnen voortkomen.

In de onderzoekswereld worden heel veel data gecreëerd. Ook bij de universiteiten groeit het bewustzijn van goed organiseren van

>>



Willem Eise Jongsma: 'Beleid moet niet te strikt worden opgezet.' Foto: Rienk Jonker.

- >> de data, van de creatie, via onderzoeksproces naar disseminatie en archivering. Dit wordt samengevat in de term *Research Data Management*. Voor onderzoek geldt een belangrijke voorwaarde, namelijk de mogelijkheid tot reproduceren van een onderzoek ten behoeve van verificatie (herhaalbaarheid). Daarbij is het van belang dat bij het opnieuw uitvoeren van een onderzoek dezelfde bronmaterialen (data inclusief algoritmes) beschikbaar zijn. Dit vraagt om een goede oplossing voor bewaren van en toegang bieden tot datasets.

Status van datasets in geval van bewaren

Zowel bij publiceren (denk aan de dataportalen) als bij intern gebruik van datasets is nog weinig oog voor het bewaren. Bewaren en langdurig toegang geven tot data verhoogt het nut van data, zeker met het oog op toekomstig historisch onderzoek en de verantwoordingsplicht van de overheid naar de burger. Een issue bij het bewaren is onder andere versiebeheer van de datasets (bij een dataset over bijvoorbeeld parkeerzones worden nieuwe gegevens over de bestaande data heen geschreven). Dit bewaren staat of valt met de selectie op datasets. Worden datasets bewaard? Wat is de bewaartermijn voor datasets? Staat de selectielijst bewaren van datasets toe? En welke datasets worden wel bewaard en welke niet?

Voor de selectie van informatie voor archieven gelden de selectielijsten. En hier zit het 'm. In de meeste selectielijsten komt het fenomeen dataset niet voor. Als begrip niet, wat geen probleem is omdat datasets als document beschouwd kunnen worden, en als soort informatie ook niet, omdat bij het opstellen van selectielijsten het bestuurlijk-juridische aspect van archieven de boventoon voert. De in 2017 vastgestelde gemeentelijke selectielijst kent wel een omschrijving waar deze sets onder kunnen vallen. Alleen is de bewaartermijn dan maximaal zeven jaar, hetgeen kort is in het licht van de behoefte om historie op te bouwen. En dat terwijl datasets van overheden wel een speciale wettelijke bescherming genieten. Ze worden in de context van de overheid als resultaat van een proces gegenereerd en vervolgens met andere datasets als bron gebruikt voor (overheids)projecten en processen. Als gevolg daarvan zijn ze te beschouwen als archiefbescheiden in de zin van de Archiefwet

1995. En deze wet bepaalt hoe overheden vanaf het moment van ontstaan tot het moment van vernietiging met hun informatie moeten omgaan.

Expertmeeting

Kortom, het was hoog tijd voor een Expertmeeting Bewaarbelangen Datasets. Doel van de bijeenkomst was het in gang zetten van een discussie vanuit de wetenschap over de bewaarbelangen en bewaartermijnen van overheidsdatasets. Sprekers en deelnemers waren afkomstig uit een breed spectrum van gemeentelijke overheden, onderzoekerswereld, archiefinstellingen, overheidsinstellingen en diverse belangengroepen. Willem Eise Jongsma (proces- en informatiearchitect, Centraal Justitieel Incassobureau) was de eerste van de serie sprekers. Hij opende de bijeenkomst als vertegenwoordiger van een uitvoeringsinstantie met zijn bijdrage 'Ontstaan van data'. Het CJIB investeert in proactief gebruik en hergebruik van data. Het gebruik en hergebruik bestaat uit het analyseren ten behoeve van beleidsmatige keuzes, het inrichten van processen en het verbeteren van dienstverlening, onder voorwaarde van privacy- en kwaliteitsaspecten. De analyse wordt een belangrijke functie. Wat betreft het bewaren zijn de eerste denklijnen om een onderscheid te maken in bewaartermijnen van de output van datagebruik, de (delen uit de) dataset en de bewerking. Een centraal register hierop is het *single point of truth* voor verantwoording. De architectuur houdt al rekening met de bewaarbelangen, ook al is er bestuurlijk nog geen besluitvorming over de afzonderlijke bewaartermijnen. Interessant hierbij is ook hoe om te gaan met bewaring van datasets, waarbij een combinatie is gemaakt van eigen data en de data van ketenpartners. Jongsma benadrukte nog eens dat het voor CJIB om nieuw terrein gaat, dat volop wordt uitgewerkt.

Bestuurlijke verantwoording

De tweede spreker, G.J. van Bussel (bedrijfskundige, bestuurlijk informatiekundige en archivaris), legde in zijn presentatie 'Waarderen, beheren en beschikbaar houden van data' nadruk op bestuurlijke verantwoording. Van Bussel maakte daarbij de parallel tussen *Research Data Management* (RDM,

onderzoeksdata) en (*Enterprise*) Information Management ('overheidsdata'). Processen uit het RDM met het oog op het realiseren van betrouwbaarheid zijn ook van toepassing op overheidsdata. Deze verantwoording zorgt voor betrouwbaarheid in het bewaren van data (onderzoeksdata en overheidsdata). Advies is om alle data te bewaren en zolang er geen bewaartermijn op datasets beschreven staat in een van toepassing zijnde selectielijst, is de enige optie blijvende bewaring. *Governance* (sturing) vindt plaats op basis van de onderdelen van de informatiewaardeketen. Bovendien moet het bestuur uitspraken doen, beleid opstellen over het samenstellen en bewaren van datasets en over de openbaarheid van data (conform de AVG), en erop toezien dat het beleid en de regels worden nageleefd, zodat uiteindelijk de goede kwaliteit van de datasets aantoonbaar is. In die zin is de AVG een zegen: zij dwingt tot het adresseren van de problematiek.

Ongekende kansen

Na het spreekwoordelijke academisch kwartiertje (dat al gauw ruim twintig minuten duurde vanwege de levendige discussies) was het woord aan Jacob Jolij (RuG, afdelingshoofd Onderzoekstechnologie en Innovatie van de faculteit Gedrags- en Maatschappijwetenschappen). Binnen de onderzoekswereld komt het besef dat toenemende beschikbaarheid van primaire onderzoeksdata, openbare bronnen, in combinatie met technische ontwikkelingen ongekende kansen biedt. Zo bieden archiefinstellingen een schat aan data die nuttig kunnen zijn voor sociaalhistorische context. Ook Jolij onderstreepte het belang van het reproduceren van onderzoek en de rol van het bewaren van datasets hierin. Daarnaast benoemde Jolij een aantal uitdagingen voor het gebruik van openbare databronnen, waaronder juridische aspecten (AVG), ethische kwesties (combinaties van datasets van openbare persoonsgegevens kunnen leiden tot *profilen*), en technische kwesties (ontsluiting in een (zo) gemeenschappelijk (mogelijk) systeem). Jolij wilde niet zo ver gaan om alle data te bewaren; verschillende onderzoekdisciplines hebben weer hun specifieke aandachtspunten als het gaat om bewaren en publiceren van datasets.

Discussie

Na deze drie presentaties was het tijd voor discussie. Dagvoorzitter Broersma trapt de discussie af met de vraag over bewaarbelangen: welk belang zou leidend moeten zijn? Al snel werd duidelijk dat hier geen eensluidend antwoord mogelijk is: context bepaalt de belangen en er is het bekende punt van openbaarheid en privacy. Zo is *profiling* vanuit de maatschappij niet wenselijk, maar kan een organisatie voor de verbetering van interne processen hier wel voor kiezen en deze keuze verantwoorden.

Moratorium: parkeren van datasets

In de discussie werd geopperd onderscheid te maken in gebruikersbelang en bewaarbelang. Onderzoekers, uitvoeringsorganisaties en archieven zijn op zoek naar criteria voor het bewaren van datasets. Vanuit de hoek van het onderzoek is er te weinig betrokkenheid in het vaststellen van de selectielijsten, waardoor er in de selectielijsten veel nadruk ligt op de beleidskant. Onderzoeksdata vereisen andere criteria dan selectiecriteria op overheidsdata. En zelfs binnen de onderzoekswereld heeft elke wetenschapsdiscipline weer haar eigen bewaarbelang.



Zo'n veertig deelnemers waren naar Groningen afgereisd om over datasets van gedachten te wisselen. Foto: Hans Laagland.

Selectielijsten hebben echter geen uitstelmogelijkheid in beslismoment van het bewaren. Vanuit de deelnemers kwam het voorstel om een moratorium in te zetten dat data voor een periode (bijvoorbeeld 25 tot 50 jaar) parkeert, zodat toekomstige onderzoekers en hergebruikers de beslissing kunnen nemen welke data voor blijvende bewaring in aanmerking kunnen komen. Een vast discussiepunt bij selectie is immers dat we nu niet kunnen bepalen welke data in de toekomst relevant worden voor onderzoek. Een uitstelmoment ondervangt dit en zo worden in deze tijd geen data vernietigd, nu er geen duidelijke richtlijnen zijn voor het bewaren van datasets. Een aspect dat overigens wel speelt bij selectie (en dus ook bij vernietiging), is dat papieren data veel eenvoudiger zijn te vernietigen dan digitale data. Systemen hebben nog geen mogelijkheid tot vernietigen; er zijn kopieën en afgeleiden in andere systemen en data bevinden zich in (meerdere) ketens. Behalve het bewaren van datasets is het creëren, standaardiseren en bewaren van metadata van die datasets cruciaal. Dit wordt opgepakt vanuit het Landelijk Coördinatiepunt Research Data Management. Belangrijk is hierbij om de onderzoekers en de ondersteuners bij elkaar te krijgen. De ondersteuners hebben een goed beeld van mogelijkheden die onderzoekers zelf niet altijd op het netvlies hebben.

In de kinderschoenen

De Expertmeeting eindigde met *take-aways* van de drie sprekers. Jolij benadrukte het gesprek tussen ondersteuners en onderzoekers: leer elkaar kennen. Volgens Jongsmas moet beleid in deze fase niet te strikt worden opgezet. Jongsmas adviseerde bovendien om in gesprek te gaan om belangen van stakeholders samen te brengen. Van Bussel drong erop aan beslissingen te durven nemen en daar niet te lang mee te wachten. Kortom: de Expertmeeting maakte duidelijk dat deze 'tak van sport' in de kinderschoenen staat en nog veel denkwerk vereist. Na afloop van de Expertmeeting werden de eerste afspraken al gemaakt voor vervolgssessies, want dat we 'iets moeten doen' met het bewaren van datasets, daarover waren alle deelnemers het meer dan eens. 📦

De auteur is dank verschuldigd aan Redmer Alma, Rienk Jonker en Erwin Karel voor hun bijdragen aan dit artikel.

Hans Laagland ■ informatiearchitect, Tresoar.

Redmer Alma ■ beleidsmedewerker informatisering, Drents Archief.

Rienk Jonker ■ gemeentearchivaris, gemeente Leeuwarden.

Erwin Karel ■ universitair docent geschiedenis, Rijksuniversiteit Groningen.