Ronald Stolk
Marco de Vos
Anwar Osseyran

Data: het nieuwe goud en de toekomst is federatief

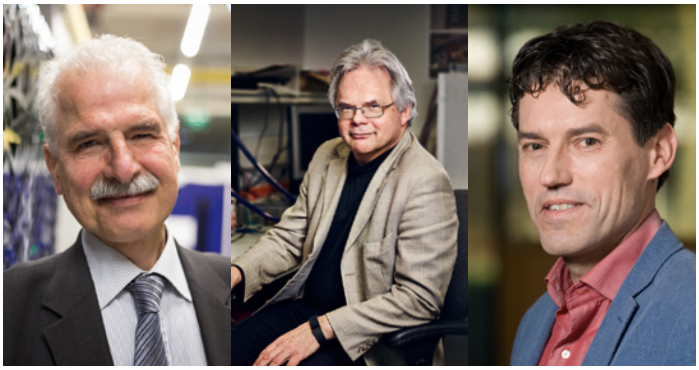
Om echt te kunnen profiteren van de enorme groei aan data in wetenschap en bedrijfsleven, moeten data gedeeld kunnen worden. De oplossing is federatief werken, waarbij kennisinstellingen hun onderzoeksdata, IT-technologie en expertise bundelen. Op die manier kunnen wij als klein land tegenwicht bieden aan de monopolisering van big data door IT-reuzen als Google.

We leven in een tijdperk waarin het volume van data exponentieel groeit en waar rekenkracht en opslagcapaciteit geen beperkingen lijken te vormen. Dit is zowel zichtbaar op data-intensieve wetenschapsterreinen zoals de astronomie, waar men werkt aan de overgang van Petabyte-systemen naar Exabytes, als in wetenschappen zoals psychologie of economie waar datasets inmiddels meerdere terabytes behelzen. De snelle ontwikkeling van dataverwerking en -analyse leidt voortdurend tot nieuwe wetenschappelijke inzichten.

Maar grote wetenschappelijke ontdekkingen en disruptieve businessinnovaties gebaseerd op grootschalige data moeten nog komen. Hiervoor zijn niet alleen veel data nodig, maar ook data gecombineerd uit verschillende bronnen, bijvoorbeeld medische gegevens, economische data en klimaatgegevens. Het combineren van gegevens uit meerdere bronnen vereist forse bewerkingen van de afzonderlijke datasets zodat deze op elkaar aansluiten.

Gestimuleerd door de Europese Commissie lopen er verschillende data science initiatieven om data veilig, betrouwbaar en toegankelijk te maken zodat delen mogelijk wordt. Het leidende principe hierbij is fair: findable, accessible, interoperable en reusable. Universiteiten en wetenschappelijke instellingen hebben het initiatief genomen om data fair te maken. Daarnaast vraagt het delen van data ook een aanpassing van de data-infrastructuur: gebruikers moeten toegang hebben tot IT-systemen van meerdere instellingen. Sommige projecten eisen zoveel capaciteit dat één centrum niet volstaat en een onderzoeker meerdere datacentra tegelijk nodig

Dit opiniestuk verscheen 5 juni jl. op de website van Computable (www.computable.nl).



V.l.n.r. Prof. dr. Anwar Osseyran, Dr. Marco de Vos en Prof.dr. Ronald Stolk

Over de auteurs:
Prof.dr. Anwar Osseyran is lid van het bestuur van SURF en algemeen directeur van SURFsara.

Prof.dr. Ronald Stolk is directeur van het Centrum voor Informatie Technologie van de Rijksuniversiteit Groningen.

Dr. Marco de Vos is managing director van Astron.

heeft. De oplossing voor beide vraagstukken is federatief werken.

Basis van kenniseconomie

De top vijf IT-bedrijven in de wereld, Apple, Google, Microsoft, Amazon en Facebook, hebben in een relatief korte periode hun machtige positie verworven door hun focus op data en disruptieve businessmodellen waarin vooral kennis en innovatie centraal staan. Hun kracht ligt in de immense schaal van data waar ze over beschikken. Hun data zijn echter niet open en dat maakt concurrentie met deze grote monopolisten steeds moeilijker. Kleinere partijen kunnen alleen over grootschalige data beschikken als ze de krachten bundelen en data met elkaar verzamelen en delen. universiteiten, umc's, hogescholen en onderzoeksinstellingen werken hiervoor samen in research data management (rdm) projecten.

Deze faciliteren het 'fair' maken van data en stimuleren hergebruik. 'Fair' vraagt namelijk een verandering in de onderzoekspraktijk. Nadat de data gebruikt zijn voor het beantwoorden van de onderzoeksvraag waarvoor ze verzameld zijn, vereist 'fair' een proces van opschonen en beschrijven van de gegevens en een procedure om ze beschikbaar te stellen aan andere onderzoekers. Rdm wordt in toenemende mate geïntegreerd in de praktijk van het wetenschappelijk onderzoek binnen onderzoeksinstellingen.

Samenwerking essentieel

Klassieke voorbeelden van bundeling van krachten in dataopslag en -verwerking zijn te vinden binnen de data-intensieve onderzoeksterreinen zoals hoge-energiefysica met het Worldwide LHC Computing Grid en astronomie met Lofar en het toekomstige SKA-project. De sterrenkunde kent een lange traditie van 'open skies' waarbij telescopentijd beschikbaar wordt gesteld op basis van de kwaliteit van voorstellen, en waarin de verzamelde gegevens na een periode van een jaar voor iedereen beschikbaar zijn.

Om deze beschikbaarheid 'fair' te maken is wel nodig dat elke dataset ook daadwerkelijk terug te vinden is en dat analysesoftware opensource beschikbaar is. Dit overstijgt het beschikbaar maken van archieven van individuele telescopen.

'Kleinere partijen kunnen alleen over grootschalige data beschikken als ze de krachten bundelen en data met elkaar verzamelen en delen.'

Bovendien willen we hergebruik van gegevens (de r in fair) ook buiten de eigen discipline bevorderen. Zo moeten Lofar-metingen gecorrigeerd worden voor ionosferische turbulentie.

Voor astronomen is dat een verstoring, of op zijn best een bijvangst. Maar voor klimaatwetenschappers is het essentiële informatie. Hetzelfde geldt voor metingen aan objecten dichterbij, zoals ruimteafval en zonnestormen. Dergelijke ontwikkelingen in data-intensieve grootschalige onderzoeksfaciliteiten geven invulling aan open science en 'fair' datagebruik. Vaak zonder dat onderzoeksgroepen zich daarvan bewust zijn overigens. Deze ontwikkelingen maken ook de noodzaak voor samenwerking en federatief werken steeds duidelijker zichtbaar.

De afgelopen jaren is er ook voor andere onderzoeksterreinen samenwerking tot stand gekomen. Deze beweging is gestimuleerd door het voorstel voor een gemeenschappelijke Europese Data Cloud (EOSC) van de Europese Commissie.

- **Samenwerken binnen één instelling:** vele universiteiten hebben een multidisciplinair Data Science Center opgericht dat als doel heeft binnen de universiteit innovatieve data science methoden te ontwikkelen en toe te passen in verschillende wetenschapsdomeinen. Voorbeelden zijn de Data Science Centers in Leiden, Eindhoven, Tilburg, Groningen en Amsterdam.
- **Samenwerken tussen instellingen:** Het Netherlands eScience Center stimuleert en ondersteunt de samenwerking tussen de data science centers in Nederland in het e-Plan initiatief. De umc's zijn gezamenlijk het Data4lifesciences-initiatief gestart, met name voor samenwerking rond privacygevoelige patiëntendata. Dit project is onderdeel van de recent gestarte nationale data-infrastructuur voor gezondheidsonderzoek Health-RI.
- **Nationaal samenwerken aan rdm:** in overleg met VSNU heeft Surf het Landelijk Coördinatiepunt Research Data Management (LCRDM) opgericht. Het LCRDM heeft tot doel het voorbereiden, faciliteren en monitoren van ontwikkeling en uitvoering van research data managementbeleid voor wetenschappelijk onderzoek in Nederland, in nauwe samenspraak met het werkveld, en het landelijk uitwisselen van kennis en ervaringen zodat Nederland tot efficiënte en effectieve ontwikkeling en uitvoering van research data management kan komen.
- **Internationale samenwerkingen:** Het eerdergenoemde EOSC wordt gesteund door verschillende Europese infrastructuren voor grootschalige opslag en verwerking van onderzoeksdata. De belangrijkste zijn EGI, Eudat Prace, Géant, Elixir, BBMRI. Nederland heeft met Lofar een concrete casus van een grootschalige onderzoeksfaciliteit die gedistribueerde en federatieve opslag- en rekencapaciteit nodig heeft. Surf Netherlands eScience Center en Astronhebben daarmee de leiding van een EOSC-pilot kunnen verwerven. Daarmee zijn we ook in een uitstekende

vervolg van pagina 15 Data: het nieuwe goud...

positie om een coördinerende rol te spelen in toekomstige faciliteiten zoals de Ska.

Datadeling met het bedrijfsleven

Vanwege de explicietere concurrentie bij het bedrijfsleven en de complexiteit van intellectuele eigendomsrechten, ligt datadeling tussen onderzoeksinstellingen en het bedrijfsleven gevoeliger dan binnen de wetenschappelijke wereld. Toch weet het bedrijfsleven dat datadeling essentieel is voor businessdoorbraken en het ontwikkelen van innovatieve toepassingen en nieuwe markten. Recentelijk is een 'smart deal' ondertekend tussen het ministerie van EZK, de provincie Noord-Brabant, TNO en Surf om gezamenlijk een Data Value Center op te richten. Het Commit2data-programma van het ministerie van Economische Zaken ondersteunt het opzetten van Big Data Value Centra. Naast Eindhoven is er een centrum opgericht in Amsterdam en zijn er concrete plannen voor Groningen.

Dergelijke kenniscentra moeten ondernemers uit het groot-, midden- en kleinbedrijf in staat stellen om met data te kunnen experimenteren en op een veilige manier data te kunnen uitwisselen. Bedrijven krijgen zo de kans van data te leren en mogelijk nieuwe businessmodellen te ontwikkelen. Daarnaast werken bedrijven en kennisinstellingen in het landelijke Big Data Alliance en het Europese Big Data Value Association samen aan kennisuitwisseling en datadeling met het doel een impuls te geven aan onderzoek, ontwikkeling en innovatie in Nederland en Europa.

In Noord-Nederland bouwt Astron met partners uit academie en industrie aan een Science Data Centre rond grote gegevensstromen. Dit centrum bouwt op drie pijlers: slimme infrastructuur, een pool van data scientists en een groep domain experts. Daarmee is er zowel een rol voor hightechindustrie als voor ICT-partners. Het is ook in een federatieve toekomst essentieel dat opslagcapaciteit en rekenkracht veel efficiënter worden in termen van energiegebruik. Bovendien groeit zowel in Smart Cities als in Smart Industry het belang van het real-time verwerken van grote gegevensstromen. Ook hiervoor is een nauwe samenwerking tussen systeemontwikkelaars, data scientists en domeinexperts essentieel. Met Lofar als concreet werkend instrument en Ska als stip op de horizon is een programma ontwikkeld waar wetenschap en bedrijfsleven beide nu al de vruchten van kunnen plukken.

'Het bedrijfsleven weet dat datadeling essentieel is voor businessdoorbraken en het ontwikkelen van innovatieve toepassingen en nieuwe markten.'

Toekomst is federatief

Uit bovenstaande voorbeelden wordt duidelijk dat samenwerking rond onderzoeksdata noodzakelijk is, zowel tussen onderzoeksinstellingen onderling als met het bedrijfsleven. De samenwerking zal uitgebreid worden richting de consument. We leven immers in een tijdperk van 'prosumers', waar de grens tussen producent en consument vervaagt en burgers, inclusief patiënten, zelf veel gegevens opslaan en gebruiken; denk aan social media, gezondheidsapps en de persoonlijke gezondheidsomgeving. Deze data zijn ook bruikbaar voor wetenschappelijk onderzoek. Door deze data te bundelen stellen we onszelf in staat om een tegenwicht te bieden aan de monopolisering van 'big data' door de grote IT-reuzen en om lokale kennisontwikkeling te behouden en uit te breiden.

Binnen de Nederlandse onderzoeksinstellingen wordt, onder regie van SURF steeds concreter gewerkt aan federatief samenwerken. Niet alleen data 'fair' maken en beschikbaar stellen voor wetenschappers van elders, maar ook toegang tot elkaars IT-systemen. Hiervoor zijn IT-aanpassingen en afspraken nodig, zoals over federatieve inlogprocedures en standaardisatie van werkwijzen. Dit maakt het bijvoorbeeld mogelijk dat een economie-onderzoeker gebruik kan maken van de 'overcapaciteit' in een systeem dat voor de astronomie is ingericht. Federatief werken is dus niet alleen het delen van data maar ook het onderling aanbieden en delen van infrastructuur, zowel op het gebied van rekenkracht, dataopslag, als lokale expertise. De enige manier voor een klein land als Nederland om de concurrentie met de grote landen en bedrijven aan te kunnen, is nationaal en internationaal samenwerken. De toekomst is federatief.

- Worldwide LHC Computing: **http://wlcg.web.cern.ch/**
- Lofar: **www.lofar.org**
- SKA-project: **www.skatelescope.org**
- Netherlands eScience Center e-Plan initiatief: **https://escience-platform.nl**
- The Netherlands Federation of UMCs Data4lifesciences: **http://data4lifesciences.nl**
- Nationale data-infrastructuur voor gezondheidsonderzoek: Health-RI: **https://www.health-ri.nl**
- Landelijk Coördinatiepunt Research Data Management (LCRDM): **www.lcrdm.nl**
- Commit2data-programma van het ministerie van Economische Zaken: **www.dutchdigitaldelta.nl/big-data/over-commit2data**
- Big Data Alliance: **www.bigdata-alliance.org**
- Het Europese Big Data Value Association: **www.bdva.eu**