

De wetenschap wordt anders

Het world wide web fungeert als wetenschapsversneller, zegt de Amsterdamse hoogleraar Kennisrepresentatie en Redeneren Frank van Harmelen. Daarmee verandert er veel. Maar dat is nog niets vergeleken bij wat ons te wachten staat.

‘Natuurlijk heeft het Web ons wetenschappelijke leven veranderd. Als wetenschappers bloggen en hyperlinken we dat het een lieve lust is,’ aldus de informaticus die, heel toepasselijk, op 20 oktober de Diërsede van de Vrije Universiteit mocht houden in aanwezigheid van eredoctor en web-uitvinder Tim Berners-Lee. En hij vervolgde: ‘Maar dat is allemaal oude koek. Ik beweer dat het Web in de nabije toekomst een veel diepgaander invloed op de wetenschap zal hebben dan we op het moment om ons heen zien.’

Twee gevolgen van de beschikbaarheid van het web zullen daarvoor zorgen, licht Van Harmelen bij navraag toe: ten eerste de manier van publiceren en de mogelijkheid om data uit te wisselen, en ten tweede de mogelijkheid om het web zelf als bron van data te gebruiken. ‘Tot nu toe wordt er gepubliceerd in wetenschappelijke artikelen, die welbeschouwd dienen als een soort ‘staatsbegrafenis voor wetenschappelijke resultaten’. Je maakt hypothesen, tabellen en plots, en vervolgens kan niemand er meer iets mee doen. Je kan de data niet checken want je kan er niet bij. Maar je kan ze ook niet gebruiken voor andere, nieuwe doeleinden. Dat wordt allemaal compleet anders. We kunnen nu de data publiceren zodat iedereen erover kan beschikken. En niet een paar jaar nadat de onderzoeker zijn of haar werk voorlopig heeft afgesloten, maar meteen.’

Als een golf

Daardoor ontstaan oneindig veel nieuwe mogelijkheden, betoogt Van Harmelen, en niet alleen om voort te borduren op het werk van een ander maar vooral ook om data te combineren en daardoor volledig nieuwe onderzoeksvragen te beantwoorden. In sommige exacte disciplines zoals de sterrenkunde, de deeltjesfysica of de levenswetenschappen, is deze manier van publiceren al heel gewoon. In andere, zoals de scheikunde en de sociale wetenschappen, nog niet. ‘Maar deze beweging trekt als een golf over de wetenschap’. Wel is er meer nodig dan alleen de mogelijkheid om uit te wisselen, benadrukt de informaticus. ‘Je moet het op zo’n manier doen dat de gegevens voor anderen ook bruikbaar zijn. Het Semantisch Web heeft daarvoor een mooie standaard opgeleverd: het Resource Description Framework of RDF. Met dat RDF kan je van elke dataset de objecten, de variabelen en de relaties daartussen beschrijven.

Dat maakt het makkelijker om ze te hergebruiken en vooral om dat door computers te laten doen.’

Het web als observatorium

In de sociale wetenschappen zijn al voorbeelden van projecten die met deze aanpak veelbelovende resultaten bereiken, aldus Van Harmelen. Communicatiewetenschappers laten bijvoorbeeld inhoudsanalyses van berichten in de media uitvoeren door hun computer – in plaats van een legertje studenten of huisvrouwen – en slaan de data op in RDF zodat ze ze gemakkelijk kunnen verbinden met data uit vele andere bronnen. Ander voorbeeld: het aan de VU verbonden Netwerk Instituut bestudeert het wetenschapsproces en heeft daarvoor nu niet alleen citatieanalyses beschikbaar – ‘de dataset daarvan is niet zo groot en loopt ook nog eens vijf jaar achter,’ – maar het hele internet. ‘Wetenschappers doen nog veel meer dan publiceren en citeren. Ze zitten op conferenties, ze bloggen, en dat zie je allemaal op internet dus je kan het ook meten.’ Een soortgelijke aanpak van het web als ‘observatorium’ toont de studie van organisatiewetenschappers aan de VU naar kennisnetwerken tussen bedrijven. ‘Vroeger stuurde je een vragenlijst op en dan mocht je blij zijn als je 30% terugkreeg. Die netwerken zijn tegenwoordig steeds meer op het net te achterhalen. Kijk maar naar LinkedIn.’

Het gaat hard en de perspectie-

MARTIJN DE GROOT



Frank van Harmelen

ven zijn stralend, vindt Van Harmelen. Maar niet alleen stralend. Het kan ook te hard gaan, citeert hij een bekende uitdrukking: ‘What to do when succes is becoming a problem?’

Lezing: Het Semantic Web als wetenschapsversneller

Digitaal onderzoek kan nog veel beter

Wetenschappers, journalisten en onderzoekers maken weliswaar volop gebruik van digitale bronnen, maar ze laten zich daarbij onnodig remmen door de relatief primitieve zoekmogelijkheden op internet. Zonder veel moeite kun je veel geavanceerder en systematischer zoeken, waardoor de kwaliteit van je onderzoek enorm kan toenemen. Dat zegt Ewoud Sanders volgende week in zijn bijdrage aan het symposium ‘Door Data Gedreven.’

Aansluitend bij de Bert van Selm-lezing, die hij eerder dit jaar hield,

Sociaal Statistisch Bestand: een voorbeeld uit de praktijk

Het Sociaal Statistisch Bestand (SSB) van het Centraal Bureau voor de Statistiek (CBS) is ontwikkeld om meer themaoverstijgende, longitudinale en gedetailleerde informatie samen te kunnen stellen, aldus projectleider Johan van Rooijen van het SSB.

Inmiddels omvat het meer dan veertig registers met informatie over uiteenlopende terreinen zoals banen, uitkeringen, zelfstandigen, processen-verbaal, opleidingen, woningen en demografie. Gegevens uit verschillende bronnen worden geïntegreerd om ze consistent te krijgen, en daarna op een gestructureerde manier beheerd en ontsloten. Standaardisatie en documentatie zijn daarbij belangrijk. De gegevens uit het SSB worden door veel gebruikers binnen en buiten het CBS gebruikt. Van Rooijen zal die toepassingen illustreren aan

de hand van een onderzoek naar de gevolgen van bedrijfseconomisch ontslag. Voor dat onderzoek werden werknemers die in 2003 op die manier hun baan verloren twee jaar gevolgd. Zo kon worden vastgesteld in welke mate zij weer toetraden tot de arbeidsmarkt. De meerderheid blijkt daar toe in staat maar een niet te verwaarlozen deel ondervindt langdurige negatieve gevolgen van het ontslag.

Lezing: Het Sociaal Statistisch Bestand: een veelzijdige bron voor onderzoek

Sprekers op het symposium ‘Door Data Gedreven’

Deze en de volgende pagina’s van e-data&research zijn gewijd aan het symposium ‘Door Data Gedreven’, volgende week in Den Haag. Een terugblik op het eerste data-archief van ons land treft u op pagina 7. Net als het symposium gaat deze bijlage verder vooral over de toekomst van het alfa- en gammaonderzoek. Welke ongedachte mogelijkheden dienen zich aan door het creatief gebruik van reeds verzamelde data? Welke kansen doen zich voor? Lees hier de vooruitblik.

Peter Doorn: Openingswoord (Interview pagina 9)

Ewoud Sanders: Zoek de vergeten dichter (pagina 7)

Willem Bouten: e-Ecologie, combinatie van natuur en techniek (pagina 8)

Henk den Heijer: Schepen, mensen en goederen, 1600- 1800 (zie pagina 8)

Esther Jansma: Duizend jaren houtgebruik (pagina 6)

Johan van Rooijen: Het Sociaal Statistisch Bestand (pagina 5)

Maarten Marx: Tekstanalyse voor de sociale wetenschappen (pagina 6)

Frank van Harmelen: Het Semantic Web als wetenschapsversneller (pagina 5)

Plaats: De Glazen Zaal, Prinsessegracht 26, Den Haag

Tijd: 2 december 2009, 12.00 – 18.00 uur

Begin november was het symposium zo goed als volgeboekt. Nog beschikbare of vrijkomende plaatsen worden uitgegeven via www.dans.knaw.nl/nl/dans_symposia/, waar u ook het volledige programma treft.



Ewoud Sanders

wil Sanders laten zien hoe je gebruikmakend van openbare bronnen in relatief korte tijd op je eigen pc of laptop een digitale materiaalverzameling kunt aanleggen van honderden miljoenen woorden. Een flexibele verzameling die voor ieder onderzoek aan te passen is en die zeer geavanceerd kan worden doorzocht. Zo kunnen de resultaten chronologisch worden geordend, omgekeerd chronologisch, op relevantie en thematisch. Ook kan, gebruikmakend van indexeringssoftware, met één zoekopdracht razendsnel worden gezocht op allerlei spellingvarianten van een woord of naam, en op talloze woordcombinaties.

Sanders laat een en ander zien aan de hand van een voorbeeld: leven en werk van een jong gestorven dichter die actief was tussen 1880 en 1925. Hij liet enkele dichtbundels, een stapel brieven en een dagboek na. Kunnen wij in grote lijnen de leefwereld van de dichter digitaal reconstrueren – de boeken, tijdschriften en kranten die hij las? En kan ons dit helpen om zijn brieven te annoteren en moeilijk leesbare passages in zijn dagboek te ontcijferen? Ja dat kan. Doordat wij op onze eigen pc zoveel bronnen uit de tijd van de jong gestorven dichter met jokertekens letter voor letter kunnen doorzoeken, blijkt het zelfs mogelijk om woorden in zijn brieven die door inkt- of wijnvlekken grotendeels onleesbaar zijn geworden, te ontcijferen.

Sanders bouwde de afgelopen drie jaar een digitale taalbibliotheek van ruim 4,5 miljoen pagina’s

en ruim twee miljard woorden, volgens hem momenteel de grootste digitale bibliotheek voor het Nederlandse taalgebied. ‘Geen mens kan zoveel lezen als een computer,’ zegt hij. ‘Mijn computer leest, gebruikmakend van optische tekenherkenning (OCR) ruim vierhonderd boeken per maand. En het geheugen van mijn pc is veel betrouwbaarder dan het mijne: iedere naam, ieder woord, ja zelfs ieder woorddeel is supersnel terug te vinden, tikfouten uitgezonderd. Je eigen digitale materiaalverzameling aanleggen, je ontworsten aan de huidige beperkingen van internet, heeft mij enorm veel gebracht. Ik kan hierdoor gerichter en beter gestructureerd zoeken dan ooit tevoren – wat vrijwel wekelijks ontdekkingen oplevert die hiervoor onmogelijk waren geweest. En als je een materiaalverzameling aanlegt op basis van openbare bronnen zijn de kosten zo gering dat dit binnen ieders handbereik ligt.’

Lezing: In het hoofd van een vergeten dichter. Een digitale reconstructie.