38

E-Science

De wetenschap in de 21ste eeuw

awt | *Adviesraad voor het* Wetenschaps- en Technologiebeleid

Colofon

Tekst:	Pieter Adriaans, Paul Diederen, Gaston Heimeriks
Illustratie:	Sylvia Weve
Vormgeving omslag:	Junior beeldvorming – Zoetermeer
Druk:	Quantes, Rijswijk

Januari 2011

ISBN 978-90-77005-53-8

Inhoudsopgave

Samenvatting

"No other country has put together such an extraordinary ecosystem of optical infrastructure and leading edge research and industrial technology transfer applications as the Netherlands. This is a true 21st century infrastructure that is, in a way, equivalent to the remarkable 17th century infrastructure of windmills and canals." (Cook Report on Internet Protocol, 2010, p. 25.)

Wie het veld van wetenschapsbeoefening in het eerste decennium van de 21ste eeuw overziet, moet concluderen dat de uitdagingen waar we voor staan niet kleiner zijn geworden. Op veel gebieden zijn wetenschappers aangelopen tegen problemen van een tot dusverre ongekende omvang en complexiteit. Voorbeelden zijn: de studie van het klimaat, de menselijke cognitie, de cel, de elementaire structuur van de materie, de taal, sociale netwerken. Bij het bestuderen van dergelijke complexe fenomenen lopen traditionele empirische methoden spaak. Het inzetten van computer- en databasetechnologie voor het bewerken en interpreteren van de data is onontbeerlijk. Het ligt daarom voor de hand dat er binnen de informatica een discipline ontstaat die zich speciaal richt op de ondersteuning van wetenschappelijk onderzoek: e-science.

Het toegenomen gebruik van computers en het internet heeft in de wetenschap veel veranderd. Computers zijn het verlengstuk van meetapparatuur geworden, maar bieden ook mogelijkheden tot het ontwikkelen van nieuwe modellen. Kennis kan beter worden opgeslagen en hergebruikt. Dankzij meer dataopslagcapaciteit kunnen wetenschappers gegevensbestanden aanleggen die, in combinatie met efficiënte zoekalgoritmen, het vinden en ordenen van informatie aanzienlijk versnellen. Supercomputers rekenen enorme problemen door. Visualisatie is standaard geworden. Via het internet zijn nieuwe kennisuitwisselingstructuren en nieuwe samenwerkingsvormen voor kennisproductie ontstaan.

In het algemeen wordt e-science gekarakteriseerd door de combinatie van drie verschillende ontwikkelingen: toegenomen rekencapaciteit, gedistribueerde toegang tot enorme hoeveelheden gegevens, en gebruik van digitale platforms voor samenwerking en communicatie. Dit heeft zijn weerslag op het gebied van de productie, de uitwisseling en het gebruik van kennis. De belangrijkste ICT-gerelateerde veranderingen betreffen de organisatie van wetenschap, vormen van samenwerking in het onderzoek, de toegenomen beschikbaarheid van gecodificeerde gegevens in databases, en het gebruik van computermodellen, simulaties en visualisaties. Dit heeft geleid tot het ontstaan van een nieuwe kijk op de aard van de werkelijkheid en van nieuwe gebieden van wetenschappelijk onderzoek. Deze ontwikkelingen roepen vragen op naar de gevolgen voor de organisatie en het functioneren van het

wetenschappelijk onderzoek in de wereld, en in het licht hiervan over de wenselijke organisatie van het onderzoek in Nederland en naar de rol van de overheid hierbij.

1 Introductie

Wetenschap en complexiteit

De wetenschap ziet zich in de 21ste eeuw geplaatst voor uitdagingen van grote omvang en complexiteit. Voorbeelden van actuele probleemgebieden zijn de studie van het klimaat, de menselijke cognitie, de cel, de elementaire structuur van de materie, van taal, van sociale netwerken. Deze domeinen onderscheiden zich alle min of meer door de volgende kenmerken:

1. *Zij zijn fundamenteel complex*. Het menselijk genoom bevat zo'n 10^{10} bits aan informatie. De opslagcapaciteit van het menselijk brein wordt geschat op 10^{14} bits. Bij het modelleren van dergelijke domeinen kunnen we niet terugvallen op de simpele, elegante, formele modellen die in het verleden kenmerkend zijn geweest voor de empirische wetenschappen (in de traditie van Galileo en Newton). Zelfs eenvoudige bewerkingen op systemen van een dergelijke complexiteit gaan de capaciteit van onze computers algauw te boven.

2. *De beschikbare data zijn altijd 'sparse'*. Deze systemen kunnen in praktische zin niet uitputtend bestudeerd worden. We kunnen waarschijnlijk in veel gevallen niet voldoende observaties verzamelen om uit alle mogelijke theorieën de juiste te selecteren. Van alle theoretisch mogelijke levensvormen die in ons universum ooit gerealiseerd kunnen worden, komt er bijvoorbeeld in de werkelijkheid maar een uiterst miniem deel voor. De rest past er gewoon niet in. Onze klimaatmodellen zijn onvolledig omdat we nog niet lang en exact genoeg hebben kunnen waarnemen. Maar als we de benodigde waarnemingen wel hadden, dan zouden onze databases te groot worden voor berekeningen van enige importantie. Er zijn in theorie zoveel mogelijkheden tot disfunctioneren van het menselijk lichaam, dat er op aarde nooit voldoende patiënten voorhanden zullen zijn om die uitputtend te bestuderen. Als de ziektes al geobserveerd worden, dan zijn de patiëntenpopulaties te klein voor het vormen en testen van een goede theorie. Maar van die kleine populaties zijn er wel weer heel veel verschillende. Alleen al in Amerika zijn er 25 miljoen patiënten met een zeldzame ziekte waarvoor het economisch nauwelijks lonend is een behandeling te ontwikkelen. Dit aspect van sparse data en slechte convergentie van theorieën is iets dat binnen de politiek voor spanningen zorgt, getuige de recente discussies rond klimaatmodellen en de kosten van de gezondheidszorg.

3. *Verdere kennisontwikkeling gaat gepaard met de opbouw van extreem grote databestanden*. Deze bestanden zijn te groot om door een enkel individu te

worden overzien. De hoeveelheid informatie groeit sneller dan het aantal experts dat haar kan interpreteren. Vaak is de groei exponentieel. Er verschijnen bijvoorbeeld in de biologie meer dan 400.000 papers per jaar. Zonder geautomatiseerde data mining en information retrieval technieken voor het ontsluiten van deze data gaat het zicht op verbanden verloren en dreigen dit soort vakgebieden te verdrinken in een data flood.

4. *Kennisontwikkeling op deze domeinen is interdisciplinair.* De domeinen zijn zo complex dat ze alleen door grote teams van internationaal samenwerkende experts adequaat bestudeerd kunnen worden.

Zonder uitgebreide inzet van computertechnologie is het nagenoeg onmogelijk vooruitgang te boeken in dergelijke domeinen.[1] Een paar ontwikkelingen zijn daarbij van centraal belang:

- In verschillende wetenschapsgebieden zijn de afgelopen jaren grote databases opgebouwd die de beschikbare kennis in hun domein samenvatten en die als basis kunnen dienen voor digitaal onderzoek (denk bijvoorbeeld aan wordnet, het menselijk genoom, medline).
- Daarnaast is door de toepassing op glasvezeltechnologie de *network latency* nagenoeg nihil geworden.[2] Het is voor een computer in Nederland makkelijker de inhoud van het intern geheugen van een andere computer in de Verenigde Staten op te vragen dan om dezelfde informatie van zijn eigen harddisk te halen. Daardoor kunnen grote hoeveelheden computers nu dynamisch in netwerken bij elkaar gebracht worden ten behoeve van complexe rekentaken.
- De ontwikkeling van *machine learning* en *data mining software* is nu zover gevorderd dat het analyseren van grote wetenschappelijke databestanden voor wetenschappers zinvol wordt. Een run van een data mining algoritme kan in principe een paper in *Nature* opleveren.

1 In bijlage 1 staat een meer technische toelichting van de grenzen die gesteld zijn aan het computationeel modelleren van complexe systemen.

2 De *network latency* is de tijd die een datapakket nodig heeft om van zender naar ontvanger te gaan.

Complexiteit: enige ruwe cijfers

10^{10} bits:	De hoeveelheid informatie in menselijke genetische code
10^{14} bits:	De opslagcapaciteit van het menselijk brein
10^{17} bits:	De opslagcapaciteit van een zoutkristal op quantumniveau
10^{92} bits:	De theoretische opslagcapaciteit op quantumniveau van het universum
10^{123} bits:	De theoretische schatting van het totaal aantal rekenstappen van het universum beschouwd als quantumcomputer sinds de Big Bang

Deze cijfers, uit Lloyd (2000), die slechts zeer ruwe (en soms ook omstreden) schattingen zijn, worden hier gegeven om een idee te geven wat de studie van complexe systemen inhoudt. Zelfs elementaire berekeningen van polynomiale complexiteit (bijvoorbeeld een legpuzzel oplossen) leggen, als de databestanden maar groot genoeg zijn, een enorm beslag op rekencapaciteit en energieconsumptie.

E-science

De toepassing van informatie- en communicatietechnologie (ICT) in de wetenschap wordt vaak met de term e-science aangeduid. Bovenstaande ontwikkelingen geven aan dat e-science meer is dan een modewoord in de wetenschap en het wetenschapsbeleid.[3] In het algemeen wordt e-science gedefinieerd als de combinatie van drie verschillende ontwikkelingen: i) de toename van de rekencapaciteit en het gebruik hiervan, ii) de gedistribueerde toegang tot enorme hoeveelheden gegevens, en iii) het gebruik van digitale platforms voor samenwerking en communicatie. Het centrale idee achter e-science is dat de productie van kennis enorm zal verbeteren door de combinatie van gebundelde menselijke expertise, van gegevens en bronnen en van digitale instrumenten voor dataverwerking en visualisatie. E-science staat in nauwe relatie met een andere IT-discipline, computational science (CS), die zich richt op specifieke simulaties van complexe systemen, waarbij men kan stellen dat e-science de infrastructuur verzorgt waarin computational science haar onderzoek kan verrichten.[4]

De term e-science staat in beleidskringen vooral in de belangstelling in verband met de financiering van grootschalige infrastructurele faciliteiten, met name in die onderzoeksgebieden waarin het onderzoek afhankelijk is van de inzet van enorme hightech voorzieningen (denk bijvoorbeeld aan CERN). Dit sluit aan op programma's van nationale overheden ter bevordering van e-science (in het Verenigd Koninkrijk) en voor het creëren van een 'cyberinfrastructuur' (in de Verenigde Staten) en een 'e-infrastructuur' (in de Europese Unie). E-science is evenwel breder. Het gaat bij

3 Zie Wouters (2006).

4 Over de Nederlands opvattingen rond computational science, zie de brief van 10 september 2009 (TWIN/SH/4681) van de KNAW aan NWO over dit onderwerp: http://www.knaw.nl/publicaties/pdf/20101025.pdf. Zie ook: Sloot et al. (2007).

e-science niet alleen om de toepassing van digitale technologie, maar ook om een transformatie van de sociale processen in de ontwikkeling en verspreiding van kennis.[5]

E-science as defined by the OECD[6]

Originally the term e-science (also called e-research or cyberinfrastructure) encompassed mainly large science projects, using grid computing in order to provide a powerful technology platform for distributed science. However, with the rise of the participative web, and more widely-available computing and more participants in science networks, including students and science libraries, the scope of discussion is expanding to cover many network-enabled science activities on the Internet. Part of the promise of e-science is the idea of a unified science workflow system, that can connect instruments with computation, data with visualisation, and allow powerful analysis. One can imagine the capabilities of such an e-science infrastructure extending beyond just researchers to decision makers and citizens as well.

ICT speelt een steeds belangrijker rol in de wetenschapsbeoefening, ook al is die rol niet voor alle wetenschappen dezelfde. Sommige disciplines danken hun bestaan aan ICT (bijvoorbeeld de bio-informatica), andere zijn door ICT nieuwe richtingen opgegaan (zoals de astronomie en de hoge-energiefysica), en weer andere hebben met ICT een krachtig hulpmiddel verworven, maar zijn intrinsiek hetzelfde gebleven (bijvoorbeeld de geschiedenis). Hoe dit ook zij, duidelijk is dat het gebruik van ICT een grote impact heeft op de inhoudelijke ontwikkeling van de wetenschap en op organisatie en functioneren van het wetenschappelijk bedrijf.
Net als andere landen en de EU zet ook Nederland in op e-science. Echter, de e-scienceontwikkelingen in Nederland zijn tot op heden meer voortgekomen uit het ondernemerschap van een aantal wetenschappelijke en dienstverlenende instituten dan uit een landelijk ontwikkelde beleidsvisie en beleidsprogramma.[7] Het is de vraag welke verantwoordelijk de overheid in dezen draagt. De overheid is immers verantwoordelijk voor de kwaliteit, de omvang, het vernieuwend vermogen en de macrodoelmatigheid van het onderzoeksbestel in Nederland. Al deze aspecten worden direct of indirect beïnvloed door ontwikkelingen in ICT. Het doel van deze achtergrondstudie is informatie te verschaffen die helpt bij het nadenken over deze vraag.

---

5 INWO stelt dat het bij e-science gaat om een nieuwe fase in de wetenschap. "E-science staat voor de koppeling tussen computertechnologie, moderne communicatiemiddelen en wetenschap. Dat gaat van het gebruik van een eenvoudige wetenschappelijke online-bibliotheek tot en met 'big science' zoals bij tienduizend wetenschappers die samenwerken aan de Europese deeltjesversneller CERN. Ook is 'in silico' onderzoek op veel gebieden een enorme stimulans voor vooruitgang." http://www.nwo.nl/nwohome.nsf/pages/NWOP_5V7LH6 (oktober 2010).

6 webnet.oecd.org/CommServerPers/blogs/participativeweb/archive/2007/09/30/backgrounder-e-science.aspx

7 Dit werd reeds geconstateerd door SURF in haar 'E-science beleidsverkenning' uit 2004.

Leeswijzer

Hoofdstuk 2 behandelt het huidige investeringsbeleid op het gebied van e-science. Veel beleid is gericht op de fysieke component van kennisproductie. Het Nederlandse investeringsbeleid in de afgelopen decennia heeft Nederland in het algemeen aan goede ICT-voorzieningen geholpen en heeft ons land op onderdelen een positie in de wereldtop verschaft.
Hoofdstuk 3 geeft een beeld van de invloed van ICT op wetenschappelijk onderzoek en op de positie van de wetenschap in de samenleving. ICT heeft het karakter van het onderzoeksproces veranderd. De belangrijkste ontwikkelingen in de kennisproductie zijn procesmatig van aard (het gebruik van communicatietechnologie voor andere vormen van organisatie en samenwerking, en de inzet van nieuwe digitale instrumenten – databases, simulatieprogramma's, visualisatie-instrumenten – voor het onderzoek zelf). Daarnaast zijn ze inhoudelijk van aard: nieuwe onderzoeksrichtingen, een nieuwe kijk op objecten van onderzoek, een nieuwe kijk op het onderzoek zelf. ICT heeft ook het karakter van de verhouding van wetenschap en samenleving veranderd. Wetenschap en samenleving zijn veel sterker op elkaar betrokken geraakt.
Hoofdstuk 4 sluit deze studie af met het identificeren en agenderen van een drietal vragen. De gepleegde investeringen in ogenschouw nemend en de fundamentele veranderingen in de wetenschap ten gevolge van de toepassing van ICT overziend, eindigen we deze achtergrondstudie – voor het moment – met vragen naar de organisatorische en institutionele gevolgtrekkingen die we aan de geschetste ontwikkelingen zouden moeten verbinden.

2 Investeringen in e-science

"Engineering breakthroughs alone will not be enough to achieve the outcomes envisaged for these undertakings. Success in realizing the potential of e-science – and other global collaborative activities supported by the 'cyberinfrastructure' – if it is to be achieved, will more likely be the resultant of a nexus of interrelated social, legal and technical transformations." (Paul A. David, Stanford University en University of Oxford)

Inleiding

E-sciencebeleid is in veel landen gericht op de infrastructuur voor kennisproductie. In de Verenigde Staten ligt het zwaartepunt bijvoorbeeld bij specifieke toepassingen die extreem veel rekenkracht of dataverwerkingscapaciteit vergen, veelal in de natuur- en levenswetenschappen. Alleen in Engeland is er een overheidsinitiatief waarmee e-science over de volle breedte van alle wetenschappelijke disciplines ondersteund wordt.[8]
Nederland behoort qua investeringen in e-science met landen als Duitsland, de Scandinavische landen en Canada tot de groep direct achter de koplopers. De overheid ontvangt permanent signalen uit het veld over noodzakelijke ingrepen en investeringen rond ICT. Het is moeilijk om deze signalen te interpreteren omdat zij elkaar vaak tegenspreken en altijd prospectief van aard zijn, dus moeilijk empirisch te staven. In dit hoofdstuk geven we een overzicht van het huidige investeringsbeleid rond e-science in de landen om ons heen en in eigen land.

E-science in onze omgeving[9]

In de Verenigde Staten is het Atkinsrapport uit 2003 bepalend geweest voor de beleidsvorming rond e-science.[10] Dit rapport schets een visie op de zogenaamde revolutie in wetenschap en techniek. Het beveelt aan om op grote schaal te investeren, één miljard dollar per jaar, in een breed programma dat bestaat uit een aantal gekoppelde grootschalige faciliteiten, nationale *repositories* en bibliotheken. Een aantal vervolgrapporten zijn sindsdien verschenen met meer toegespitste aanbevelingen voor verschillende wetenschapsgebieden.[11] De roep om meer investeringen in cyberinfrastructure werd vervolgens ook gehoord in de sociale wetenschappen,

8 Het programma is in 2009 positief geëvalueerd op initiatief van een van de UK research councils (http://www.epsrc.ac.uk/research/intrevs/escience/Pages/default.aspx) . Op het moment van schrijven is er nog geen definitieve duidelijkheid over de continuering van de projecten.
9 Deze paragraaf is gebaseerd op Wouters (2006) en Schroeder and Fry (2007).
10 Atkins, D.E. (2003).
11 Bijvoorbeeld de publicaties van de National Science Foundation over Cyberinfrastructure, zie www.nsf.gov/publications/index.jsp?org=OCI.

waar men de mogelijkheden van nieuwe ICT-hulpmiddelen bij het analyseren van sociaal gedrag begon te ontdekken.[12] Maar vooralsnog ligt de nadruk in de Verenigde Staten op informatica, natuurwetenschappen en techniek, en op business-toepassingen. Er wordt niet zozeer gesproken van e-science maar over *Grid computing*-toepassingen.

In het Verenigd Koninkrijk kwam de belangrijkste impuls voor e-science in 2000 met de oprichting van een programma voor de financiering en van het Nationaal e-Science centrum van de *Engineering and Physical Science Research Council* (EPSRC) en de *Office of Science and Technology* (OST). Sindsdien zijn er projecten gekomen in verscheidene wetenschappelijke en technologische gebieden. Eveneens is er een initiatief van latere datum voor de sociale wetenschappen van de Economische en Sociale Onderzoeksraad (ESRC) dat ook nu een nationaal centrum coördineert. In de praktijk blijkt de nadruk te liggen op de technologie. In het Verenigd Koninkrijk ligt net als in de Verenigde Staten veel nadruk op de vraag hoe verschillende instrumenten kunnen worden ingezet om een meer wijdverbreid gebruik van e-science te bewerkstelligen. Het beleid richt zich steeds meer op de integratie van e-science in andere beleidsinitiatieven.

E-science heeft met zijn nadruk op samenwerking en communicatie per definitie een wereldwijde dimensie, maar de verschillende nationale en supranationale (EU) e-science-initiatieven weerspiegelen de structuur van de bestaande nationale systemen van wetenschap en innovatie. Het Amerikaanse cyberinfrastructure-initiatief is bijvoorbeeld gericht op de versterking van de nationale kennisinfrastructuur in de traditie van 'big science', missiegedreven nationale laboratoria en de grootschalige aanpak van specifieke uitdagingen (ruimtevaart, de oorlog tegen kanker, het menselijk genoom, etc).[13] Het Verenigd Koninkrijk kent een vergelijkbare traditie, maar heeft ook een serie van onderzoeksprogramma's die gericht zijn op gebieden als generieke grid-middleware en interdisciplinaire onderzoekssamenwerking. Tegelijkertijd is er in de EU een opeenvolging van kaderprogramma's waarin het belangrijkste doel is om een geïntegreerde infrastructuur te creëren voor onderzoeksinstellingen uit de deelnemende landen.

Een recent e-scienceproject van de Europese Commissie stoelt expliciet op de Amerikaanse en Britse ervaringen.[14] De e-science programma's in het Verenigd Koninkrijk en de Verenigde Staten hebben geleid tot initiatieven in Europa (op EU-niveau via de e-Infrastructures Reflection Group)[15], in Azië en Australië, en elders.

---

12 Zie Berman en Brady (2005); zie ook het rapport uit 2005 over 'Cyberinfrastructure voor de geesteswetenschappen en de sociale wetenschappen' van de American Council of Learned Societies (ACLS): www.acls.org/cyberinfrastructure/cyber_report.htm.

13 Zie Galison en Hevly (1992) en Westwick (2003).

14 Het draagt de titel: A study on requirements and options for accelerating the transition from traditional research to virtual research organisations through e-infrastructures. Zie Barjak (2006).

15 www.e-irg.org.

De verwachting is dat nationale onderzoeksinitiatieven in de toekomst voornamelijk bestaan uit onderzoeksprogramma's die gericht zijn op bepaalde vakgebieden en met beperkte termijnen.

Hoewel er nationale accentverschillen zijn, zijn de overeenkomsten groot. In veel landen is er een centrale organisatie ter bevordering en coördinatie van e-science initiatieven opgericht, als onderdeel van de nationale (of, in geval van de EU, supranationale) onderzoeksinfrastructuur.[16] Ook schuilt achter vrijwel alle initiatieven ter wereld het idee van een *technology push*: de beschikbaarheid van de nieuwe instrumenten versnelt het onderzoeksproces. Dit brengt grote voordelen voor onderzoekers met zich mee en stimuleert daarmee de vernieuwing in het onderzoek.

E-science in ESFRI

Europa heeft een lange traditie in het gezamenlijk investeren in grootschalige onderzoeksinfrastructuur. Een recent vehikel hiervoor is het *European Strategy Forum on Research Infrastructures* (ESFRI), een overlegorgaan dat een coherente aanpak van de beleidsvorming inzake onderzoeksinfrastructuren in Europa ondersteunt en optreedt als een incubator voor de internationale onderhandelingen over concrete initiatieven. ESFRI werd gelanceerd in april 2002 en bestaat uit vertegenwoordigers van de EU-lidstaten en de Europese Commissie. Het houdt zich voornamelijk bezig met de voorbereiding van een Europese routekaart voor nieuwe onderzoeksinfrastructuren van pan-Europees belang. Deze routekaart identificeert nieuwe onderzoeksinfrastructuren met een verschillende mate van rijpheid, die in de komende 10 tot 20 jaar gerealiseerd moeten kunnen worden. De ESFRI-routekaart wordt periodiek geactualiseerd. Veel van de ESFRI-initiatieven hebben een sterke e-sciencecomponent.

In 2007 stelde de minister van OCW de Nederlandse *Commissie Nationale Roadmap Grootschalige Onderzoeksfaciliteiten* in (naar de voorzitter ook wel de *Commissie Van Velzen* genoemd).[17] De commissie adviseert de minister over de belangrijkste investeringen in grootschalige onderzoeksfaciliteiten in de komende vijf tot tien jaar en zoekt daarin aansluiting bij de Europese routekaart van ESFRI. Het gaat om grootschalige faciliteiten die grensverleggend onderzoek mogelijk maken en door hun schaalgrootte niet door een individuele universiteit of instelling gefinancierd kunnen worden, en waarvoor dus samenwerking binnen Nederland of Europa noodzakelijk is. Hiermee zijn bedragen van ten minste 40 miljoen euro per faciliteit over een periode van tien jaar gemoeid.

16 Zie Drori, Meyer, Ramirez en Schofer (2003).
17 De commissie is mede ingesteld naar aanleiding van de adviezen in het rapport Nijkamp (2005).

De roadmapcommissie heeft het eerste deel van haar advies in december 2007 aan de minister van OCW aangeboden. Hierin gaat het om een aantal faciliteiten die op de ESFRI-lijst staan. In juli 2008 heeft SURF een ICT-roadmap gepubliceerd met een beschrijving van een nationale ICT-onderzoeksinfrastructuur voor Nederland. Het advies werd opgesteld in samenwerking met stichting NCF-NWO, BigGrid, VL-e, SARA en NBIC (Netherlands Bioinformatics Centre). In 2009 is Surfnet met GigaPort 3 van start gegaan.

Nationale Routekaart

Binnen het domein van de Geestes- en Maatschappijwetenschappen
- CLARIN (Common Language Resources and Technology Infrastructure)
- ESS (European Social Survey)

Binnen het domein van de Natuur- en Technische wetenschappen
- European XFEL (X-ray Free Electron Laser)
- KM3NET (Cubic Kilometre Neutrino Telescope)
- ELT (European Extremely Large Telescope)

Binnen het domein van de Milieuwetenschappen
- ICOS (Integrated Carbon Observation System)
- LIFE WATCH (Research Infrastructures Network for Research in Biodiversity)

Binnen het domein van de Levens- en Medische wetenschappen
- European Biobanking and Biomolecular resources

E-science in Nederland

Nederland heeft met een aantal grote ICT-infrastructuren inmiddels concurrerende voorzieningen gecreëerd voor wetenschappelijk onderzoek.[18] Het meest intensieve en geavanceerde gebruik treft men aan onder fysici, die zich in het hart van de Nederlandse gridgemeenschap bevinden. Ook zien we een concentratie binnen de *life sciences*, met een aantal links naar gedragswetenschappen.[19]

Bij de ontwikkeling van de infrastructuur voor e-science in Nederland is een reeks van (koepel)organisaties en regisseurs betrokken. De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) stimuleert bijvoorbeeld via het NCF-fonds voor gridinfrastructuur de ontwikkeling van het Nederlandse *science grid*. De Stichting Universitaire Research Faciliteiten (SURF) ontwikkelt en onderhoudt via SURFnet en programma's als SURFworks de universitaire netwerkfaciliteiten. De Koninklijke Nederlandse Academie van Wetenschappen (KNAW) is samen met NWO betrokken bij het initiatief *Data Archiving and Networked Research* (DANS). Bij ICT-Regie, het

18 Op sommige punten neemt een aantal andere landen Nederland als voorbeeld, bijvoorbeeld Frankrijk met zijn Grid5000 project.

19 Zie Kircz (2004) en KNAW (2004) over initiatieven van stichting SURF en KNAW ter bevordering van e-geesteswetenschappen.

nationaal regieorgaan voor ICT-onderzoek en -innovatie was tot medio 2010 'het mobiliseren en stimuleren van creativiteit op het gebied van ICT' belegd.[20] Inmiddels heeft de minister na een externe evaluatie het mandaat van ICT-regie met ingang van 2011 opgeheven. In de recente brief waarin dit besluit kenbaar gemaakt is, wordt de volgende schets van de huidige situatie gegeven:[21]

"Het ICT-onderzoek achten wij van onverminderd belang voor de Nederlandse economie en wetenschap. In 2009 is SURFnet van start gegaan met het GigaPort3-project als opvolger van het GigaPort Next Generation Netwerk-project, en heeft als doel om de SURFnet netwerkinfrastructuur naar een hoger plan te brengen. SURF heeft voor dit project € 32 mln uit het FES gekregen. Op dit moment werkt NWO samen met SURF aan de oprichting van een e-Science Research Center en neemt Nederland vooralsnog deel aan PRACE als Principal Partner. Voor het creëren van innovatiekansen speelt het op peil houden van een ICT-kennisbasis een grote rol. Stimulering hiervan, via NWO maar ook, zoals vorig jaar, via toegang tot de WBSO en recenter de programma's Service Innovation and ICT en COMMIT, blijft nodig. Voor het COMMIT-voorstel heeft het kabinet op 28 mei een reservering gemaakt van € 50 mln, waarbij een onafhankelijke internationale adviesraad per COMMIT-deelproject een bindend advies zal geven of het betreffende deelproject in aanmerking komt".[22]

Over de de definitieve toekenning van deze projecten is op dit moment geen duidelijkheid. De opheffing van ICT-regie is een indicatie van het feit dat coördinatie van e-science-initiatieven maar beperkt van de grond is gekomen. Van een heldere onderzoeks- en innovatiestrategie voor de langere termijn is vooralsnog geen sprake. Kennisinstellingen volgen de ontwikkelingen op afstand. Wel heeft de Universiteit van Amsterdam (UvA) e-science tot speerpunt gekozen en een e-sciencecoördinator aangesteld.
Nederland investeert jaarlijks enige tientallen miljoenen in supercomputers, netwerken, dataopslag en integratie. De middelen daarvoor komen voor een beperkt deel uit structurele bronnen en voor een groot deel uit incidentele fondsen. Vaak zijn deze wisselende regelingen en tijdelijke subsidiebronnen niet optimaal op elkaar afgestemd. Aan de structurele kant heeft de Stichting Nationale Computerfaciliteiten (NCF), onderdeel van NWO, een jaarlijks budget van 6 miljoen euro voor met name supercomputers. SURFnet heeft jaarlijks 32 miljoen euro te besteden, waarvan 13 miljoen euro subsidie van de overheid voor innovatie van het netwerk.

De belangrijkste grootschalige ICT-investeringen zijn de afgelopen jaren gepleegd in netwerken, *gridcomputing*, e-scienceplatforms en data-infrastructuur. Van deze onderdelen geven we hieronder een korte karakterisering.

---

20 In maart 2009 heeft IC-regie het COMICT-voorstel in het kader van de FES-regeling ingediend. Hierin vormt e-science in een integraal onderdeel.

21 http://85.17.160.176/publicaties%202010/10i-NROI-136%20brief%20kamer-def.pdf

22 Het COMMIT-deelproject is eind 2010 voor een groot deel goedgekeurd.

Netwerken

Nederland kent al decennialang een aaneenschakeling van netwerkprojecten onder leiding van de SURFnet-organisatie. SURFnet maakt deel uit van de SURF-organisatie, waarin Nederlandse universiteiten, hogescholen en onderzoeksinstellingen nationaal en internationaal samenwerken aan innovatieve ICT-voorzieningen. Dat heeft geleid tot een (erkend) uitstekende landelijke netwerkinfrastructuur voor onderzoek. Het laatste project, GigaPort, zorgt bovendien voor een *leading edge* optische *'back bone'* in die infrastructuur. SURFnet ontwikkelt en exploiteert het hybride netwerk SURFnet6 en biedt innovatieve diensten op het gebied van beveiliging, authenticatie en autorisatie, groepscommunicatie en video. Bijna 1 miljoen wetenschappers, docenten en studenten van het Hoger Onderwijs en onderzoek in Nederland hebben via SURFnet dagelijks toegang tot het internet. Het maakt daarmee grensverleggend onderwijs en onderzoek mogelijk. SURFnet jaagt bovendien naar eigen zeggen de Nederlandse telecommunicatie- en internetmarkt aan. De nieuwe mogelijkheden hebben een uitstraling naar andere toepassingsdomeinen en versterken de Nederlandse kenniseconomie.

Gridcomputing

Een *grid* is een relatief nieuwe toepassing van het internet die data, rekenkracht en meetapparaten van over de hele wereld combineert en daarmee zeer veel data uit verspreide bronnen kan ontsluiten, een enorme gedistribueerde rekenkracht kan genereren en daardoor bijzondere toepassingen mogelijk maakt.

De NCF besteedt van haar budget 2,8 miljoen euro aan *grids* voor de Nederlandse wetenschap. NWO heeft dit bedrag aan de NCF toegewezen. Met dit geld moet de ICT-infrastructuur voor universiteiten en onderzoeksinstellingen verder in gereedheid gebracht worden voor het gebruik van grids. Daarnaast heeft het programma BigGrid ongeveer 25 miljoen euro voor gridtechnologie toegewezen gekregen over een periode van 5 jaar.

De afgelopen jaren speelde Nederland een vooraanstaande rol in de internationale internet- en gridwereld. Het computercentrum SARA (Stichting Academisch Rekencentrum Amsterdam) in Amsterdam, waarin de nationale supercomputer is gehuisvest, en het computercentrum in Groningen, waar de LOFAR-computer staat, beschikken over een aanzienlijke expertise om gebruikers van *'high performance distributed computing'* te ondersteunen. Bovendien zorgt SARA voor de ondersteuning van een aanzienlijk deel van de huidige Nederlandse netwerkinfrastructuur. Al deze ontwikkelingen hebben ertoe geleid dat Nederland (aangevoerd door Nikhef, het nationaal instituut voor subatomaire fysica) in een heel vroeg stadium naar *gridcomputing* kon overstappen en nu beschikt over leidende ervaring in enkele van de benodigde specifieke technologieën.[23]

23 Amsterdam organiseerde de eerste Global Grid Forum conferentie, en ons land nam deel in de allereerste internationale wetenschappelijke gridprojecten waardoor we in staat zijn geweest een aantal topexperts in de gridtechnologie op te leiden.

E-science

Omdat e-science steunt op gridtechnologie, is het niet verrassend dat op het Science Park Amsterdam de eerste e-science projecten in Nederland zijn gestart: VLAM (Virtual Laboratory Abstract Machine) en VL-e (Virtual Laboratory for e-science). Die projecten waren, samen met het Engelse e-science-initiatief, zelfs de eerste ter wereld. Dat heeft de Nederlandse onderzoekers aanzienlijke ervaring en internationaal gezien een uitstekende positie opgeleverd, zoals een internationale evaluatiecommissie van VL-e heeft vastgesteld. Het programma VL-e, dat een budget van 40 miljoen euro voor 6 jaar had, waarvan 20 miljoen euro subsidie, en waarin onder andere de UvA, de VU, Nikhef en SARA een belangrijke rol speelden, heeft verscheidene werkende prototypes van e-science-eindgebruikertoepassingen gecreëerd, alsook twee experimenteeromgevingen. Inmiddels is VL-e afgerond en is de nieuwe subsidieaanvraag COMMIT gedeeltelijk goedgekeurd. Verdere voortgang is ondermeer hiervan afhankelijk.

Data-infrastructuur

Open Access speelt een steeds grotere rol in de onderzoekswereld. *Open Access* is het principe dat kennis (in de vorm van publicaties, data, software of anderszins), en dan zeker publiek gefinancierde kennis, openbaar beschikbaar behoort te zijn en ontsloten moet worden. Terwijl het programma *Digital Academic Repositories* (DARE) vele tienduizenden publicaties via internet ontsluit, zijn NWO en KNAW begonnen primair wetenschappelijk materiaal *online* toegankelijk te maken.[24] Daarnaast zorgt *Data Archiving Network Services* (DANS), sinds zijn oprichting in 2005, voor de opslag en blijvende toegankelijkheid van onderzoeksgegevens in de alfa- en gammawetenschappen. DANS ontwikkelt zelf duurzame archiveringsdiensten, bevordert dat anderen dat ook doen, en werkt samen met databeheerders om ervoor te zorgen dat zoveel mogelijk data vrij beschikbaar komen voor gebruik in het wetenschappelijk onderzoek.

Universitaire Medische Centra (UMC's) bundelen intussen hun datakrachten in het Parelsnoerproject. Met dit initiatief willen de UMC's alle kennis en patiëntengegevens bundelen die de afzonderlijke centra hebben op het gebied van bepaalde aandoeningen. Dat gebeurt via landelijke 'biobanken': grote databanken waarvoor de acht instellingen gegevens en materiaal van hun eigen patiënten geanonimiseerd aan elkaar beschikbaar stellen. Vervolgens wordt deze informatie via een speciaal digitaal netwerk toegankelijk gemaakt voor onderzoekers.

24 DARE is een initiatief van de gezamenlijke Nederlandse universiteiten om hun onderzoeksresultaten digitaal toegankelijk te maken. Zie ook NWO's incentive fund open access publications (http://www.nwo.nl/nwohome.nsf/pages/NWOP_82LC99_Eng).

Conclusie

E-science is een belangrijke ontwikkeling waarin in Nederland, net als in andere landen, veel wordt geïnvesteerd. Nederland heeft de afgelopen jaren voortvarend geïnvesteerd in hardware en netwerkinfrastructuur. Ook zijn belangrijke stappen gezet op het gebied van data-infrastructuren, gridcomputing en digitale onderzoeksinstrumenten. Ons land is daarmee goed gepositioneerd om een belangrijke rol in de verdere ontwikkelingen op e-scienceterrein te spelen. Een goede nationale en internationale coördinatie van initiatieven is daarbij van vitaal belang.

3 De invloed van ICT op de wetenschap

"Even though the buzz about e-science often focuses on massive hardware, user interfaces, storage capacity and other technical issues, in the end, the ability of eScience to serve the needs of scientific research teams boils down to people: the ability of the builders of the infrastructure to communicate with its users and understand their needs and the realities of their work cultures." (Alex Voss, National Center for e-Social Science, Manchester, UK)[25]

Inleiding

Het gebruik van ICT in de wetenschap heeft ingrijpende gevolgen voor de manier waarop nieuwe kennis tot stand komt. Het toegenomen belang van ICT daarin komt tot uitdrukking in groeiende hoeveelheden data en in toenemende complexiteit van methoden om informatie te verwerken en te analyseren.[26] ICT-gebruik in de wetenschap heeft daarnaast belangrijke gevolgen voor de relatie tussen wetenschap en samenleving (*in casu* bedrijfsleven, overheid en burgers). Dit hoofdstuk geeft daar een beeld van.

Waar het gaat om de productie van nieuwe wetenschappelijke kennis, onderscheiden we hieronder de *procesmatige* veranderingen in de kennisproductie van de *instrumentele* en de *conceptuele* ontwikkelingen. De belangrijkste *procesmatige* veranderingen zijn de opkomst van nieuwe samenwerkingsvormen en de intensivering van de extramurale samenwerking. Dit heeft geleid tot een veranderende organisatie van het 'wetenschapsbedrijf'. De belangrijkste *instrumentele* ontwikkelingen in de wetenschap hangen samen met het kunnen gebruiken van steeds omvangrijkere en complexere datasets en met de toegenomen mogelijkheden om onderzoeksobjecten te simuleren (onderzoek 'in silico' in plaats van in vitro, in het laboratorium) en om informatie te visualiseren. De belangrijkste *conceptuele* veranderingen die mede onder invloed van deze nieuwe processen en instrumenten in de wetenschap zijn opgetreden, hebben niet alleen te maken met het ontstaan van nieuwe onderzoeksthema's en disciplines, maar ook met het tot ontwikkeling komen van een andere manier van kijken naar objecten van onderzoek en naar de wetenschap zelf. Deze drie veranderingen komen hieronder achtereenvolgens aan de orde. Tenslotte staan we stil bij de veranderingen in de verhouding tussen wetenschap en samenleving.

---

25 http://www.hpcwire.com/hpc/1853013.html

26 De termen gegevens, informatie en kennis onderscheiden wij als volgt van elkaar. Gegevens (of data) worden gegenereerd door menselijke zintuigen, meetinstrumenten of machines, zoals computers. Informatie ontstaat wanneer er door ordening en structurering betekenis aan gegevens is toegekend. Kennis is gekende informatie, is sociaal, veronderstelt een kennend subject.

In silico

The National Virtual Observatory describes itself as "a new way of doing astronomy, moving from an era of observations of small, carefully selected samples of objects in one or a few wavelength bands, to the use of multiwavelength data for millions, if not billions of objects. Such datasets will allow researchers to discover subtle but significant patterns in statistically rich and unbiased databases, and to understand complex astrophysical systems through the comparison of data to numerical simulations." (www.us-vo.org)

Procesmatige veranderingen

Digitalisering vergroot de schaal van netwerkvorming ingrijpend: wie online is, heeft toegang tot miljoenen websites op vele tienduizenden computers en kan met enorm veel mensen via e-mail in contact treden of bestanden uitwisselen. De mogelijkheden van interactie, tussen personen en met machines, zijn door het internet aanzienlijk vergroot. Samenwerking in de wetenschap is daardoor de laatste jaren dan ook enorm toegenomen en van structuur veranderd. Geografische afstanden zijn minder belemmerend geworden; data, gegevens en resultaten worden op steeds grotere schaal gedeeld. Ook de coördinatie van onderzoek is radicaal veranderd met de komst van onmiddellijke terugkoppelingsmogelijkheden die ICT biedt. Onderzoek vindt steeds meer plaats in geografisch verspreide teams waarin expertises vanuit verschillende vakgebieden bijeen worden gebracht. En zelfs gebruikers van kennis – soms op grote afstand – krijgen een steeds grotere rol, bijvoorbeeld in een *'living lab'*: met behulp van ICT biedt het concept *living lab* een nieuwe manier om grote gebruikersgroepen en belanghebbenden te betrekken bij de ontwikkeling van kennis, producten en diensten. *Living labs* genieten in veel gevallen de voorkeur boven gesloten laboratoria, omdat ze een beter beeld geven van menselijk gedrag en sociale processen.

Technologie is maar een deel van het verhaal

Creating and sustaining effective virtual organizations, especially those spanning many traditional organizations, is a complex technical and social challenge. It requires an open technological framework consisting of, for example, applications, tools, middleware, remote access to experimental facilities, instruments and sensors, as well as monitoring and post-analysis capabilities. An operational framework from campus level to international scale is required, as well as a need for partnerships between the various cyberinfrastructure stakeholders. Overall effectiveness also depends upon the appropriate social, governance, legal, economic and incentive structures. Formative and longitudinal evaluation is also necessary both to inform iterative design as well as to develop understanding of the impact of virtual organizations on enhancing the effectiveness of discovery and learning.
(NSF 07-28, Cyberinfrastructure Vision for 21st Century Discovery)

Instrumentele ontwikkelingen

Toepassing van ICT heeft de gegevens waar de wetenschap mee werkt in een nieuwe vorm gegoten en nieuwe onderzoeksinstrumenten in het leven geroepen. Deze komen achtereenvolgens hieronder aan de orde.

Nieuwe gegevensstructuren

Het gebruik van ICT heeft de mogelijkheden informatie te codificeren doen toenemen. Met het codificeren van informatie bedoelen we het gestructureerd vastleggen van gegevens en van betekenisvolle verbanden die tussen die gegevens (verondersteld worden te) bestaan. Ontologiëen, thesauri en semantische modellen spelen in dit verband een belangrijke rol evenals het concept van het Semantic Web.[27] Daarbij kan het gaan om opslag van causale verbanden (zoals in het geval van verklarende modellen) of om normatieve verbanden (zoals bij expertsystemen). Codificatie helpt kennis te objectiveren die anders alleen impliciet bij individuele personen zou berusten. Toepassing van ICT heeft de mogelijkheden verruimd om tegen beperkte kosten gecodificeerde informatie te delen en te gebruiken. Met een model kan eenvoudig worden gerekend en met een expertsysteem kunnen beslissingen worden gesimuleerd.

Met name in de levenswetenschappen zijn enkele succesvolle data-infrastructuren ontwikkeld.[28] Dit is echter in lang niet alle vakgebieden het geval. In de sociale wetenschappen komen initiatieven maar moeizaam van de grond. De ontwikkeling van dit soort gegevensstructuren is complex om ten minste twee redenen. Op de eerste plaats stellen verschillende soorten vakgebieden en toepassingen verschillende eisen aan datastructuren. Er is enorm veel variatie in soorten data en er bestaat geen algemene aanpak die past op alle datacategorieën. Datasets kunnen langs allerlei technische dimensies van elkaar verschillen: omvang, structuur, verbanden tussen de elementen, complexiteit. Daarnaast kunnen ze verschillen waar het gaat om commerciële waarde of privacygevoeligheid. Vergelijk bijvoorbeeld sterrenkundige gegevens of genoomdatabases met volkstellinggegevens of patiëntgegevens, of met webpagina's, gedigitaliseerde boeken of educatieve courseware.

Dit leidt ertoe dat de ontwikkeling van e-science per wetenschapgebied een ander karakter krijgt. Weliswaar kan er een laag van gemeenschappelijke routines ontwikkeld worden die een breed spectrum aan wetenschapsgebieden kan ondersteunen (zogenaamde *'middleware'*, een laag van meer generieke functies tussen bijvoorbeeld het operating systeem en de eigenlijke applicaties), maar verschillende disciplines vereisen daarbovenop hun eigen technieken. Vergelijk bijvoorbeeld de ont-

27 "Semantic Web is a group of methods and technologies to allow machines to understand the meaning – or "semantics" – of information on the World Wide Web", W3c Frequently Asked Questions: http://www.w3.org/2001/sw/SW-FAQ.

28 Zie bijvoorbeeld "The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship. Report of a workshop held in Phoenix, Arizona April 17-19, 2007.

wikkeling van e-science binnen de fysica en de biologie. Beide hebben de beschikking over veel data, maar er is een verschil. Fysici werken vaak deductief, startend vanuit theoretisch afgeleide hypothesen en modellen, met de bedoeling deze te toetsen aan waarnemingen. Biologen werken vaak meer inductief: de causale verbanden worden gezocht door verzamelde gegevens te analyseren met statistische methoden. De modelvorming loopt achter de datacollectie aan. Dit betekent dat biologen andere databasestructuren en softwaregereedschappen nodig hebben dan fysici. Dergelijke verschillen zullen in veel vakgebieden optreden. Het valt dan ook te verwachten dat de ontwikkeling van e-science binnen de diverse alfa-, beta- en gammadisciplines een eigen karakter krijgt en dat deze ontwikkeling tot nieuwe specialisaties binnen de informatica zal leiden.

Op de tweede plaats is de ontwikkeling van nieuwe gegevensstructuren complex omdat voortschrijdende technologische ontwikkeling leidt tot nieuwe meetinstrumenten en -methoden, hetgeen resulteert in een exponentieel groeiende hoeveelheid data. De explosief toegenomen beschikbaarheid van digitale databestanden heeft tot gevolg dat het werken met databases steeds ingewikkelder is geworden. Onderzoek in steeds meer (deel)disciplines is afhankelijk van opbouw, onderhoud en koppeling van grote databestanden. Dit speelt bij uitstek in de levenswetenschappen, vooral het genoomonderzoek, maar ook in disciplines als de hoge-energiefysica, de astronomie, de economie en de sociologie.

Nieuwe methoden en instrumenten

De beschikbaarheid van grote hoeveelheden digitale informatie en van nieuwe gegevensstructuren heeft de ontwikkeling van nieuwe methoden en instrumenten voor onderzoek gestimuleerd. We noemen hieronder *data mining*, simulatie en visualisatie als voorbeelden van technieken die onder invloed van ICT-ontwikkelingen de afgelopen jaren een snelle ontwikkeling hebben doorgemaakt.

Data mining poogt op een geautomatiseerde manier patronen en relaties te ontdekken in grote hoeveelheden gegevens. Het maakt gebruik van database management, statistiek, *machine learning* en patroonherkenning. Het kan nieuwe inzichten in correlaties en suggesties voor mogelijke causale verbanden opleveren, die op basis van gangbare theorieën over het hoofd gezien worden. *Data mining* wordt vaak toegepast op grote hoeveelheden biologische, chemische en medische data.

Een simulatie is een dynamische nabootsing van de werkelijkheid, meestal op basis van een model van die werkelijkheid. Vanuit een gegeven startpunt laat een simulatie zien hoe een situatie zich in de loop van de tijd ontwikkelt. Het model geeft daarbij de regels aan volgens welke deze verandering plaatsvindt. Voordelen van een simulatie zijn dat deze plaatsvindt in een gecontroleerde, welomschreven omgeving, en dat deze kan worden uitgevoerd zonder de werkelijkheid te beïnvloeden.

Veel vakgebieden kennen modellen die zo uitgebreid of complex zijn, dat simulaties zonder een krachtige computer eenvoudig niet mogelijk zouden zijn. Voorbeelden hiervan zijn de simulatiemodellen die gebruikt worden voor weersvoorspellingen, voor beursontwikkelingen of voor kuddegedrag.

Van groot belang binnen e-science zijn zogenaamde workflowmodellen, electronische modellen die de werkwijze van de wetenschapper bij het experimenteren vastleggen.[29] Een verre ambitie van het onderzoek is om via workflowmodellen het doen van experimenten te automatiseren, tot en met het aansturen van de instrumenten in de laboratoria. Een meer korte termijn ambitie is het structureren en standaardiseren van de communicatie tussen wetenschappers via gestandaardiseerde workflows.

Visualisatie, tenslotte, is een onmisbaar instrument van kennisproductie geworden in diverse vakgebieden, bijvoorbeeld de astrofysica. Visualisatie is het vertalen van gegevens of concepten naar beelden als uitdrukkingsvorm. Vanuit de natuurwetenschappen is er steeds meer behoefte aan visualisaties van omvangrijke hoeveelheden reken- en meetgegevens, en in de levenswetenschappen (de geneeskunde, het genoomonderzoek) spelen interactieve visualisatiemethoden en beeldvormende technieken zoals PET, CT of MRI een steeds grotere rol.[30] In weer andere gebieden (zoals scientometrie) is een nieuw type cartografie ontstaan om situaties en analyses inzichtelijk te maken die moeilijk op een andere manier gecommuniceerd kunnen worden.[31] Met name de mogelijkheden om ruimtelijke structuren en processen weer te geven zijn verruimd van het afbeelden op het platte (papieren) vlak tot het gebruik van multimedia- en hypermediatechnieken inclusief Virtual Reality.[32]

In de technische, natuur- en levenswetenschappen zijn deze technieken vooralsnog dieper doorgedrongen dan in de sociale en menswetenschappen. Naar verwachting zal het gebruik van dergelijke methoden in deze wetenschappen flink toenemen. De potentiële impact van ICT op deze disciplines is groot, gegeven de groei van beschikbare databestanden in deze wetenschappen en de complexiteit van de verschijnselen waarmee ze zich bezig houden.

---

29 Zhao et al. (2005).

30 Positron Emission Tomography (PET), Computerized Tomography (CT) en Magnetic Resonance Imaging (MRI) zijn alledrie niet-invasieve methoden die scans opleveren van hetgeen zich binnen het lichaam bevindt.

31 Zie bijvoorbeeld Map of Science voor netwerkvisualisaties van wetenschapsgebieden op http://mapofscience.com/.

32 De term hypermedia is een verbreding van de term hypertekst. Hypertekst is tekst zoals gebruikt op internet, die toegang biedt tot achterliggende tekst door erop 'door te klikken'. Analoog bieden hypermedia (behalve tekst ook grafische voorstelling, audiofragmenten en videofragmenten) toegang tot achterliggende informatie van allerlei vorm.

Kanttekeningen bij *data mining*

De evolutie heeft het menselijk brein niet erg goed geëquipeerd voor de analyse van grote complexe data bestanden en het nemen van beslissingen op basis van statische informatie.[32] Dat komt bijvoorbeeld aan het licht wanneer we te maken hebben met hoogdimensionale datasets (een probleem bekend als de *curse of high dimensionality*). Onze intuïties over patronen in datasets zijn gevormd door het feit dat we in een relatief simpele wereld met drie ruimtelijke dimensies leven. Die intuïties gaan echter niet meer op in complexe, hoogdimensionale datasets. Daarom is voorzichtigheid geboden bij de interpretatie van de resultaten van *data mining* op deze datasets. Toevallige fluctuaties in de distributie van de data kunnen grote gevolgen hebben voor de patronen die *data mining* algoritmen vinden. Bijna al onze medische, economische, wetenschappelijke en forensische datasets hebben een dergelijk hoogdimensionaal karakter. Ze zijn bijgevolg ongeschikt om in ruwe vorm als basis te worden gebruikt voor het via *data mining* vinden van zinvolle patronen die als leiddraad voor beleid kunnen dienen.

Het is daarnaast een illusie dat we veel kennis kunnen vergaren, puur en alleen door veel data te verzamelen. Politici en beleidsmakers lijken soms het idee te hebben dat het verzamelen van veel databestanden en het daarop toepassen van data mining technieken ongekende mogelijkheden biedt voor het op het spoor komen van nieuwe inzichten: ter bestrijding van fraude en terrorisme, voor het doorrekenen van effecten van fiscale maatregelen, voor het vinden van geneesmiddelen voor ziektes, enzovoort. De verwachtingen daarover blijken vaak te hoog gespannen. Als onze *ex ante* kennis over causale verbanden en functionele samenhang in een complexe dataset beperkt is, lopen we al snel aan tegen problemen van *sample bias* in de data en beperkingen van beschikbare rekencapaciteit en energieverbruik, die ons verhinderen valide inzichten rechtstreeks uit de data te destilleren. Het aantal potentieel 'interessante' patronen in een database neemt immers exponentieel toe met de grootte, maar de overgrote meerderheid van dergelijke patronen is ofwel triviaal ofwel een toevalligheid.

Om bovenstaande redenen moet in de wetenschap, de rechtspraak en de politiek uiterst terughoudend worden omgegaan met inzichten verkregen op basis van data mining technieken. In een complete database met sterfgevallen in ziekenhuizen, gekoppeld aan werktijden van verplegend personeel in Nederland, zou men bijvoorbeeld via *data mining* al gauw vele tientallen individuen kunnen vinden, die veel stelliger verdacht zouden kunnen worden dan Lucia de Berk. Ook bij de discussie over ons klimaat spelen dit soort slecht gefundeerde interpretaties van wetenschappelijke data een rol.

33 Een bekend voorbeeld is het zogenaamde driedeurenprobleem (in het Engels Monty Hall problem): Stel je mag in een spelprogramma kiezen uit drie deuren. Achter één deur staat een auto; achter de andere een geit. Je kiest een deur, zeg deur 1, en de presentator, die weet wat achter de deuren zit, opent een andere deur, zeg deur 3, met een geit. Hij zegt dan tegen je: "Zou je deur 2 willen kiezen?" Is het in je voordeel om van keuze te wisselen? Intuïtief lijkt het niet uit te maken – de kans dat de auto achter elk van de overblijvende deuren staat, lijkt een half – maar in feite is de kans dat de auto achter deur 2 staat 2/3 (bij aanvang is de kans dat de auto achter deuren 2 en 3 staat samen 2/3, en dat is na de ingreep van de spelleider nog steeds zo).

Conceptuele ontwikkelingen

Een laatste gevolg van nieuwe toepassingen van ICT in de wetenschap is het stimuleren van inhoudelijke veranderingen. Aan de ene kant heeft het gebruik van ICT-instrumenten ertoe geleid dat nieuwe aspecten van de werkelijkheid in het vizier van de wetenschap terecht zijn gekomen, net zoals in het verleden de microscoop en de telescoop de blik van wetenschappers op nieuwe werkelijkheden hebben gericht. Aan de andere kant heeft datzelfde gebruik van ICT-instrumenten ertoe geleid dat wetenschappers op een nieuwe manier zijn gaan kijken naar aspecten van de werkelijkheid die ze al vanouds in het vizier hadden.

Studie van complexe systemen en processen binnen allerlei disciplines is door toepassing van ICT binnen het bereik van wetenschappelijke analyse gekomen. De hierboven beschreven ontwikkeling van methoden voor *data mining*, simuleren en visualiseren hebben hieraan bijgedragen. Maar er zijn ook nieuwe onderzoeksterreinen opgebloeid. Sommige daarvan zijn het gevolg van de digitalisering van de samenleving, bijvoorbeeld een onderzoeksgebied als *'webometrics'*, dat internetdata gebruikt om sociaalwetenschappelijk onderzoek mee te doen. Online databases hebben disciplines als biomedicine een nieuwe richting doen uitgaan.[34] De onderzoeksobjecten van mediastudies, kunstmatige intelligentie, artificial life, virtuele etnografie, internetstudies en natuurlijk de informatica zelf zijn vakgebieden die hun bestaan mede aan ICT te danken hebben.

Daarnaast is er ook binnen de wetenschap een nieuwe manier van kijken naar de werkelijkheid ontstaan. Zoals de ontwikkeling van de statistiek ons de ogen heeft geopend voor het stochastische karakter van veel van de verschijnselen die we om ons heen zien, zo heeft de ontwikkeling van ICT onze aandacht gericht op de complexe onderlinge samenhang van allerlei aspecten van de wereld om ons heen. Deze samenhang is complex in de zin dat er heel veel interacties tussen de elementen binnen een systeem bestaan. In de biologie, in de politieke en sociale wetenschappen, in de economie en in vele andere disciplines heeft ICT de aandacht van wetenschappers op deze interacties gericht en hen de instrumenten verschaft om deze te bestuderen en te modelleren. Dat heeft geleid tot een zekere verschuiving in het wereldbeeld, van een beeld dat gebaseerd is op objecten en functies naar een beeld dat gebaseerd is op complexe, evoluerende netwerken waarin het gaat om patronen, correlaties, samenhang en emergente verschijnselen op systeemniveau.[35] Deze verschuiving is niet alleen in de natuurwetenschappen waar te nemen, maar ook in de sociale wetenschappen.[36]

---

34 Lenoir, T. (1999).
35 Ihde, D. (2000).
36 Zie bijvoorbeeld Rheingold, H. (2003), Johnson, S. (2001)

De metamorfose van de biologie en het ontstaan van de bio-informatica

In het midden van de jaren '60, toen Jacob en Monod werkten aan de genetische code, begonnen de ontwikkelingen in ICT die leidden tot computerarchitecturen en algoritmen voor het modelleren van chemische structuren en voor het simuleren van chemische interacties die het mogelijk maakten theorie en experiment op geheel nieuwe manieren te combineren. De nieuwe computational science heeft, gecombineerd met visualisatietechnieken, een enorme impact gekregen op het gebied van de biochemie, de moleculaire dynamica en moleculaire farmacologie. Hiermee werd het mogelijk beelden en signalen uit camera's en sensoren om te zetten in geabstraheerde visuele representaties in de vorm van symbolen en structuren.

Maar ook andere informatietechnologieën hebben invloed gehad op de metamorfose van de biologie. In de jaren '90 werden moleculair biologen overspoeld met nieuwe gegevens door het gebruik van nieuwe instrumenten als klonen, restrictie-enzymen, eiwit sequencing, genexpressie en productamplificatie. Deze gegevens werden opgeslagen in steeds grotere elektronische databanken van genetische kaarten, atomaire coördinaten voor chemische- en eiwitstructuren, en eiwitsequenties.

Deze ontwikkelingen transformeerden de biologie tot een datagedreven wetenschap, een wetenschap waarin het beschikbaar komen van de empirische gegevens van een domein (bijvoorbeeld een genoom) sterk vooruitloopt op het begrip van de structuur en de wetmatigheden van dit domein. Biologen hebben zich met deze gegevens begeven op het pad van de informatiewetenschap, de bio-informatica. Dat betrof de ontwikkeling van zoekalgoritmen om structuren en patronen in biologische gegevens te identificeren en de toepassing daarop van kunstmatige intelligentie en expertsystemen. Aanvankelijk richtte de bio-informatica zich vooral op het genomicsonderzoek. Nu heeft het gebied zich verbreed tot de levenswetenschappen als geheel.

Om levensprocessen werkelijk goed te begrijpen is het noodzakelijk biologische systemen integraal te beschouwen. Het gaat daarbij om het met elkaar in verband brengen van informatie op elk biologisch aggregatieniveau in mathematische modellen die in staat zijn eigenschappen van biologische systemen als geheel weer te geven. Dat is de uitdaging waar de levenswetenschappen, en de systeembiologie in het bijzonder, voor staan.

Niet alleen onze kijk op de kenobjecten van wetenschappelijk onderzoek is veranderd onder invloed van het ter beschikking komen van ICT, maar ook onze kijk op het proces van wetenschappelijk onderzoek zelf. Een blijk daarvan is te vinden in de vaak gesignaleerde overgang in de aard van kennisontwikkeling van 'mode 1' naar 'mode 2'. In de woorden van de bedenkers van de term: *"Mode 1 is discipline-based and carries a distinction between what is fundamental and what is applied;*

this implies an operational distinction between a theoretical core and other areas of knowledge such as engineering sciences, where the theoretical insights are translated into applications. By contrast, Mode 2 knowledge production is transdisciplinary. It is characterised by a constant flow back and forth between the fundamental and the applied, between the theoretical and the practical. Typically, discovery occurs in contexts where knowledge is developed for and put to use, while results – which would have been traditionally characterised as applied – fuel further theoretical advances"[38]

Sinds deze transformatie van de wetenschap voor het eerst gesignaleerd is, heeft deze zich verder uitgekristalliseerd. Waar wetenschap 1.0 omschreven kan worden als reductionistische kennisproductie in de goed gedefinieerde omstandigheden van gecontroleerde experimenten, daar is wetenschap 2.0 het bestuderen van snel veranderende socio-technische systemen die niet reproduceerbaar zijn in een laboratorium. Wetenschap 1.0 blijft essentieel, maar ontwikkeling van kennis op systeemniveau vraagt om wetenschap 2.0. Dat vereist de nodige veranderingen binnen het wetenschappelijk bedrijf: *"Advancing Science 2.0 will require a shift in priorities to promote integrative thinking that combines computer science know-how with social science sensitivity. Science 2.0 researchers who develop innovative theories, hypothesis testing based on case study research methods, and new predictive models are likely to lead the way. The quest for empirical validity will drive research beyond what laboratory-based controlled studies can provide, while replicability and generalizability will be achieved with greater effort through multiple case studies."*

Kenmerkend voor de transitie van een reductionistische 'mode 1' benadering van kennisproductie naar een meer omvattende en holistische 'mode 2' benadering is op de eerste plaats dat de aandacht in het onderzoek op systeemniveau komt te liggen. Het gaat daarbij vooral om analyse van complexe, adaptieve (zelforganiserende, lerende) systemen. Om deze te analyseren, zoekt men nieuwe benaderingen in het modelleren en simuleren van die systemen op de computer. Het is daarbij de uitdaging het systeem simultaan op meer niveaus van analyse te bestuderen, waarbij de interactie tussen al die niveaus wordt meegenomen.[39] Ook kunnen evolutie of leerprocessen in het model worden opgenomen, op individueel niveau en op systeemniveau.
De noodzaak van een integrale systeembenadering in de wetenschap wordt door de indieners van een voorstel voor een Nederlands centrum voor e-science als volgt omschreven: "In de moderne wetenschap en techniek en in de maatschappij worden we meer en meer geconfronteerd met complexe problemen die niet afzonderlijk kunnen worden bestudeerd, maar alleen binnen de context van het volledige

38 Shneiderman, B. (2008)..
39 Zie voor een voorbeeld het tekstkader over biologie. In de sociale wetenschappen kan het bijvoorbeeld gaan om interactie tussen het individuele niveau, dat van sociale netwerken, en dat van organisaties.

systeem waarvan ze deel uitmaken. De studie van dit soort problemen wordt vaak aangeduid als system-level science. Een van de beste definities hiervan is: de integratie van verschillende beschikbare kennisbronnen omtrent de samenstellende delen van een complex systeem met het doel begrip te verkrijgen van de systeemeigenschappen van het totale systeem."[40]

Kenmerkend voor 'mode 2' wetenschap is op de tweede plaats dat naast natuurlijke aspecten ook sociale en culturele aspecten in beeld komen. Een holistische wetenschap is transdisciplinair. Tenslotte komt ook de onderzoeker zelf en zijn manier van onderzoek doen in beeld. Daarmee krijgt wetenschappelijk onderzoek een reflexief element: het proces van kennisontwikkeling wordt mede onderwerp van kennisontwikkeling.

Aandacht voor dynamica van complexe systemen in Nederland

In lijn met het bovenstaande, heeft NWO recent een programma gestart dat tot doel heeft complexe systemen met nieuwe methoden te onderzoeken. NWO beschrijft dit programma op zijn website als volgt: "Nieuwe complexe methodes, strategieën en benaderingen zijn in toenemende mate vereist voor de sociale, economische en financiële analyses van de toekomst. Onderzoek naar complexe systemen startte vanuit de exacte wetenschappen, maar is uitgebreid naar onder meer de gedragswetenschappen en de economie, en is nu een van de snelst groeiende wetenschapsgebieden. Het onderzoek richt zich op systemen van verschillend karakter, zoals complexe processen bij energiedistributienetwerken, de verspreiding van micro-organismen, reactiesnelheden in een chemische reactor, fluctuaties in aandelenkoersen en veranderingen in klimaat.

Complexe systemen gedragen zich deterministisch, maar hun evolutie is toch niet voorspelbaar vanuit de begincondities. Hiervoor is meer inzicht nodig in hun dynamica en de concepten van deterministische chaos. Ondanks de enorme variëteiten zien we soms gelijksoortig gedrag. Dat is aanleiding om te verwachten dat door een juiste meting van de vaak gigantische hoeveelheden gegevens samen met wiskundige modellering en analyse, een beter inzicht in complexe processen valt te verkrijgen. De productiesector kan hiervan bijvoorbeeld profiteren via een verbeterde procescontrole, de biosector door een verkregen begrip van complexe organismen, en de energiesector door een verbeterde beheersing van dynamische processen."[41]

---

40 I Uit het voorstel "Nederlands centrum voor e-Science ten behoeve van systeemniveau-onderzoek". De Commissie van Wijzen heeft de VL-e-directie geadviseerd het initiatief te nemen voor een Nederlands centrum voor e-Science ten behoeve van systeemniveau-onderzoek. Onder auspiciën van de UvA en de VU is daartoe deze notitie voor de Commissie Nationale Roadmap Grootschalige Onderzoeksfaciliteiten voorbereid.

41 Zie NWO (2006); zie voor actuele informatie http://www.nwo.nl/nwohome.nsf/pages/NWOA_7BUJ6J.

De laatste zin hierboven en de noot bij dit programma dat "het aansluit bij de thema's van VNO/NCW, de prioriteiten van het ministerie van EZ en de thema's van TNO en de GTI's", illustreert de veranderende verhouding tussen het (fundamenteel) wetenschappelijk onderzoek en de samenleving (zie het volgende hoofdstuk).

Veranderingen in de verhouding van wetenschap en samenleving

Waar de vorige paragrafen keken naar de veranderingen die mede onder invloed van ICT binnen de wetenschap zijn opgetreden, gaat deze laatste paragraaf over de veranderingen in de maatschappelijke inbedding van de wetenschap en over de veranderingen in de interactie tussen wetenschap en samenleving. Rondom kennisontwikkeling zijn de laatste jaren nieuwe organisatievormen tot ontwikkeling gekomen. We zijn in een informatie- en netwerksamenleving terecht gekomen, en dat geldt *a fortiori* voor het wetenschappelijk onderzoek.[42] Steeds meer werken kennisinstellingen, overheden en bedrijven samen in kennisontwikkeling en innovatie, waarbij deze samenwerking in open netwerkverband wel wordt aangeduid als het functioneren van een *Triple-Helix of university-industry-government relations*.[43] Langs deze weg kunnen overheden en bedrijven in toenemende mate actief betrokken raken bij onderzoek en ontwikkeling. Het feit dat overheden en bedrijven een meer actieve rol op zich hebben genomen in netwerken van kennisproductie ondersteunt uiteenlopende ontwikkelingen als de transitie naar 'mode 2' wetenschap die hierboven is beschreven, maar ook het cyclische innovatieproces en de trend naar 'open innovatie'.

De toegenomen mogelijkheden voor interactie en terugkoppeling betekenen voor de kennisontwikkeling dat de gebruikers van kennis centraler komen te staan: *"In modern times, science has always 'spoken' to society; indeed science's penetration of society is close to being a defining characteristic of modernity. But society now 'speaks back' to science."*[44] Het nut en de toepasbaarheid van wetenschappelijke kennis worden steeds meer van belang geacht. De eigen dynamiek van onderzoek en technologie zijn dan ook steeds minder het startpunt voor verdere kennisontwikkeling en innovatie, en de behoeften van gebruikers steeds meer. Bovendien wordt onderzoek veel sterker integraal en cyclisch georganiseerd. Deze ontwikkeling wordt gefaciliteerd door de terugkoppelingsmechanismen die ICT mogelijk maakt. Samenvattend kan men stellen dat de samenleving kennisintensiever is geworden, en de kennisproductie 'samenlevingsintensiever'. In een kennissamenleving wordt de wetenschap nadrukkelijker aangesproken op haar taak maatschappelijke ontwikkelingen te doordenken, te voorspellen en te begeleiden.

---

42 Zie over de netwerksamenleving bijvoorbeeld Castells (1996).
43 Etzkowitz and Leydesdorff (2000).
44 Aldus Nowotny, Scott en Gibbons (2001), p. 50.

Conclusie

De toepassing van ICT in de wetenschap heeft een aantal radicale gevolgen gehad voor de manier waarop wetenschap wordt bedreven, voor de instrumenten die worden gebruikt, en zelfs voor de manier waarop wetenschappers tegen hun object van onderzoek aankijken. Procesmatig heeft het geleid tot veel meer en veel snellere samenwerking over instellingsmuren en landsgrenzen heen. Belemmeringen van tijd en afstand in wetenschappelijke netwerken zijn sterk gereduceerd.
In de onderzoekspraktijk zijn nieuwe technieken en instrumenten in zwang geraakt die wetenschappers in staat hebben gesteld grote en complexe datasets te analyseren en om onderzoeksobjecten te visualiseren en hun gedrag te simuleren. De deductief-analytische en de inductief-empirische aanpak van wetenschappelijk onderzoek kunnen nu worden aangevuld met een op exploratie en op simulatie *in silico* gebaseerde aanpak.

Dit alles heeft zijn invloed gehad op de ontwikkelingsrichting van allerlei wetenschappen en op het ontstaan van nieuwe disciplines. Het heeft ook onze kijk op de aard van de werkelijkheid aangevuld en ons methodologisch repertoire in de ontwikkeling van nieuwe kennis fundamenteel verbreed.

Tenslotte heeft het ertoe bijgedragen dat de wetenschap een nieuwe plaats in de samenleving heeft verworven. Het heeft geholpen de wetenschap uit zijn ivoren toren te halen en maatschappelijk in te bedden, met alle voordelen en wederzijdse verwachtingen die daarmee gepaard gaan.

4 Vragen tot besluit

Hierboven hebben we gezien dat Nederland veel in de toepassing van ICT in de wetenschap heeft geïnvesteerd en een goede uitgangspositie heeft op het gebied van e-science. Daarmee is ons land uitstekend geëquipeerd om een belangrijke rol te spelen in de verdere ontwikkeling van e-science en in de toepassing van de vruchten daarvan in het wetenschappelijk onderzoek.

Vervolgens hebben we beschreven hoe onder invloed van ICT de laatste decennia een revolutie in de wetenschap op gang is gekomen. Deze revolutie heeft allereerst betrekking op de manier waarop de wetenschap functioneert en waarop wetenschappers met elkaar samenwerken en op de methoden en instrumenten die wetenschappers in het onderzoek gebruiken. Daarenboven heeft hij betrekking op het epistemologische karakter van wetenschappen, op de aard van het wetenschappelijk proces en op het functioneren van de wetenschap in de samenleving.

Dit roept een paar vragen op die wij hier graag tot besluit willen agenderen:

- De aard van het onderzoeksproces verandert. Kennisontwikkeling vindt steeds meer plaats in ruimtelijk gedistribueerde, door ICT gefaciliteerde netwerken. Wat impliceert dit voor *de optimale organisatie van het onderzoek in kennisinstellingen*? Hoe kunnen kennisinstellingen het best inspelen op de nieuwe mogelijkheden tot samenwerking in de wetenschap? Wat betekent het transdisciplinaire karakter van 'wetenschap 2.0' voor de organisatie van de universiteit? Steeds sneller komen nieuwe onderzoeksthema's en onderzoeksvelden op die een beslag leggen op middelen en onderzoeksinfrastructuur. In hoeverre zijn de huidige institutionele arrangementen toegerust om deze snelle dynamiek in goede banen te leiden?
- De veranderende aard van de wetenschap en de veranderende verhoudingen tussen wetenschap en samenleving komen tot uitdrukking in de ontwikkeling van 'mode 2' kennisontwikkeling , van cyclische innovatieprocessen en van open innovatie. Wat betekent dit voor *de rol die de overheid tegenover de wetenschap zou moeten spelen*? In het bijzonder, wat betekent het i) voor de gewenste betrokkenheid van de overheid bij de kennisontwikkeling, ii) voor de aard van de aansturing en de financiering van het wetenschappelijk bedrijf, en iii) voor de verwachtingen die de overheid tegenover de wetenschap mag koesteren, bijvoorbeeld ten aanzien van het maatschappelijk nut van op te leveren kennis?
- De toenemende openheid van het wetenschapssysteem, het groeiende belang van netwerken en de groter wordende schaal van de systemen waarin kennis wordt geproduceerd maakt dat Nederland steeds minder in staat is een eigenstandig wetenschapsbeleid te voeren en steeds meer onderdeel

wordt – en afhankelijker wordt – van wat elders in de wereld gebeurt. Hoe kan de Nederlandse overheid zich het best opstellen om ervoor te zorgen dat de integratie van de Nederlandse wetenschap in internationale netwerken optimaal gewaarborgd blijft? Wat dient zij te doen om te bewerkstelligen dat Nederlandse wetenschappers toegang hebben tot kennis en faciliteiten wereldwijd? Welke positie kan ze het best innemen als het gaat om *open access* en open standaarden?

De toepassing van ICT in de wetenschap heeft gezorgd voor een geleidelijke maar zeer ingrijpende revolutie. De mogelijkheden in het onderzoek zijn enorm toegenomen.[45] De praktische consequenties van deze ontwikkeling voor kennisinstellingen, voor de overheid, voor bedrijven en voor burgers tekenen zich langzaam af.
De vraag naar de beste strategieën voor deze partijen om van de geboden kansen gebruik te maken hangt vooralsnog boven de markt.

45 Zonder daarmee te willen suggereren dat de wetenschap zich langs deze weg aan alle beperkingen kan ontworstelen – zie bijlage 1.

b1 Grenzen aan e-science

E-science staat momenteel sterk in de belangstelling. Een gevaar van deze aandacht is dat de verwachtingen te hoog gespannen raken. Bij sommige beleidsmakers en onderzoekers bestaat de indruk dat e-science een soort panacee is. Het risico van nieuwe hypes met overdreven claims, zoals die rond de ontrafeling van het menselijk DNA, is niet denkbeeldig. Het managen van de verwachtingen is in deze fase van de ontwikkeling van e-science van vitaal belang. De inzet van e-science is essentieel voor de voortgang van de wetenschap maar sommige domeinen zijn zo complex dat we niet op korte, of zelfs lange, termijn mogen verwachten dat onze theorieën convergeren. Met een variant op de bekende skeptische houding uit de antieke filosofie wordt het universum van de informaticus gedomineerd door de volgende obstakels:

- De meeste problemen zijn onberekenbaar.
- Als ze al berekenbaar zijn, dan hebben we vaak te weinig data.
- Als er genoeg data zijn, dan kost het vaak teveel tijd.
- Als er genoeg tijd is, dan kost het vaak teveel energie.

E-science confronteert ons echter ook met fundamentele grenzen van de wetenschappelijke onderneming. Het risico van 'verdrinken in data' is reëel. Een elementair begrip van resultaten uit de informatica en leertheorie kan ons helpen te begrijpen welke fundamentele grenzen er aan de mogelijkheden van e-science gesteld zijn en daarmee al te hoog gespannen verwachtingen te temperen. Hieronder bespreken we kort enige inzichten rond complexiteits issues, energieverbruik en VC-dimensie en hun mogelijke implicaties voor beleidsontwikkeling.

De notie van een algoritme is zo fundamenteel en algemeen in de informatica dat ze niet exact gedefinieerd kan worden. De beste omschrijving is iets als: een systematische methode om een probleem op te lossen volgens vaste regels. Dat kan variëren van het bijhouden van pokerscores op een bierviltje tot het oplossen van een complex planningsprobleem. Allemaal hebben we op school formele algoritmen geleerd voor het oplossen van eenvoudige wiskundige vergelijkingen: optellen, aftrekken, delen, vermenigvuldigen, enzovoort. Met de introductie van een formeel machinemodel door Alan Turing in 1936 beschikken we over een abstract referentiemodel voor het bestuderen van algoritmes. Fundamentele vragen over de kracht en efficiëntie van berekeningen kunnen op die manier geanalyseerd worden. Voor de analyse van vraagstukken rond e-science die over het algemeen het gebruik van complexe algoritmen op grote dataverzamelingen met zich meebrengen, is een elementair begrip van deze materie belangrijk. We geven een aantal kernbegrippen:

1. *Onbeslisbaarheid.*
Turing beschreef zelf al een eerste resultaat: niet alles wat we zouden willen weten is ook berekenbaar. Er bestaan onbeslisbare problemen. Zelfs al bestaat het antwoord voor een bepaald wiskundig probleem in een Platonische ideale wiskundige wereld, dan hoeft dat nog niet te betekenen dat er een eindig computerprogramma bestaat dat een dergelijk probleem ook in een eindig aantal stappen kan oplossen. Onbeslisbaarheid is het algoritmische analoog van Gödels onvolledigheidsstelling. Dit soort resultaten deden de wiskunde op zijn grondvesten schudden, maar ze leken aanvankelijk weinig praktische waarde te hebben. De gevolgen voor ambities van e-science zijn nihil. De klasse van onbeslisbare problemen was vrij overzichtelijk en speelde in de praktische wetenschap (net als in de economische praktijk) nauwelijks een rol. Dat veranderde toen men in de zestiger jaren van de vorige eeuw met toen nog vrij logge computers werkelijke problemen probeerde op te lossen. Er bleek een aantal problemen te zijn die in theorie wel beslisbaar waren, maar waar computers in de praktijk hun tanden op stukbraken. Dat leidde tot de studie van de complexiteit van algoritmen en problemen.

2. *Complexiteitsklassen.*
De studie van complexiteitsklassen probeert een antwoord te vinden op de volgende vraag: als ik weet hoe groot het probleem is, kan ik dan voorspellen hoe lang een computer erover doet om het op te lossen? Een paar ruwe complexiteitsklassen kan men eenvoudig beschrijven op basis van een relatie tussen de tijd t die het kost om een probleem op te lossen en twee andere parameters, n = de grootte van het probleem en c = een constante:
   - Logaritmische complexiteit: $t = c \log n$
     Voorbeeld: het zoeken in een beslissingsboom.
   - Lineaire complexiteit: $t = c\, n$
     Voorbeeld: het zoeken in een lijst.
   - Polynomiale complexiteit: $t = n^c$
     Voorbeeld: het oplossen van een legpuzzel.
   - Exponentiële complexiteit: $t = c^n$
     Voorbeeld: het oplossen van een logisch probleem (zoals het maken van beladingschema's of roosters en dergelijke).

   Problemen met exponentiële complexiteit worden *intractable* genoemd. Ze kunnen in principe in eindige tijd worden opgelost maar voor problemen van enige omvang is deze tijd zo groot dat het in de praktijk ondoenlijk is omdat de rekentijd die nodig is op zelfs de sterkst denkbare computer de leeftijd van ons universum overschrijdt. Helaas hebben de meeste problemen die economisch en wetenschappelijk interessant zijn *(planning, scheduling, hypotheseconstructie)* naar het zich laat aanzien exponentiële complexiteit. Dit leidt tot een ook voor beleidsmakers interessant inzicht: *voor de meeste*

economische en wetenschappelijke problemen van enige complexiteit zullen computers in de praktijk nooit de theoretisch optimale oplossing vinden, ook al bestaat die wiskundig wel en is die berekenbaar. Dat laat onverlet dat computers erg nuttig zijn bij het benaderen van oplossingen, maar het noopt ons wel de verwachtingen wat te temperen. Voor de praktijk van e-science ligt de zaak zelfs nog wat ongemakkelijker. Zie het volgende punt.

3. *Het Landauer principe.*

Al in 1961 heeft de IBM-er Rolf Landauer op basis van een eenvoudig thermodynamisch gedachte-experiment een ondergrens geformuleerd voor de minimale hoeveelheid warmte die vrijkomt bij het wissen van een bit aan informatie. Dit is bekend geworden als *Landauer's bound: kT ln 2*. Hier is k de Boltzmannconstante ($1.380650 \times 10^{-23}\ JK^{-1}$) en T de absolute temperatuur. In theorie geeft dit ons een mogelijkheid het minimale energieverbruik te schatten van het laten draaien van bepaalde algoritmen op een computer. In de praktijk zal het energieverbruik vele malen groter zijn, maar voor zogenaamde niet-reversibele algoritmen geeft het een benedengrens. Qua complexiteit is de factor 10^{-23} uit de Boltzmannconstante doorslaggevend. Voor normale administratieve toepassingen is deze factor verwaarloosbaar, maar binnen de context van de extreem grote databestanden van e-science begint het een factor van betekenis te worden. De complexiteit van het menselijk brein wordt bijvoorbeeld geschat op 10^{14} bits. Het draaien van een simpel patroonherkenningalgoritme met een derdemachtscomplexiteit op een dergelijke data verzameling, leidt tot een minimale warmteproductie van $10^{19}\ J$. Dat is in de orde van grootte van de totale energie consumptie van de wereld per jaar.[46] Ook als we alle data die een menselijk brein karakteriseren zouden hebben, dan nog hoeft het niet zo te zijn dat we modellen die de data verklaren eenvoudig kunnen vinden met behulp van een computerprogramma. Zelfs het draaien van simpele algoritmen op extreem grote bestanden leidt tot een enorme energiebehoefte. *Adequaat omspringen met energie is een van de belangrijkste agendapunten voor e-science in e 21^{e} eeuw.*

4. *De VC-dimensie.*

Stel dat we willen leren hoe we een vliegtuig moeten repareren zonder handboek. Hoeveel reparaties moeten we dan observeren voordat we een redelijk beeld hebben van die taak? Het is duidelijk dat het antwoord op die vraag ondermeer afhangt van de complexiteit van het vliegtuig. Stel dat het vliegtuig een miljoen onderdelen heeft die allemaal onafhankelijk van elkaar onder bepaalde omstandigheden stuk kunnen gaan. De leertheorie vertelt ons dat we dan in de orde van 100 miljoen reparaties moeten evalueren voordat we een redelijk beeld hebben van wat er allemaal mis kan gaan. Zoveel reparaties zijn er echter nog nooit aan vliegtuigen gedaan. We hebben dus niet voldoende observaties om zonder verdere modelinfor-

46 Zie http://en.wikipedia.org/wiki/Orders_of_magnitude_%28energy%29..

matie een goede theorie over het repareren van dergelijke vliegtuigen te ontwikkelen. Daarom levert de bouwer van het vliegtuig reparatiehandboeken. Daarmee kunnen reparaties gepland worden op basis van modelinformatie over het vliegtuig. Met het bestuderen van dit soort leerproblemen houdt de theorie van het PAC (Probably Approximately Correct) leren zich bezig. Een belangrijke maat is de zogenaamde VC (Vapnik-Chervonenkis) dimensie. De VC-dimensie van een probleem geeft aan hoeveel verschillende theorieën we over de oplossing van dat probleem kunnen hebben. Een interessante vraag in verband met e-science is wat de VC-dimensie is van de verschijnselen die men in dat verband bestudeert. Wat is bijvoorbeeld de medisch relevante VC-dimensie van de mens of het klimaat. Het menselijk lichaam is een zeer complex systeem. Stel dat we die complexiteit, heel conservatief, schatten in termen van de complexiteit van ons brein. In dat geval moeten we, naar analogie met ons vliegtuigvoorbeeld 10^{16} patiënten zien voordat we enig beeld hebben van wat er allemaal mis kan zijn. Dat zijn meer mensen dan onze aarde ooit zal kunnen herbergen. We zitten nu op een totale wereldbevolking van $6{,}8 \times 10^9$ personen. Dat wil zeggen dat we het leeuwendeel van de ziekten die mensen in principe kunnen krijgen nooit zullen tegenkomen. Eenzelfde argumentatie geldt voor de studie van ons klimaat, de menselijke cel, het leven, et cetera. Als we het zogenaamde sequenome beschouwen als de totale ruimte van genetisch mogelijke levensvormen, dan heeft de evolutie van alle leven op aarde tot dusver slechts een zeer klein deel van die ruimte verkend. Het is niet duidelijk of dat voldoende kan opleveren om de onderliggende mechanismes ook te begrijpen. Dit is een inzicht dat ook voor beleidsmakers van belang is: *de meeste databases met observaties van complexe verschijnselen waar de wetenschap momenteel mee werkt, zijn fundamenteel 'sparse'. Dat wil zeggen dat we niet noodzakelijkerwijs in staat zijn voldoende observaties te doen om tot adequate theorievorming te komen*.

5. *Powerlaws.*

Een typische indicatie van complexiteit van een probleemgebied, dat samenhangt met *sparse data*, is het optreden van zogenaamde powerlaws in de distributie van de data. Er is bijvoorbeeld een powerlawrelatie tussen de grootte van steden en de macht van hun frequentie van de vorm: $f(x) = ax^k$ (k en a constant). Dit soort distributies komt veel voor bij complexe problemen: de frequentie van woorden in tekst, de frequentie van ziekten in een populatie, het optreden van storingen in telefoonverkeer, de frequentie van extreme weersverschijnselen, enzovoort. Deze distributies hebben een aantal interessante eigenschappen: ze hebben geen gemiddelde (ofwel het gemiddelde is 'oneindig') en onze leeralgoritmen convergeren er niet op omdat er steeds een vaste fractie nieuwe gevallen optreedt als we meer data vergaren. Voor teksten geldt bijvoorbeeld een dergelijke *powerlaw*. Hoeveel tekst we ook verzamelen, als we meer tekst erbij krijgen zit er altijd een vast percentage nieuwe woorden bij die we nog nooit gezien hebben. Dit geldt zowel voor een moeder die met een baby spreekt als voor de wetenschappelijke artikelen in een encyclopedie. *Voor beleidsmakers is het herkennen van domeinen met powerlawverdelingen van*

belang omdat een aantal voor de hand liggende beslissingsmodellen in deze domeinen niet werkt. Neem bijvoorbeeld ziektekosten. Het is bekend dat de distributie van ziektes een powerlawverdeling volgt.[47] Dit heeft gevolgen voor het beleid. De beslissing alle patiënten te helpen met behulp van wetenschappelijk onderzoek leidt tot oncontroleerbare kosten, puur vanwege het feit dat powerlawverdelingen geen (ofwel, een oneindig) gemiddelde hebben. Er komen altijd nieuwe patiënten met nieuwe soorten ziektes bij. De neiging om een *evidence based policy* te volgen leidt echter tot het systematisch uitsluiten van een vast, substantieel, percentage van de bevolking, simpelweg omdat er te weinig cases zijn. De politiek lijkt zich te weinig bewust te zijn van dit soort inzichten en komt vaak met rigoureuze voorstellen die geen recht doen aan de onderliggende complexiteit van deze materie.
Het verplichtstellen van standaard behandelcodes is een voorbeeld hiervan. Voor verschijnselen met een powerlawverdeling bestaan er geen dekkende eindige codesystemen. De medische wetenschap ontkomt, gegeven de complexiteit van haar domein, niet aan de noodzaak naast *evidence based* technieken ook terug te grijpen op *model based* redeneren en redeneren naar analogie.

47 http://healthcare-economist.com/2006/08/10/the-long-tail-of-healthcare-why-25-million-americans-have-a-rare-disease/

b2 Literatuuroverzicht

- Atkins, D.E. (ed.), (2003) 'Revolutionizing Science and Engineering Through Cyberinfrastructure. Cyberinfrastructure for Education and Learning for the Future: a vision and research agenda.'
- Barjak, F. (2006) 'Research productivity in the internet era', Scientometrics, 68(3), 343-360.
- Berman, F. and Brady, H. (2005) 'Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences', Retrieved October 3, 2006.
- Castells, M. (1996) 'The Rise of the Network Society', Oxford, Blackwell.
- Cook Network Consultants (2010) 'Cook Report on Internet Protocol'.
- Cozzens et al. (1990) 'Research systems in transition'.
- Drori, G.S., Meyer, J.W., Ramirez, F.O. and Schofer, E. (2003) 'Loose Coupling in National Science: Policy versus Practice', in G.S. Drori, J.W. Meyer, F.O. Ramirez and E. Schofer (eds.) Science in the Modern World Polity: Institutionalization and Globalization , pg. 155-73. Stanford, CA:, Stanford University Press.
- Etzkowitz, H. and Leydesdorff, L. (1997) 'Triple Helix'.
- Etzkowitz, H. and Leydesdorff, L. (2000) 'The Dynamics of Innovation: From National Systems and 'Mode 2' to a Triple Helix of University-Industry-Government Relations.' Introduction to the special 'Triple Helix' issue of Research Policy 29(2): 109-123.
- Galison P. and Hevly B. (eds.) (1992) 'Big Science: The Growth of Large-Scale Research', Stanford: Stanford University Press.
- Gibbons, M. C. and Limoges, H., et al. (1994) 'The New Production of Knowledge'. London, Sage.
- Ihde, D. (2000) 'Epistemology engines', Millennium Essay, Nature 406, 21.
- Heimeriks, G. and Vasileiadou, E. (2008) 'Changes or Transition? Analysing the Use of ICT's in Science.' Social Science Information 47(1).
- Johnson, S. (2001) 'Emergence: The connected lives of ants, brains, cities and software'.
- Kircz, J. (2004) 'E-based Humanities and E-humanities on a SURF platform', Amsterdam: KRA Publishing research.
- Kling, R. and McKim, G. (1998) 'The Shaping of Electronic Media in Supporting Scientific Communication: The Contribution of Social Informatics', presented at 'European Science and Technology Forum: Electronic Communication and Research in Europe' Darmstadt/Seeheim, 15 to 17 April 1998. http://academia.darmstadt.gmd.de/
- KNAW (2004) 'Developing e-science in the humanities', Amsterdam: KNAW.
- Lenoir, T. (1999) 'Shaping Biomedicine as an Information Science', Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems, M. E. Bowden, T. B. Hahn and R. V. Williams. Medford, NJ, Information Today, Inc.: 27-45.

- Leydesdorff, L. (2006) 'Knowledge-based economy'.
- Lloyd, S. (2000) 'Ultimate physical limits to computation'. Nature 406: 1047-1054.
- Nentwich, M. (2003) 'Cyberscience, Research in the Age of the Internet', Austrian Academy of Sciences.
- Nijkamp, P. (2005) 'Kennisambitie en researchinfrastructuur', Innovatie Platform.
- Nowotny, H., Scott, P. en Gibbons, M. (2001) 'Re-thinking science'.
- NSF (2007) 'The future of scholarly communication: building the infrastructure for cyberscholarship'. Report of a workshop held in Phoenix, Arizona, April 17 to 19, 2007.
- NWO (2006) 'Wetenschap gewaardeerd', NWO Strategie 2007-2010. http://www.nwo.nl/files.nsf/pages/NWOA_6PXJ9W/$file/wetenschap_gewaarderd_lowres.pdf.
- Odlyzko, A. M. (1996) 'On the road to electronic publishing', Euromath Bulletin, vol. 2, no. 1, p. 49-60.
- *PRIME* ERA-Dynamics Policy Workshop (2007) 'Beyond the dichotomy of national vs. European science systems. Configurations of knowledge, institutions and policy in European research', http://www.czelo.cz/dokums_raw/ERA-Dynamics.pdf.
- Raad voor Cultuur en Raad voor het openbaar bestuur (2008) 'Informatie: grondstof met toekomstwaarde. Contouren van een visie op de rol en betekenis van informatie'.
- Rheingold, H. (2003) 'Smart Mobs: The Next Social Revolution'.
- Rip, A. en Van der Meulen, B. (1996) 'The post-modern research system'.
- Rip, A. (2000) 'Fashions, Lock-Ins, and the Heterogeneity of Knowledge Production. The Future of Knowledge Production in the Academy', M. Jacob and T. Hellström. Buckingham, Open University Press.
- Schroeder, R. and Fry, J. (2007) 'Social Science Approaches to e-Science: Framing an Agenda'.
- Shneiderman, B. (2008) 'Science 2.0', Science (AAAS).
- Sloot, P.M.A., Frenkel, D., Van der Vorst, H.A., Van Kampen, A., Bal, H.E., Klint, P., Mattheij, R.M., Van Wijka, J., Schaye, J., Van Langevelde, H.J., Bisseling, R.H., Smit, B., Valenteyn, E., Sips, H.J. Roerdink, J.B.T.M. en Langedoen, K.G. (2007) 'Computational e-science: Studying complex systems in silico. A national coordinated initiative'. White paper. Technical Report, Informatics Institute, Universiteit van Amsterdam. (http://www.science.uva.nl/research/pscs/papers/archive/Sloot20-07a.pdf).
- Van Alstyne, M. and Brynjolfsson, E. (1997) 'Widening access and narrowing focus: Could the Internet balkanize science?', Science 274 (5292): pg. 1479-80.
- Van den Besselaar, P. and Heimeriks, G. (2001) 'Disciplinary, Multidisciplinary, Interdisciplinary: Concepts and Indicators', In M. Davis and C.S. Wilson (eds), ISSI 2001, 8th international conference of the Society for Scientometrics and Informetrics. Sydney: UNSW 2001. pg. 705-716.
- Westwick, P. (2003) The National Labs: Science in an American System, 1947-1974. Cambridge, MA: Harvard University Press.

- Wouters, P. F. (1999) 'The Citation Culture', PhD thesis University of Amsterdam.
- Wouters, P. (2006) 'What is the matter with e-science?' http://www.pantaneto.co.uk/issue23/wouters.htm.
- WRR (2002) 'Van oude en nieuwe kennis: de gevolgen van ICT voor het kennisbeleid' (r61 2002).
- Zhao, Z., Belloum, A.S.Z., Wibisono, A., Terpstra, F., De Boer, P.T., Sloot, P.M.A. and Hertzberger, L.O. (2005) 'Scientific workflow management: between generality and applicability', In: Proceedings of the 5th international conference on quality software, Melbourne, Australia, Sep. 19 -20, pg. 357-364.