Infrastructuur duurzaam op maat

Een analyse van de vraag naar ICT-infrastructuur en services in Nederland

Uitgevoerd door ePLAN

Platform van eScience-Data Research Centers in Nederland

Versie 1.0 d.d. 29-02-2016

Patrick J.C. Aerts

1 Inhoudsopgave

Inhoudsopgave

2 Conclusies

Onderstaand treft u de conclusies van ePLAN over de toekomst van de nationale ICT-Infrastructuur voor de Academische publieke sector, gebaseerd op een uitgebreide enquête onder de Universitair Docenten (UD), Universitaire Hoofddocenten (UHD) en hoogleraren in Nederland en een viertal workshops langs disciplinaire lijnen.

- Algemeen
  - De uitkomsten van de enquête hebben aanzienlijke draagkracht op grond van het aantal respondenten, hun achterban en de verdeling over (sub-)disciplines;
  - De ontwikkelingen in het onderzoekdomein worden door een overzienbaar aantal kernwoorden gekarakteriseerd: vooral meer data, maar ook rekenen, speciaal ook meer echt *grootschalig* rekenen (HPC[1] en incidenteel GPU[2]-gebaseerd), meer beeldverwerking en visualisatie, meer wensen voor (ondersteuning bij) analyse van data en zorgen over security en privacy;
  - Er vindt veel multidisciplinaire samenwerking plaats, met een sterk accent op internationale samenwerking;
- Data-domein
  - De toename van het belang van en het volume aan data bij het onderzoek betreft alle disciplines en alle types data;
  - Het delen van data is geen algemeen gebruik. Slechts een kwart deelt data zonder voorbehoud;
  - Voor zover data worden gedeeld, gebeurt dat op veel en uiteenlopende manieren, waarbij clouddiensten[3] (zoals Dropbox en Googledocs en SURFdrive) het vaakst worden aangegeven;
  - Van de respondenten is het grootste deel niet bekend met wat er lokaal[4] aan data-opslag of reken-capaciteit beschikbaar is;
  - Er is een concrete verwachte enorme groei in behoefte aan data-resources (naar het Exabytes[5] domein) en –diensten;
- Reken-domein
  - Grote zorgen worden geuit over het algemeen achterblijven van het investeringsniveau van de nationale ICT-infrastructuur in relatie tot de sterk toegenomen behoefte aan alle soorten resources en in vergelijking met het aanbod elders in de wereld waarmee men in competitie is. Ook het gebrek aan reserveringen voor toekomstige investeringen, onder meer in een nieuwe nationale supercomputer is reden voor die zorg;
  - De afstand van een nationale HPC-voorziening tot de internationale top[6] zou niet groter moeten zijn dan een factor 10 voor onderzoekers om op competitief niveau te kunnen meekomen. Dit is ook nodig om (recht op) toegang te houden tot Europese voorzieningen, zoals PRACE, de Europese infrastructuur voor supercomputers (Partnership for Advanced Computing in Europe);
  - Van de meerderheid die aangeeft meer te zullen gaan rekenen geeft één derde ook aan dat dit geheugenintensief werk betreft. De verdeling over lange of juist veel korte jobs[7] is daarbij ongeveer gelijk;
  - Er is, naast een supercomputer[8], behoefte aan een robuuste tweede laag[9] van clustervoorzieningen, centraal of gecoördineerd bij de instellingen;
- Cloud-domein
  - 80% van de respondenten op de vraag over clouddiensten geeft aan van commerciële clouddiensten gebruik te maken. Dit betreft vooral data-opslag en hosting en iets minder rekenen;
  - Het belangrijkste motief voor het gebruik van clouddiensten zijn niet de kosten maar is het gebruiksgemak;
  - Over het algemeen zou de beschikbaarheid van een nationale clouddienst, of een op Europese grondslag, worden gewaardeerd, in verband met de verwachting dat privacy en security beter kunnen worden geregeld dan via commerciële diensten;

[1] HPC staat voor High Performance Computing, doorgaans geassocieerd met “supercomputing”
[2] GPU staat voor Graphical Processor Unit, in de praktijk vaak gebruikt als rekenversneller voor speciale functies
[3] Clouddiensten zijn on-demand diensten voor rekenen, data-opslag of hosting (van eigen diensten of websites). Het is daarbij voor de gebruiker doorgaans niet transparant waar ter wereld die fysieke voorzieningen staan.
[4] Onder “locaal” wordt in dit document verstaan: bij de onderzoekgroep, vakgroep, faculteit of instelling.
[5] Exa is het voorvoegsel dat staat voor 10^{18}. Een Exabyte is dus 10^{9} Gigabytes.
[6] De echte topvoorzieningen staan overwegend in de VS, maar ook in Japan of China. Zie www.TOP500.org
[7] Een “job” is de eenheid waarin werk dat aan een computer ter verwerking wordt aangeboden wordt uitgedrukt
[8] Eigenlijk een voorziening voor *capability* computing, het soort rekenwerk dat niet op traditionele clusters of mainframes kan worden uitgevoerd wegens de omvang of de communicatie-intensiteit.
[9] Onder “tweede laag”, met een piramide-model voor de computerinfrastructuur als uitgangspunt, wordt verstaan de verzameling computervoorzieningen onder de “top”. De top wordt dan gevormd door de nationale supercomputer.

- Ondersteuning, diensten en informatie
  - Er is een grote wens naar verbetering van het gebruiksgemak van de ICT-infrastructuur en de toegang daartoe. Gebruiksgemak wordt als veruit het belangrijkste argument voor het gebruik van commerciële clouds aangemerkt
  - Drie-kwart van de respondenten geeft aan niet of weinig op de hoogte te zijn van de nationale ICT-infrastructuur, maar er bestaat wel veel behoefte aan meer informatie over het bestaan ervan, de resources, de services, de kosten en de ondersteuning. Dat geldt ook voor het bestaan van de ePLAN-partijen (de data research en escience centra);
  - Ondersteuning door partijen als SURFsara en NLeSC wordt genoemd als waardevolle en noodzakelijke voorziening voor een efficiënte inzet van ICT-infrastructuur in het onderzoek;
  - De kennis over en deelname in Europese ICT-infrastructuren is beperkt tot een bescheiden groep deelnemers.
  - De Computational Science community vraagt om en zoekt naar een passend ophangpunt voor coördinatie en waarbinnen zij hun belangen en behoeftes beter kunnen kanaliseren.
- Data Stewardship en Software Sustainability
  - Ruim meer dan de helft van de respondenten geeft aan nog nooit gehoord te hebben van de termen data stewardship, research data management en/of software sustainability. Maar wel vindt 90% (!) deze onderwerpen van groot belang, groot genoeg dat hiervoor landelijk beleid op zijn plaats is (80%).

Dankwoord
De enquête, uitgezet door ePLAN, zou niet mogelijk zijn geweest zonder de bijzondere medewerking van NLeSC en DANS, zowel financieel als organisatorisch. Speciale dank past de auteur aan Niels Drost voor de proeflezing van de vragen en de ondersteuning en rapportage bij de workshops, aan Lode Kulik voor de hulp bij het gebruik van SurveyMonkey en aan Lars Ridder voor de close reading van het eindrapport.
Patrick Aerts

3 Inleiding

Nederland beschikt sinds de tweede helft van de jaren tachtig over een nationale ICT-infrastructuur (ook wel e-infrastructuur genoemd). Gedragen door een internationaal vermaard netwerk, SURFnet, is deze infrastructuur in de loop van de tijd voorzien van tal van resources, waaronder één of meer supercomputers, een groot rekencluster, dataopslagvoorzieningen op diverse locaties, een grid-infrastructuur, met clouddiensten, ondersteuning, eScience support met personele inzet, voorlichting, soms visualisatietools en meer.

We leven momenteel in een Big Data maatschappij. De Big Data wereld is bijzonder actief en velen voelen zich aangetrokken tot dit nieuwe domein, zowel in het bedrijfsleven als in het onderzoek, zoals uit de enquête zal blijken. Nieuwe instituten worden opgericht met data science als bindend element en Big Data neemt een steeds belangrijkere plaats in in industrie en het bedrijfsleven, meestal in de vorm van Business Intelligence. Ook sterk in opkomst is het concept van het *Internet of Things*: witgoed (wasmachines, koelkasten, etc.), bruingoed (radio, TV, DVD/Blue Ray spelers, ect.), installaties (CV, denk aan "Toon", electriciteitsnet, "smart grid") worden voorzien van internetconnectiviteit en wisselen data uit met eigenaren op afstand, fabrikanten voor services, etc. Daar komen bij de talloze andere apparaten in fabrieken, in het veld, bij transport en logistiek en andere sensoren die data uitwisselen over hun toestand of hun taken. Dit geeft het belang aan van een goede e-Infrastructuur ter ondersteuning van deze ontwikkelingen, met een sterke maatschappelijke component en van ondersteunende diensten op het gebied van analyse en *decision support*.

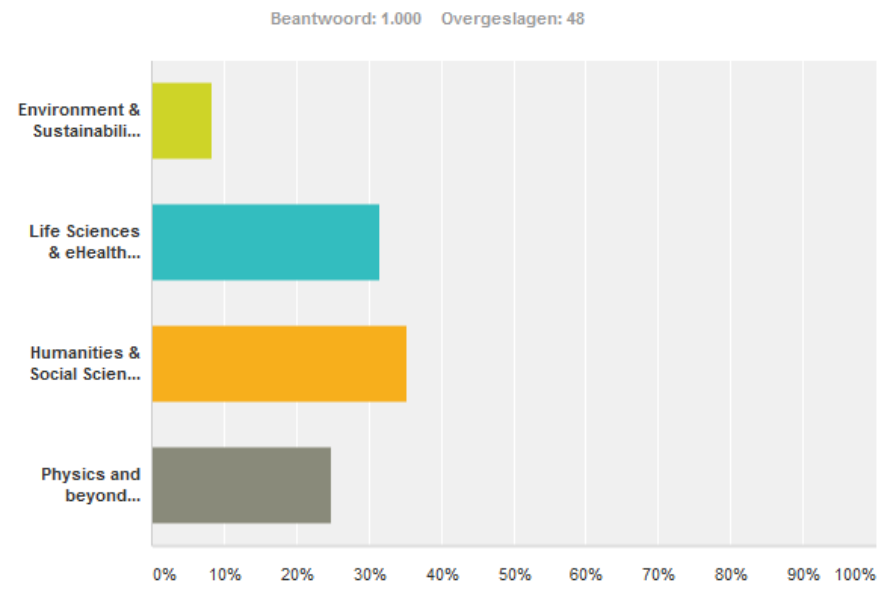
Ook onderzoek waar traditionele geschriften, boeken of handschriften nog essentieel zijn, of waar de ontwikkeling van een theorie vooral berust op denkwerk en niet op digitale data, heeft belang bij een nationale e-infrastructuur, al is het met een andere intensiteit, voor de bevordering en versnelling van de communicatie over het onderzoek of voor het delen van resultaten of de ontwikkeling van nieuwe onderzoeksmethoden. De nationale e-Infrastructuur voor het publieke onderzoek is er dan ook voor alle onderzoekers, ongeacht discipline of belangstelling.

De Nederlandse Data Science centra en onderzoekers, groepen en instituten die eScience centraal stellen in hun onderzoek, hebben hun krachten gebundeld in een (sociaal) platform, ePLAN (www.escience-platform.nl). ePLAN heeft zich een aantal taken gesteld, waaronder het in kaart brengen van *white spots* (blinde vlekken) in de nationale e-infrastructuur en het analyseren van de verwachte gevolgen voor de benodigde e-infrastructuur door de ontwikkelingen die de verschillende disciplines in hun onderzoek doormaken. De partners in ePLAN hebben door hun werkdomein tussen onderzoek en e-infrastructuur een meer dan gemiddeld goed zicht op deze ontwikkelingen, maar vele onderzoeksrichtingen staan ver af van de wereld van ICT, informatica en infrastructuur. Daarom is gekozen voor een enquête onder -in beginsel- alle disciplines en niet slechts die welke toch al van nature een nauwe betrokkenheid bij ICT en infrastructuur hebben of voelen.
Met de uitkomsten van de enquête bij de hand is vervolgens een viertal workshops georganiseerd rond het thema "de toekomstige behoefte vanuit de wetenschap aan nationale ICT-voorzieningen en diensten". Die vier workshops hadden elk één van de vier hoofdthema's die het Netherlands eScience Center (NLeSC) hanteert om het totale palet aan onderzoekdomeinen in op te delen:

- Environment&Sustainability;
- Life Sciences&eHealth;
- Humanities&Social Sciences;
- Physics&Beyond.

Deze workshops maakten het mogelijk de concept conclusies uit de enquête aan te scherpen of te verduidelijken, de mate van herkenbaarheid binnen de vier domeinen van de hoofdconclusies uit de enquête vast te stellen, omissies in de uitkomsten van de enquête (bv. door de gehanteerde vraagstelling) helder te krijgen en om de prioriteringen per domein boven te krijgen.

In hoofdstuk 12 wordt nader ingegaan op de workshops en wordt per domein een verslag gegeven over het proces en de belangrijkste bevindingen.
Het resultaat van het uitzetten van de enquête bij het hele onderzoeksveld is boven verwachting: veel respondenten, ook uit de geestes- en maatschappij- en gedragswetenschappen, alsmede onderzoekers uit de medische wereld hebben de moeite genomen de aanzienlijke lijst vragen te beantwoorden. Een spontane opmerking over de enquête als geheel, van een respondent die kennelijk maar op een beperkt aantal vragen kon antwoorden, was niettemin: “mooi dat deze enquête ook de geestes-wetenschappen aanspreekt”. Een andere opmerking was: “het is eigenlijk fantastisch dat hier nationaal over nagedacht wordt”. Wij zijn alle respondenten, zowel de genaamde als de anonieme, dan ook bijzonder erkentelijk voor de genomen moeite.



Figuur 1 Verdeling over hoofddomeinen (door respondent opgegeven)

Resteert ons als dank een nuttig verslag te presenteren waaruit duidelijk wordt wat de publieke researchsector van een nationale ICT-infrastructuur en zijn brede omgeving van diensten verwacht.

4 Representativiteit

Over de representativiteit is het volgende te zeggen. De enquête is aan 11.500 onderzoekers gestuurd, hoogleraren, UHD en UDs, met de vermelding dat de enquête gedeeld mocht worden met andere onderzoekers uit de omgeving van de aangeschreven onderzoeker.

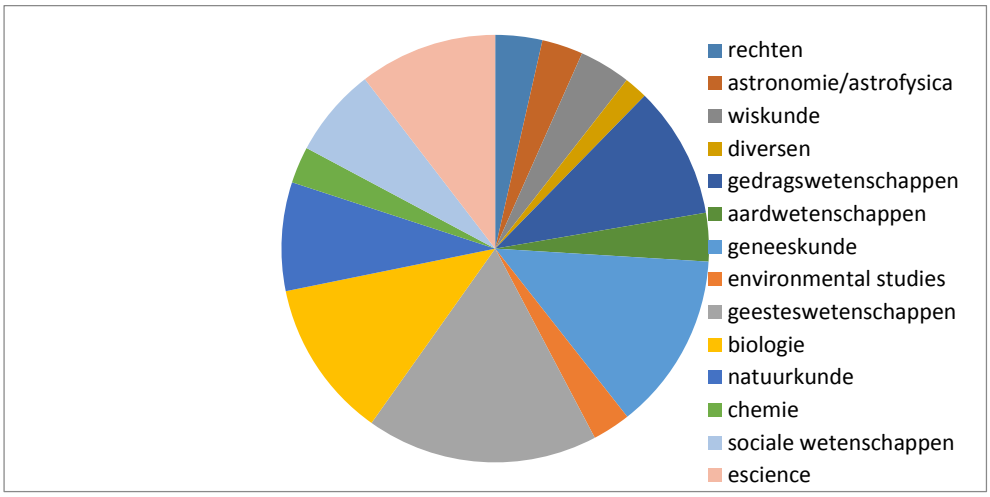
4.1 Respons

Het aantal respondenten bedraagt 1048. Dat is ruim negen procent van het aantal aangeschreven personen.
795 personen hebben opgegeven de enquête namens zichzelf in te vullen en 242 (23%) doet dat namens zijn/haar groep. De som van de opgegeven groepsgroottes bedraagt 5356 fte. In een aantal gevallen werd een hele faculteit of instituut meegeteld, wat een enigszins vertekend beeld op kan leveren. Het aantal maal dat een groepsgrootte groter dan 50 is opgegeven bedraagt 14. Als de maximale groepsgrootte wordt beperkt tot 50 -op zich een willekeurige maar toch duidelijk beperkende grens- bedraagt het virtuele aantal respondenten nog altijd ruim 3800, naast het aantal respondenten dat de enquête namens zichzelf heeft ingevuld.

4.2 Verdeling over disciplines

Belangrijk is ook de vaststelling dat de respondenten uit een breed aantal disciplines afkomstig zijn. In tegenstelling tot het overzicht bij figuur 1, is de disciplineverdeling in figuur 2 met de hand samengesteld om tot een overzichtelijk aantal te komen, op basis van vele tientallen opgegeven subdisciplines. Hierbij valt ondermeer “Informatica” onder de kop “eScience”. Op grond van al deze

punten mag worden geconcludeerd dat de verzamelde antwoorden voldoende draagkracht hebben om van betekenis te zijn als richtinggevend voor de nationale ICT-infrastructuur.



Figuur 2 Verdeling over sub-domeinen. Geaggregeerd op basis van opgaven respondenten

Passend is om hier te vermelden, dat er enkele respondenten zijn die bij diverse gelegenheden bij open vragen of opties zoals “anders, nl.”, aan hebben gegeven dat ze van de ICT-infrastructuur geen gebruik maken en ook wel dat men zulke overkoepelende voorzieningen niet ziet zitten en liever zelf extra middelen krijgt om zaken naar eigen inzicht in te richten.

5 Globale analyse van de enquête

5.1 Wat houdt de mensen bezig?

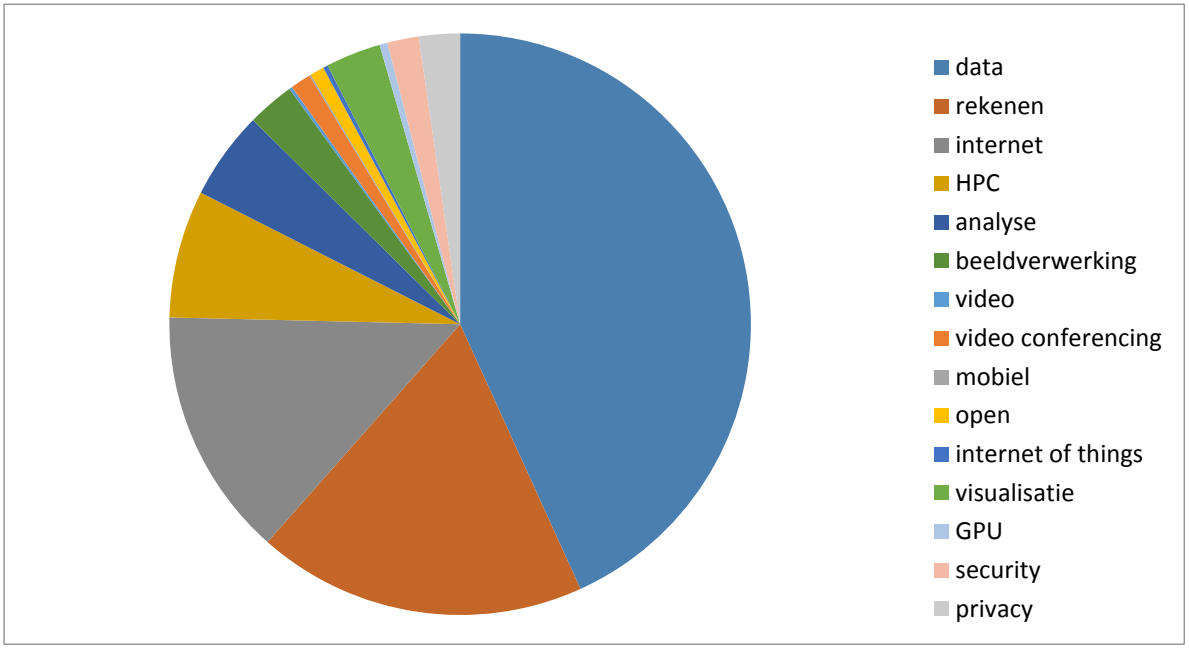
Het veruit meest genoemde onderwerp in de lijst van kernbegrippen die uit de open vraag over het onderzoek naar voren komen is “data”. Waarschijnlijk niet verrassend, gezien de tijd waarin we leven en onderzoek verrichten, maar toch goed om eens op deze wijze formeel -en met referentie naar de gehele onderzoekswereld- vastgesteld te zien. Het onderwerp is gekoppeld aan aspecten van ambitie, nieuwe drive, complexiteit in volume en inhoud en aan zorgen. Rekenen, zeker als ook de verbijzonderingen “HPC[10]” en “GPU[11]” worden meegenomen neemt een sterke tweede plaats in, gevolgd door “internet”.

De enquête omvatte zowel gesloten als open vragen, maar ook bij de gesloten vragen was veel ruimte voor toelichtingen of extra commentaar. Daar is veel gebruik van gemaakt. De eerste open vraag betrof een korte omschrijving van het vakgebied in relatie tot het huidige of een eventueel toekomstig gebruik van de nationale ICT-infrastructuur. Uit deze teksten kwam een aantal niet vooraf meegegeven kernwoorden met aanzienlijke frequentie voor. In onderstaande grafiek zijn deze kernwoorden weergegeven.

743 respondenten hebben de betreffende vraag ingevuld met antwoorden variërend van uitgebreide omschrijvingen tot één enkel trefwoord. Alle vragen zijn inhoudelijk geïnterpreteerd en samengevat in het voorkomen van kernwoorden. De kernwoorden zelf zijn gedestilleerd uit de teksten en niet vooraf geselecteerd.

[10] High Performance Computing

[11] Graphical Processing Units, tegenwoordig steeds vaker ook gebruikt als versnellers voor bepaalde rekenprocessen



De kernwoorden geven helder aan waar door de onderzoekers de ontwikkelingen gezien worden en ook welke zorgen ze hebben. Zie met name de frequentie waarmee privacy en veiligheid (security) genoemd worden als zorgwekkende hindernissen die de nuttige toepassing van data in de weg (kunnen) staan.

Er is -op basis van de geschreven tekst- een onderscheid gemaakt tussen “rekenen” en “HPC” of “GPU” in gevallen waarin duidelijk is aangegeven dat het om rekenwerk gaat dat het gebruikelijke overstijgt. Rekenen wordt zowel als *zelfstandige* behoefte aangegeven als een direct gevolg van de toename aan data die verwerkt moeten worden of ook als gevolg van de behoefte aan visualisatie en/of beeldverwerking.

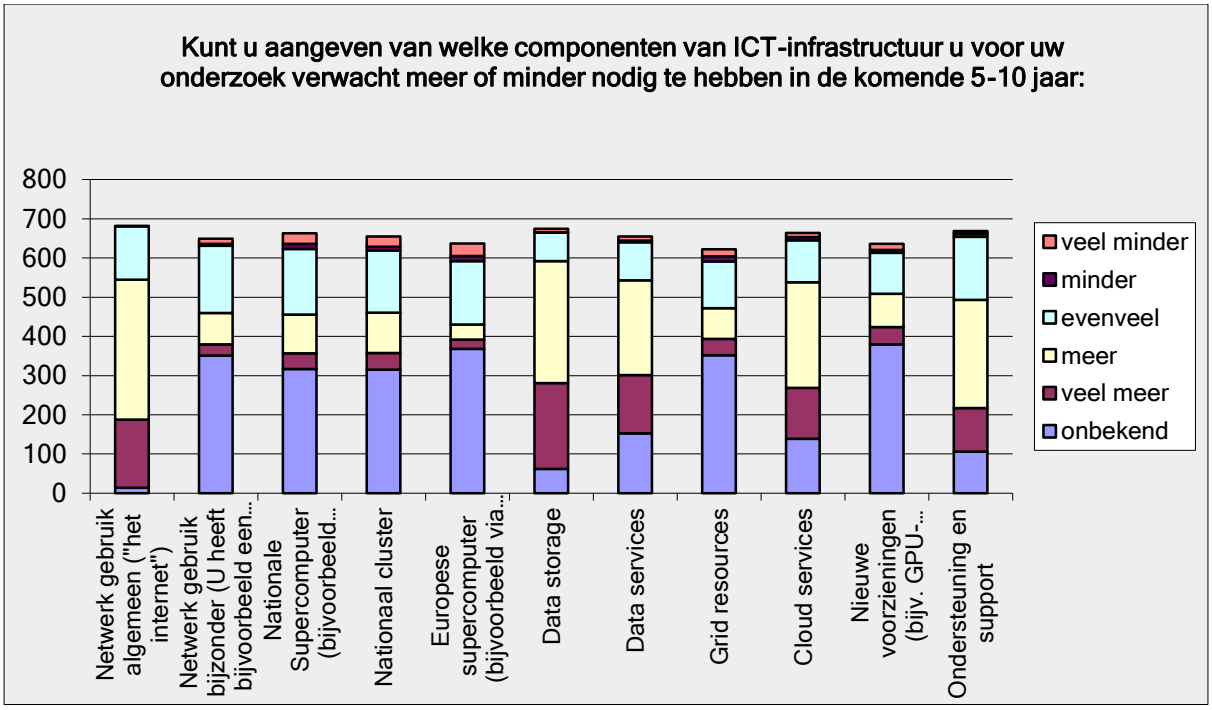
Er wordt meermaals melding gemaakt van de behoefte aan (hoge-resolutie) video conferencing. Zowel als onderdeel van een efficiëntere werkwijze als onderdeel van de behoefte tot intensievere samenwerking.

Er wordt onderscheid gemaakt tussen beeldverwerking en visualisatie. De begrippen worden elk apart of soms in samenhang benoemd. Beeldverwerking is een specifieke vorm van “analyse”, een term die ook veelvuldig voorkomt. “Meer data” vraagt ook om meer of geavanceerdere analysetechnieken, omdat het niet om de data zelf gaat maar om wat er in de data besloten ligt.

“Internet” is de term die gebruikt is om de behoefte aan goede, snelle en responsieve netwerkvoorzieningen aan te geven. Daarbij worden zowel de termen lichtpad, bandbreedte en latency incidenteel apart benoemd.

Incidenteel wordt ook gerefereerd aan de behoefte naar *open* data of de openheid van data. Zowel wat betreft overheidsdata als verzamelingen uit Facebook en Twitter. Ook de toename in het gebruik van mobieltjes en de ontwikkeling van het Internet of Things worden door onderzoekers incidenteel benoemd.

In een separate vraag kon men aangeven van welke componenten in de ICT-infrastructuur men verwachtte meer of minder nodig te zullen hebben. Dit was, in tegenstelling tot de open vraag boven, een geleide vraag met een beperkt aantal antwoordmogelijkheden, maar wel met een open “anders, nl.” optie. De uitkomsten van deze vraag versterken het beeld dat boven is geschetst.



Figuur 3 Detailvraag naar de toekomstverwachting van het gebruik van de nationale ICT-infrastructuur op basis van 698 respondenten

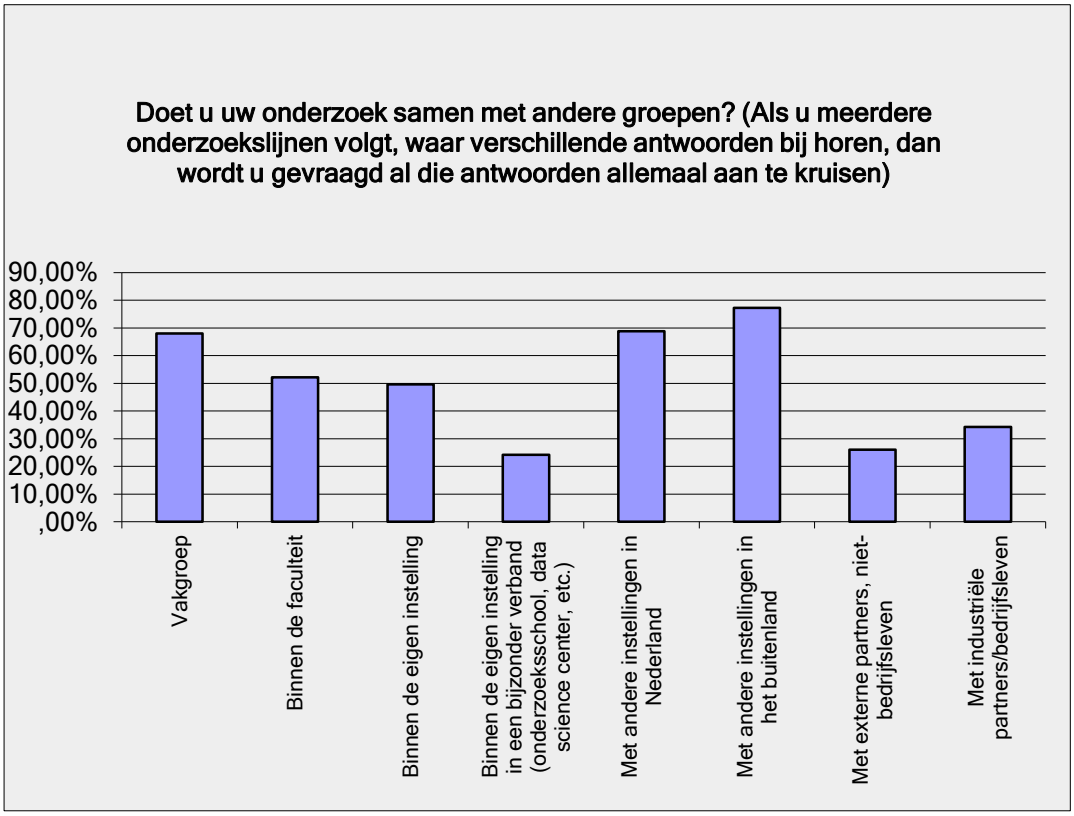
De trends zijn duidelijk: men heeft over de volle breedte van het onderzoeksveld meer tot veel meer behoefte aan netwerkcapaciteit, dataopslag en –services, clouddiensten, rekenvoorzieningen en ondersteuning. Maar ook de behoefte aan supercomputers en nieuwe voorzieningen (zoals GPU-gebaseerde systemen) neemt duidelijk toe, rekening houdend met het typisch kleinere aantal onderzoekers dat hiervan afhankelijk is (in deze telling wordt dit door 139 onderzoekers aangegeven, terwijl 167 gebruikers aangegeven dat hun gebruik op peil zal blijven. 46% van de respondenten heeft hier kennelijk een duidelijke opinie over).
33% van de 81 respondenten die de optie voor aanvullende suggesties heeft ingevuld noemt GPU's als gewenste voorziening (andere bijzondere processortypes als FPGA en Phi worden ook genoemd). Meer dan de helft van de respondenten kent geen "bijzondere netwerkdiensten", waarvan lichtpaden als voorbeeld in de vraag werden genoemd. Wel is opgemerkt, dat in meerdere gevallen de levering van een lichtpad niet het gewenste succes heeft gehad. In gevallen waarin SURFnet een lichtpad kosteloos heeft aangeboden en éénzijdig ook gerealiseerd, werd aan de gebruikerskant óf het fysieke eindpunt niet gevonden, óf lag dit binnen een datacentrum's beschermd gebied en waren de kosten voor verlenging naar de eindgebruiker's lokatie prohibitief hoog (tienduizend euro per meter is wel genoemd).
Op sommige punten zijn er, zoals verwacht mag worden, verschillen tussen de resultaten voor de vier domeinen. Zo geven alle domeinen op behoefte te hebben aan meer data opslag capaciteit, maar geven de domeinen Life Sciences&eHealth alsmede Environment&Sustainability in aanvulling daarop ook aan (46,6%) behoefte te hebben aan *veel* meer data opslag en iets minder uitgesproken ook aan data services.
Alle partijen geven duidelijk aan behoefte te hebben aan meer support. Enviroment&Sustainability is daar met 53,2% nog het meest uitgesproken in.

5.2 Samenwerking

Om inzicht te krijgen in de uitgebreidheid van samenwerkingsverbanden, waaraan mogelijk conclusies verbonden kunnen worden over het delen van resources, software en het belang van een goed functionerend netwerk is gevraagd naar alle vormen van samenwerkingen die respondenten

hebben in het kader van hun onderzoek. Als er meerdere onderzoeken plaatsvinden in meerdere samenwerkingsverbanden werd men gevraagd die allemaal te benoemen.

De vraag werd door 718 respondenten beantwoord. Daarin vallen een paar zaken op. Het is misschien vanzelfsprekend dat de samenwerking binnen de eigen vakgroep groter is dan binnen de faculteit of instelling. Maar dat de samenwerking met andere instellingen groter is dan met de eigen vakgroep is misschien minder voor de hand liggend. Mogelijk speelt hierbij een rol dat de samenwerking het intensiefst is met vakgenoten uit dezelfde (sub-)discipline. Dat zou dan ook kunnen verklaren dat de samenwerking met instellingen in het buitenland het grootst is van alle samenwerkingsverbanden die men aangeeft (namelijk ruim 77%). Een andere reden kan zijn, dat internationale samenwerking meer wordt gewaardeerd bij subsidieaanvragen. Toch karakteriseert ruim drie-kwart van de respondenten zijn of haar onderzoek als multi-disciplinair. Slechts 21% noemt zijn of haar onderzoek mono-disciplinair. Ruim één derde (34%) geeft ook aan met industriële partners samen te werken.



Figuur 4 Samenwerking met andere groepen

Antwoordkeuzen		Reacties	
mono-disciplinair		21,10%	154
multi-disciplinair		75,34%	550
Geen van beide, maar als:	Reacties	3,56%	26
Totaal			730

Figuur 5 Disciplinair karakter van het onderzoek

3,6% geeft een andere karakteristiek aan dan mono- of multi-disciplinair, meestal om nog een nuance te kunnen aanbrengen, zoals trans-disciplinair of mono-disciplinair in een breder multi-disciplinair samenwerkingsverband.

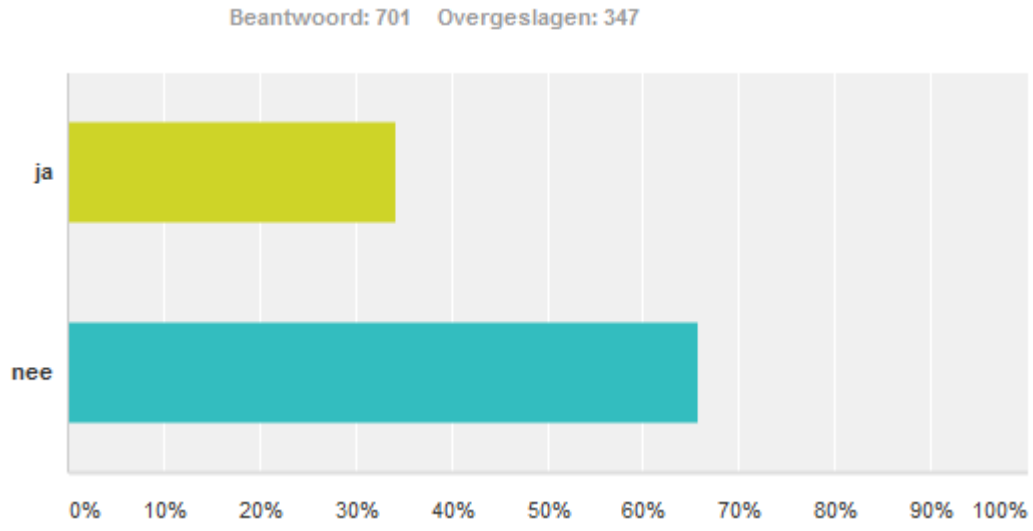
In meerdere onderzoekdomeinen treedt het fenomeen op dat samenwerking in een groot (internationaal) verband een zekere beperking oplegt aan de vrijheid om zaken in te richten naar eigen inzicht. De gezamenlijke ontwikkeling van een grote (software)applicatie met meerdere modules en over meerdere wetenschappelijke specialismen, zoals bijvoorbeeld in de klimatologie gebeurt, vraagt om taakverdeling en standaardisatie van bijvoorbeeld interfaces en data formats. Dergelijke “beperkingen” dienen dan een hoger doel. De vraag is, of dergelijke samenwerkingsverbanden vaak voorkomen. Uit onderstaand overzicht blijkt, gebaseerd op 721 respondenten, dat dat het merendeel van de onderzoekers een dergelijke coördinatie niet ervaren. Maar ruim dertig procent heeft wel degelijk te maken met formele of niet-geformaliseerde afspraken in internationaal verband. Een kleine vier procent geeft aan de vraag niet te begrijpen of een nuance te willen aanbrengen in de keuzemogelijkheden: soms heeft men van doen met zowel formele als informele afspraken, of het varieert per project.

Wordt uw hoofdlijn van onderzoek internationaal gecoördineerd via een formele, informele of project-organisatie?

Answer Options	Response Percent	Response Count
nee	62,8%	453
via een formele (project-)organisatie	14,7%	106
via een informele (project-)organisatie	16,6%	120
op grond van een internationale conventie	1,9%	14
Ja, op een andere manier, namelijk	3,9%	28
answered question		721
skipped question		327

Figuur 6 (Inter)nationale coördinatie

Tenslotte werd speciaal gevraagd naar het delen van data binnen het onderzoeksgebied. Daarop werd door ruim één derde (34%) positief geantwoord.Kennelijk is bij een meerderheid van de respondenten het delen van data niet de meest gebruikelijke procesgang. Daarover later meer onder het hoofdstuk “Data”.



Figuur 7 Delen van data

Een dergelijke vraag is ook gesteld over het samen delen van rekenresources. Maar slechts 7,7% van 680 respondenten doet binnen zijn vakgebied zoiets.

6 Data

Naast de algemene inventarisatie van toekomstige behoeftes zijn voor een aantal onderwerpen verdiepende vragen gesteld. Zo ook met betrekking tot “data”. Er is een onderscheid gemaakt tussen gestructureerde, ongestructureerde, sensor-, simulatie en gestreamde data. Men kon niet-exclusief opgeven waar men belangrijke veranderingen verwacht en waar niet. Voor mensen die met dit onderwerp niets van doen hebben was er de optie “n.v.t./onbekend”. Er hebben 767 personen antwoord op deze detailvraag gegeven.

6.1 Algemene vragen over data

Enkele respondenten geven aan over de linie minder data te verwachten.
Met *gestructureerde data* verwacht 74% van de respondenten meer tot veel meer te maken te krijgen. 53% verwacht meer tot veel meer *data uit simulaties* te gaan verwerken en zelfs als het gaat om *gestreamde data* verwacht 42% een toename in het onderzoek. Dus aanzienlijk meer mensen in het onderzoek verwachten met meer *gestructureerde* data te maken te krijgen dan met meer *ongestructureerde* data, ondanks de geweldige toename van het aantal sociale-communicatiewebsites en -systemen dat zich buiten de wetenschap ontwikkelt en dat momenteel de grote aandacht van de Business Intelligence wereld heeft. Maar ook daar wordt altijd nog een toename tot sterke toename voorzien door 41% van de respondenten.

	minder	even veel	meer	veel meer	onbekend/n.v.t.	Total Respondents
Gestructureerde data, zoals databases en geordende bestanden	**1.97%** 15	**20.39%** 155	**38.82%** 295	**34.74%** 264	**6.32%** 48	760
Ongestructureerde data, zoals emails, Facebook of Twitter data	**8.85%** 67	**40.29%** 305	**27.48%** 208	**13.08%** 99	**10.83%** 82	757
Sensordata of data uit instrumentatie	**4.64%** 35	**18.15%** 137	**24.24%** 183	**24.50%** 185	**28.87%** 218	755
Data uit simulaties	**4.39%** 33	**22.87%** 172	**31.78%** 239	**21.01%** 158	**20.88%** 157	752
Gestreamde data	**3.79%** 28	**23.82%** 176	**28.55%** 211	**13.26%** 98	**31.12%** 230	739

Er zijn wel duidelijke verschillen tussen de antwoorden per hoofddiscipline. Zo geven de disciplines Life Sciences&eHealth alsmede Environment&Sustainability aanzienlijk hogere verwachtingen voor meer en veel meer gestructureede data (tot 88,7%) dan bij Humanities&Social Sciences of Physics&Beyond. Het veld Environment&Sustainability verwacht bovendien veel meer (48,5%) tot meer (25-30%) sensordata of data uit simulaties; duidelijk hogere percentages dan bij de overige domeinen.

6.2 Data delen en openheid

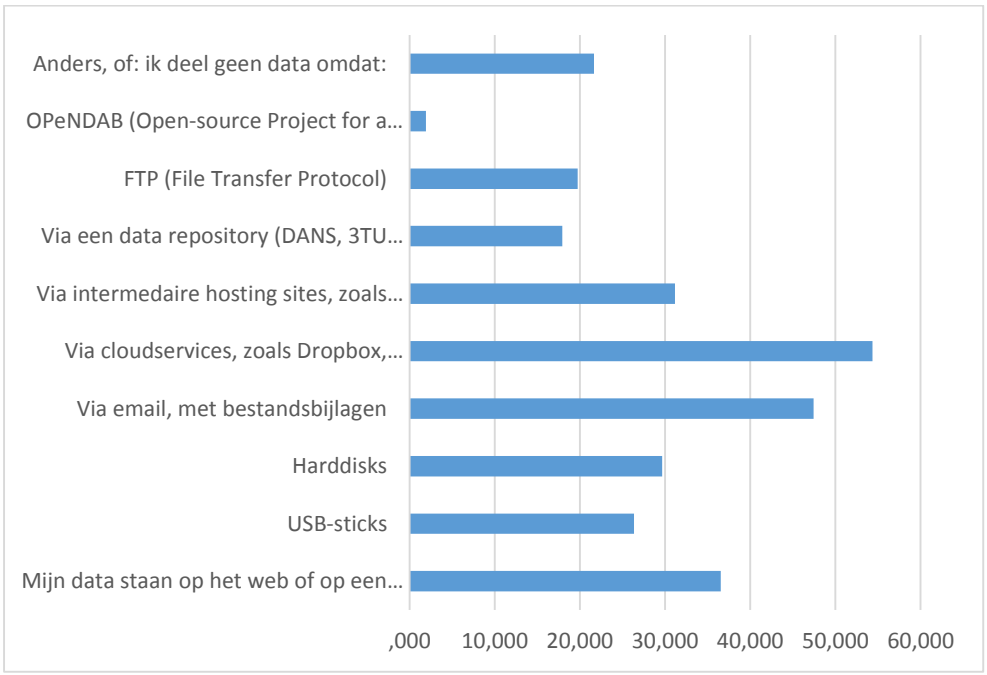
Tegenover het tamelijk intensieve beleid dat door overheden, nationaal en internationaal wordt gevoerd met betrekking tot *openheid en delen van data* zijn de onderstaande antwoorden, gebaseerd op 728 respondenten, opmerkelijk: slechts een kwart deelt zijn of haar data zonder veel voorbehoud met iedereen. 36% deelt zelfs nooit data. Daar worden wel verschillende redenen voor aangegeven, waarvan de meeste met privacy te maken hebben. Maar ook gebrek aan wil om de moeite te nemen de data publiek te maken of angst voor misinterpretatie van de data door anderen komen voor.

Antwoordkeuzen	Reacties	
Ja, voor iedereen open	25,14%	183
Alleen voor collega onderzoekers	38,74%	282
Nee, mijn data zijn te privacygevoelig	12,09%	88
Nee, mijn data zijn niet openbaar, omdat:	24,04%	175
Totaal		**728**

Van de argumenten die gegeven worden voor niet- of pas latere openbaarmaking komen de volgende drie het vaakste voor, uit 175 gegeven extra toelichtingen: 1) privacy of beperkingen door opdrachtgever/andere partij (33 maal), 2) openbaarmaking pas na publicatie van een artikel of het einde van een project (30 maal), 3) geen passende mogelijkheden voor publicatie (14 maal). Bij dit laatste argument kan het zijn dat het om te veel data gaat, de aard van de data (handgeschreven en geannoteerd) of geen middelen (geld of personeel).

6.3 Wijze van delen van data

Als data worden gedeeld, gebeurt dat op elke denkbare manier. De vraag is beantwoord door 725 respondenten en meerdere opties waren mogelijk. Het valt op dat een groot aantal respondenten gebruik maakt van Dropbox of daarmee vergelijkbare oplossingen. Het aantal respondenten dat ook gebruik maakt van wat specialistischer mogelijkheden, zoals FTP is nog aanzienlijk.
Veel data worden ook gedeeld door persoonlijke contacten, zoals bijvoorbeeld via email. De antwoorden bij “anders, nl.” leveren niet veel extra informatie op, maar de hoge frequentie van mensen die via emailbijlagen hun data uitwisselen geeft het belang aan van het delen op basis van onderling vertrouwen, alsmede van het profijt dat mensen willen kunnen hebben van de soms jarenlange inspanningen om data te verzamelen. Daarnaast worden vooral systemen genoemd die alleen bekend en toegankelijk zijn voor de onderzoekers die bij projecten betrokken zijn. Die zijn soms openbaar, maar alleen als men de weg kent.



Figuur 8 Manieren voor het delen van data. Uitgedrukt in percentages van het aantal respondenten van deze vraag (725). Meerdere antwoorden waren mogelijk.

6.4 Lokale datafaciliteiten

Op de vraag of men beschikt over lokale datafaciliteiten, antwoordt een duidelijke meederheid bevestigend. Onder lokale faciliteiten werd expliciet niet verstaan een dropbox-achtige oplossing,

Antwoordkeuzen	Reacties	
ja	60,47%	387
nee	30,94%	198
onbekend	8,59%	55
Totaal		640

Figuur 9 Beschikt men over lokale data-faciliteiten

maar een fysieke faciliteit.
Van degenen die hebben opgegeven wel over lokale opslagmedia te beschikken, wordt de volgende opgave verstrekt van de capaciteiten waarover het dan gaat. Uitgedrukt in Terabytes ligt het accent op enkele tientallen. Ongeveer 10% beschikt over opslag op Petabyte-niveau.

	Geen/onbekend	<10	10-100	100-1000	>1000	Totaal
Op de afdeling	44,51% 158	27,04% 96	20,00% 71	6,20% 22	2,25% 8	355
Op de faculteit	64,06% 205	16,88% 54	9,38% 30	6,56% 21	3,13% 10	320
Op de instelling	63,45% 217	14,62% 50	6,43% 22	5,85% 20	9,65% 33	342

Figuur 10 Opgegeven omvang van lokale data-faciliteiten

6.5 Data Stewardship en Software Sustainability

De vraag of men bekend is of zelfs betrokken bij de ontwikkelingen rond Data Stewardship en Software Sustainability levert het interessante inzicht op dat een ruime meerderheid van de respondenten nog nooit van de termen Data Stewardship, Research Data Management of Software Sustainability heeft gehoord. Maar dat men dit belangrijke onderwerpen vindt (ruim 90%), belangrijk genoeg zelfs om aan te geven dat hier landelijk beleid op zijn plaats is, blijkt uit de overweldigende positieve respons: bijna 80% van de respondenten ondersteunt dit.

	locaal (bij de instelling)	nationaal	internationaal	niet mee bekend	Totaal
Data Stewardship	28,10% 163	8,97% 52	11,55% 67	51,38% 298	580
Research Data Management	33,73% 196	10,84% 63	13,08% 76	42,34% 246	581
Software Sustainability	18,74% 107	5,78% 33	11,03% 63	64,45% 368	571

Figuur 11 Is men bekend of zelfs betrokken bij de onderwerpen Data Stewardship en Software Sustainability

Antwoordkeuzen	Reacties	
ja	90,24%	536
nee	2,36%	14
niet relevant voor mij	7,41%	44
Totaal		594

Figuur 12 Vindt u het belangrijk dat er tijdens en na afloop van projecten aandacht wordt besteed aan de omgang met data en het behoud van belangrijke software?

Antwoordkeuzen	Reacties	
ja	77,84%	439
nee	10,99%	62
geen mening	11,17%	63
Totaal		564

Figuur 13 Bent u van mening dat de nette omgang met data en het behoud van relevante data belangrijk genoeg zijn om nationaal of internationaal beleid voor te ontwikkelen?

De parallelle vragen over software sustainability worden op dezelfde wijze beantwoord, met dien verstande, dat het aantal respondenten dat daarop geen mening heeft veel groter is.

Het is opvallend, dat de antwoorden op deze vragen qua verdeling vrijwel gelijk zijn voor de vier hoofddomeinen.

De open vragen met opmerkingen over data stewardship en software sustainability leveren een beeld op waaruit blijkt dat veel deelnemers van mening zijn dat er op die vlakken te weinig gebeurt, dat men er langzaamaan aan begint, maar vooral dat men de huidige processen te amateuristisch vindt. Men is op zoek naar veilige manieren om professioneel met data en software om te gaan en mist in de eigen omgeving een goede aanpak.

62% van de 562 respondenten geeft aan bereid (28,5%) of misschien bereid (33,6%) te zijn mee te werken aan de beleidsvorming op de gebieden data stewardship en software sustainability.

7 Rekenen

7.1 Algemeen

Van 764 respondenten geeft 56% aan meer tot veel meer te zullen gaan rekenen. En ruim één derde van deze respondenten geeft aan behoefte te hebben aan machines voor geheugenintensief rekenwerk. Dat past voor wat de supercomputergebruikers betreft bij het profiel van de nationale supercomputer als machine met relatief veel direct adresseerbaar (werk-)geheugen. Maar kennelijk hebben ook overige gebruikers behoefte aan machines met extra veel geheugen.

Gebruikers hebben ook behoefte aan machines waarop lange runs gedaan kunnen worden, zowel op de supercomputer (14%) als op een cluster (32%). Maar het meeste rekenwerk bestaat uit korte runs, maar dan heel veel (47%). (Dat is evenveel als het aantal keer dat lange runs worden opgegeven voor supercomputer en cluster bij elkaar, maar dat zijn niet per se verschillende gebruikers).

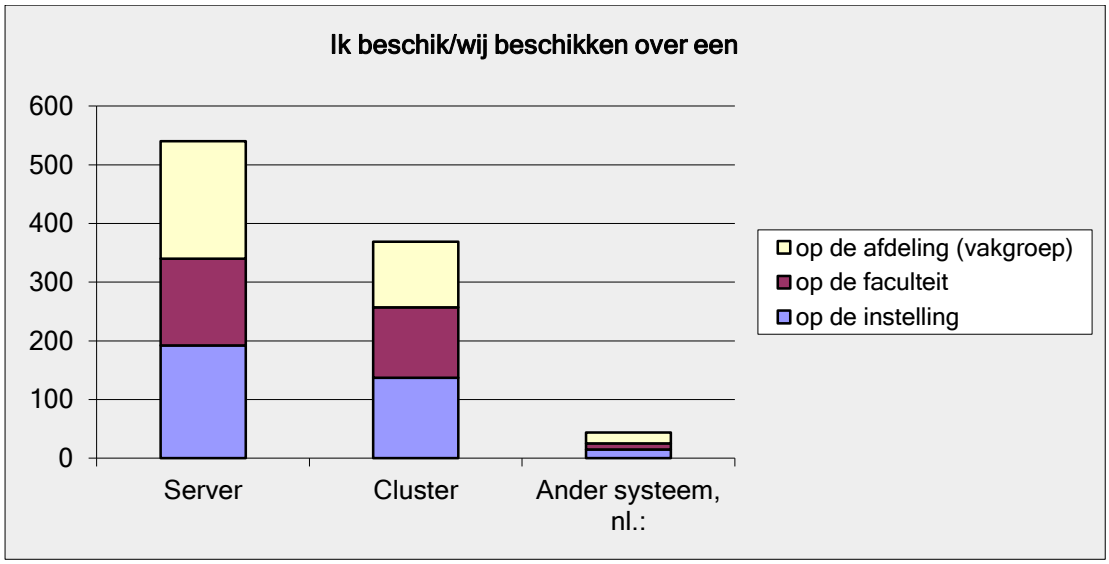
25% van de 764 respondenten op deze vraag geeft aan niet of nauwelijks te rekenen, maar 9,5% heeft juist tijdkritisch rekenwerk. 20% heeft rekenwerk met veel IO (input-output). Dat is dus data-intensief rekenwerk.

Antwoordkeuzen	Reacties	
Geen, ik reken nauwelijks	24,69%	182
Kort rekenwerk, maar wel vaak	46,54%	343
Lang rekenwerk, op een cluster	32,70%	241
Lang rekenwerk, op een supercomputer	14,25%	105
Doorgaans gedistribueerd rekenwerk	10,31%	76
Tijdkritisch rekenwerk	9,36%	69
Rekenwerk met veel I/O	20,22%	149
Geheugenintensief rekenwerk	35,01%	258
Totale aantal respondenten: 737		

Figuur 14 Wat is de aard van uw rekenwerk?

7.2 Lokale rekenvoorzieningen

Onder lokale rekenvoorzieningen worden een gedeelde server of rekencluster verstaan. 687 respondenten hebben de vraag beantwoord over lokale rekenvoorzieningen. Daarvan geeft 66,8% aan daarover te beschikken. 14,7% geeft expliciet aan daarover niet te beschikken en 18,5% weet het niet.
Verder gevraagd naar de aard en omvang van deze voorzieningen komt het volgende beeld naar voren:



Figuur 15 Beheersniveau lokale voorzieningen

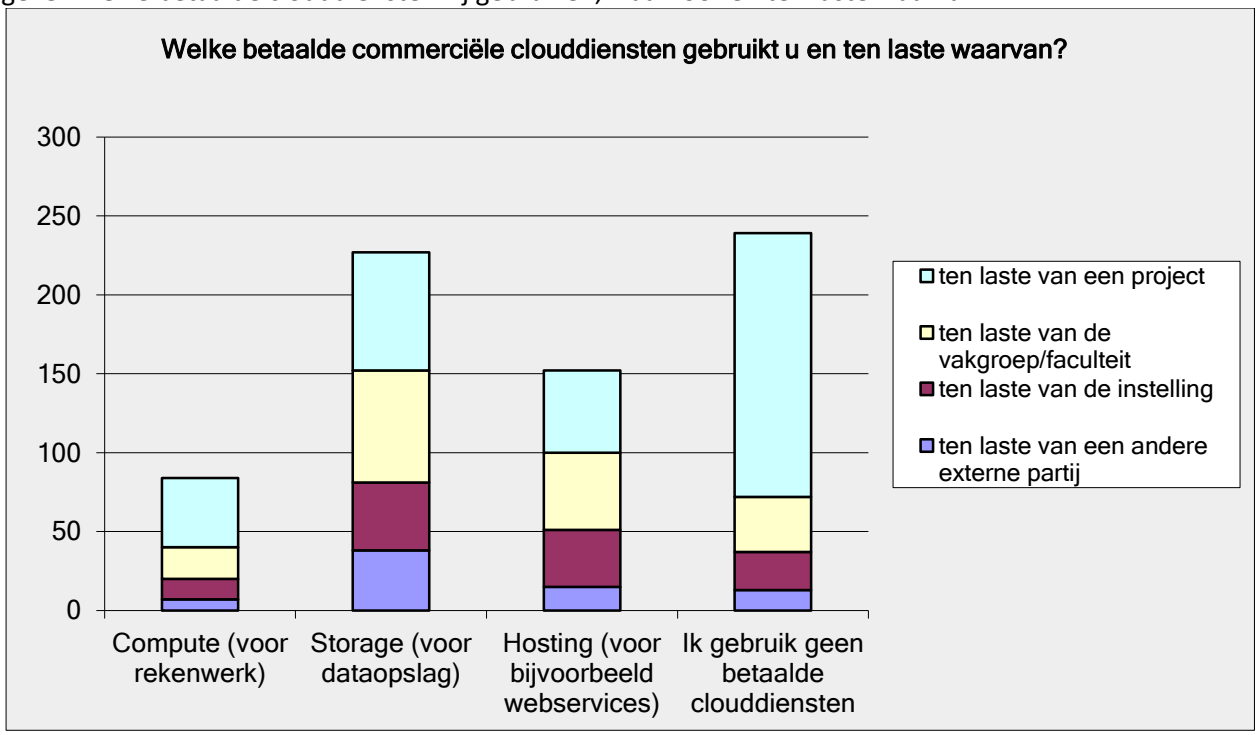
De systemen behoren eigenlijk allen tot de bekende merken (HP, Clustervision, Dell met Intel, AMD) en de eigengebouwde systemen en Beowulfclusters zijn opgebouwd uit de bekende merken processoren. Het aantal personen dat specifieke informatie heeft gegeven over de lokale systemen bedraagt zo'n 73. Het aantal opgegeven "cores" bedraagt in totaal 98354, waarbij aangetekend wordt dat één respondent het (geverifieerde) aantal van 50.000 heeft opgegeven.
Van 680 respondenten geeft 7,7% aan dat er in het vakgebied afspraken bestaan over het delen van rekentijd. Aangenomen mag worden (maar niet geverifieerd) dat dit vooral grid-gebruik betreft.

Gevraagd werd ook naar andere ICT-voorzieningen waarover men kon beschikken of waartoe men toegang heeft. Daarop heeft 28,5% van een minderheid van 390 van de respondenten een inhoudelijk antwoord gegeven. Afgezien van antwoorden in de sfeer van "cloud" of "toegang tot

HPC" betroffen de meeste antwoorden de toegang tot externe voorzieningen van derden, zoals van ECMWF, CBS, CERN, toegang tot supercomputers in de VS of Rusland, TRAIT, de NCDR-infrastructuur of ook voorzieningen thuis. Een enkeling noemt een eigen GPU-cluster, een FPGA-platform of toegang tot externe databases.

8 Clouddiensten

In termen van ICT-infrastructuren vormen clouddiensten de nieuwste ster aan het firmament. Er zijn betaalde en onbetaalde diensten. En er zijn diensten voor rekenen, data-opslag of combinaties daarvan. Het gebruik van het woord "ster" bij de introductie hierboven is niet toevallig: van het aanzienlijke aantal van 619 respondenten die de vraag over clouddiensten hebben beantwoord geeft maar liefst 80% aan van commerciële clouddiensten gebruik te maken! Ook opmerkelijk is dat minder dan één procent van deze respondenten aangeeft nog nooit van clouddiensten te hebben gehoord. Clouddiensten hebben dus een opmerkelijk groot bereik. Het is dan misschien weer merkwaardig dat slechts 18 respondenten aangeeft gratis clouddiensten te gebruiken (waarvan 83% Dropbox gebruikt, 20% SURFdrive en een handvol gebruikt GoogleDocs), terwijl 458 respondenten aangeven welke betaalde clouddiensten zij gebruiken, waarvoor en ten laste waarvan.



Figuur 16 Opgave gebruik commerciële clouds

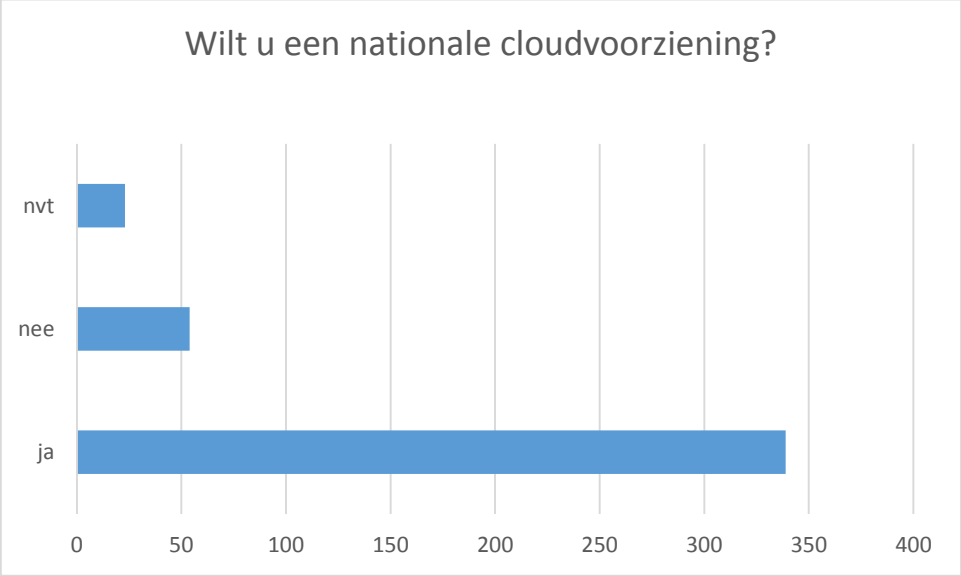
Het gebruik van (commerciële) clouddiensten heeft een duidelijke positie ingenomen: jaarlijks wordt door de respondenten op deze enquête mogelijk tot € 5 miljoen[12] aan commerciële clouddiensten besteed. Voor de gehele onderzoekgemeenschap wordt dan een aanzienlijk bedrag uitgegeven aan diensten die niet via de huidige nationale ICT-infrastructuur worden betrokken. Dit valt aan te merken als een lacune in het voorzieningenniveau.

Clouddiensten worden het meest gebruikt voor dataservices (49%), gevolgd door hosting van diensten (zoals webservices), gevolgd door rekendiensten (18%). Het meeste daarvan komt ten laste van een project (37%), gevolgd door de vakgroep (30%), de instelling (20%) en diverse andere partijen (13%).

[12] Er zijn respondenten die hun opgave bij de enquête later naar beneden hebben aangepast.

Waarom gebruiken mensen commerciële clouddiensten, tegenover andere alternatieven? De *prijs* of "zelfstandig willen zijn" is voor ruim 9% van 362 respondenten het belangrijkst. Maar voor 81% is dat het gebruiksgemak. Dat sluit aan bij open tekst opmerkingen over onvoldoende gebruiksgemak bij bestaande ICT-infrastructuurvoorzieningen, waaronder SURFdrive.

Zou u voorstander zijn van een nationale cloudvoorziening? Het antwoord hierop is gegeven door 416 respondenten die daarop met ruim 81% positief antwoorden. Bij dat positieve antwoord zijn wel een aantal voorwaarden opgegeven: het moet goed werken (minstens zo goed als "Dropbox"), het moet goedkoop zijn, het moet vooral ook veilig zijn, men moet er internationaal data mee kunnen delen en ook commerciële partijen waarmee vanuit de wetenschap wordt samengewerkt moeten voor het doel van die samenwerking ook toegang kunnen krijgen, om maar een paar voorbehouden te noemen. Deze eigenschappen worden momenteel aan SURFdrive niet toegedicht, hoewel die voorziening qua gevoel van veiligheid bij onderzoekers de voorkeur heeft boven Dropbox.
Degenen die "nee" antwoorden vinden dat er al prima commerciële voorzieningen bestaan, zijn bang dat men de concurrentie niet aan zal kunnen, of heeft gewoon geen behoefte aan "weer een dienst". Onder "nvt" vallen ook degenen die het niet weten of er geen oordeel over hebben. Niettemin is de uitkomst van deze vraag wel scherp: met alle mitsen en maren is er onder degenen die zich hebben uitgesproken een duidelijke vraag naar een nationale cloudvoorziening.

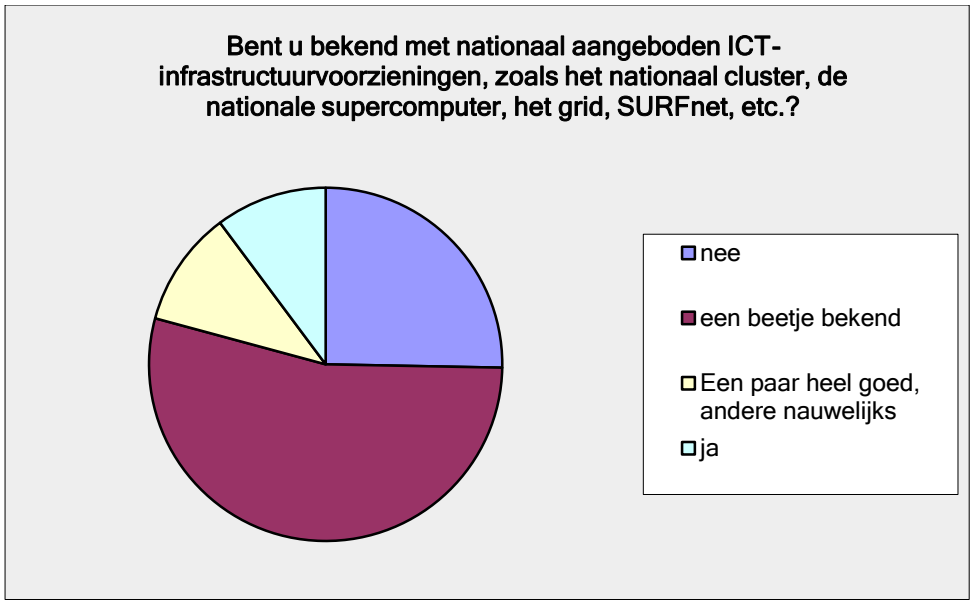


Figuur 17 Zou u voorstander zijn van een nationale cloudvoorziening?

9 De Nationale ICT-infrastructuur

De nationale ICT-infrastructuur is ontstaan in de tweede helft van de jaren tachtig. Deze bestond in eerste instantie uit twee hoofdelementen: een netwerk (SURFnet) en een supercomputer. Die infrastructuur is mettertijd uitgebreid en diverser geworden, met specialistische datacommunicatievoorzieningen van SURFnet, toegang tot andere supercomputers dan de "nationale", een groot rekencluster, dataopslagvoorzieningen, een grid-infrastructuur, cloud-diensten en meer. Met deze verbreding van het aanbod werd en wordt beoogd een grotere en meer diverse groep onderzoekers te kunnen bedienen. Maar mogelijk kan hier nog een slag gemaakt worden. In ieder geval wat betreft de bekendheid met het bestaan van deze nationale ICT-infrastructuur. In onderstaand overzicht, gebaseerd op 644 respondenten, is het aantal onderzoekers dat zich goed op de hoogte acht van de nationale ICT-infrastructuur duidelijk in de minderheid. Er van uit gaande dat alle respondenten gebruik maken van de toegang tot het internet,

die buiten de eigen instelling in Nederland door SURFnet geleverd wordt, mag misschien worden aangenomen dat “een beetje bekend” slaat op dat gebruik. Dat meer dan een kwart geheel niet bekend is met het fenomeen is hopelijk een beetje veranderd door deze enquête, maar lijkt toch wel opmerkelijk.



Figuur 18 Gevraagd naar de bekendheid van de nationale ICT-infrastructuur

Een deel van de respondenten (113) heeft de vragen beantwoord over de toegang tot de ICT-infrastructuur en het gebruiksgemak. Daarvan hebben er 14 ook een toelichting gegeven op hun antwoorden. Uit het feit dat de respons op deze vraag beperkt is moet wellicht worden geconcludeerd dat de strekking van deze vraag de meeste respondenten al onvoldoende helder is. Diegenen die met de infrastructuur bekend zijn of althans enige componenten daarvan goed kennen, alsmede een deel van diegenen die de infrastructuur een beetje zeggen te kennen hebben blijkbaar op deze vraag een antwoord gegeven. Een ruime meerderheid daarvan (72%) zegt de weg te kennen voor toegang tot de verschillende faciliteiten in de infrastructuur en 23% van deze respondenten geeft ook aan makkelijk te kunnen “schakelen” tussen de verschillende faciliteiten. Met schakelen werd bedoeld de onderlinge uitwisseling van data tussen de resources of het laten samenwerken van verschillende resources in één verband. Maar een bijna even groot deel geeft aan niet te weten hoe er tussen de faciliteiten geschakeld kan worden. Er is ook een groot aantal respondenten die hier geen expliciet antwoord op heeft.

Vindt u de toegang tot nationale ICT-faciliteiten voldoende transparant (meerdere antwoorden mogelijk)?

Antwoordopties	Respons Percentage	Respons aantallen
Ik weet hoe ik toegang kan krijgen en toegang moet aanvragen	72,2%	96
Ik kan voldoende makkelijk schakelen tussen verschillende faciliteiten	23,3%	31
Ik weet niet hoe ik toegang kan krijgen of moet aanvragen	12,8%	17
Ik weet niet hoe ik makkelijk kan schakelen tussen verschillende faciliteiten	20,3%	27

Figuur 19 Transparantie toegang en gebruik

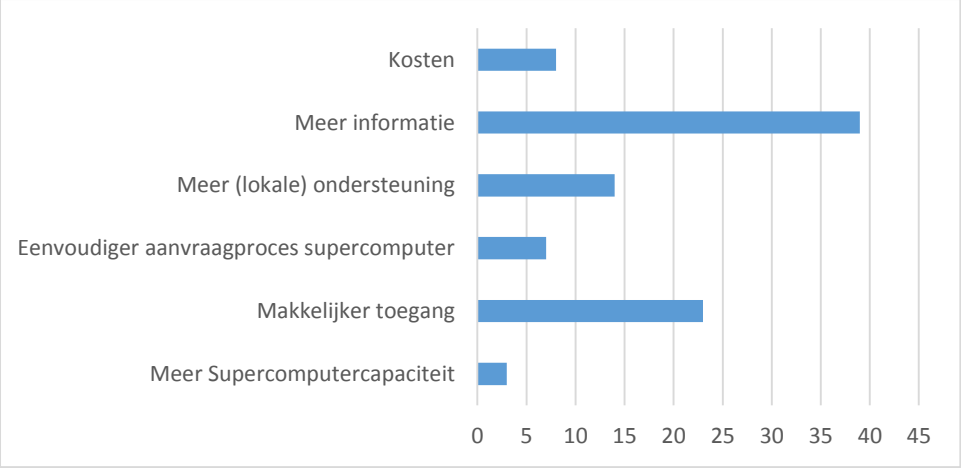
De antwoorden ”ik weet niet hoe ik toegang kan krijgen” of “ik weet niet hoe ik kan schakelen tussen de verschillende voorzieningen” werden met ongeveer 25% betrekkelijk consistent over de verschillende disciplines heen beantwoord. Alleen de groep Physics&Beyond leek hier duidelijk geen probleem mee te hebben.

243 respondenten hebben van de gelegenheid gebruik gemaakt om aan te geven wat er volgens hen wenselijk is in aanvulling op de bestaande infrastructuur (of, indien zij eerder aangaven niet op de hoogte te zijn van de nationale ICT-infrastructuur: waaraan zij in ieder geval behoefte hebben).

Uit deze diverse verzameling antwoorden en suggesties worden de onderstaande meermaals genoemd of vallen op:

- Een git-server;
- Een (nationale) voorziening die net zo goed werkt als Dropbox;
- Een nationale cloud;
- Een grote elektronische bibliotheek;
- Grootschalige dataopslag tegen lage kosten;
- In de toekomst (5-10 jaar) hebben we meer rekenkracht en dataopslag nodig;
- Behoefte aan lokale voorzieningen met snel internet;
- Behoefte aan snel datatransport (bijvoorbeeld voor meer dan 5TB);
- (Meer) Supercomputer/Supercluster;
- Snellere toegang tot supercomputer voor meer jaren tegelijk (geen jaarlijkse aanvraag);
- Meer gestructureerde informatie over de nationale ICT-infrastructuur;
- Meer ondersteuning;
- Geen of niet meer behoefte dan nu.

Desgevraagd kon men nog een laatste open opmerking maken over de nationale ICT-infrastructuur. Van deze gelegenheid is door 223 personen gebruik gemaakt. 44 antwoorden vallen in de categorie “divers”, d.w.z. niet rubriceerbaar. 90 personen meldden geen commentaar te hebben. Van de overige is een categorie-indeling gemaakt die er als volgt uit ziet:



Figuur 20 Gerubriceerde antwoorden op de open vraag over de ICT-infrastructuur

Twee dingen vallen hierin in het bijzonder op. Op de eerste plaats vraagt een groot aantal respondenten expliciet om meer informatie. Meer nog over de ICT-infrastructuur als zodanig dan over specifieke aspecten. Maar ook de roep om meer ondersteuning en het commentaar op het gebruiksgemak, zowel technisch als administratief, zijn duidelijk.

Ten aanzien van de beschikbaarheid en toegankelijkheid van nationale supercomputercapaciteit worden opnieuw ook bij de open vragen opmerkingen gemaakt, met name ook onder verwijzing naar de situatie in het buitenland. Men vindt de toegang te complex, technisch of administratief.

10 Internationaal

Sinds omstreeks 2003 is gewerkt aan een Europese ICT-infrastructuur. Er waren voordien ook al Europese projecten op het gebied van ICT, en natuurlijk was er al een Europese netwerkvoorziening

gedragen door de verzameling van nationale research and education network organisations (NRENs), maar de komst van grids bracht de meer geavanceerde ICT-infrastructuurontwikkeling in een stroomversnelling. De aanleg van de zogenaamde grid-infrastructuren, door via netwerken computerresources met elkaar te delen, toonde dat het steeds makkelijker werd gezamenlijk in Europa gebruik te maken van nationaal beschikbare voorzieningen. Daarop werd voortgebouwd met de komst van DEISA, Distributed European Infrastructuur for Supercomputer Applications, een grid van bestaande supercomputers, en PRACE, het Partnership for Advanced Computing in Europe voor nieuwe supercomputers van een orde groter. Tenslotte is daar EUDAT bijgekomen als generieke infrastructuur voor data. En ook maken de ESFRI-infrastructuren in toenemende mate gebruik van de mogelijkheden om bestaande of eigen infrastructuren op te bouwen voor gezamenlijk gebruik.

De vraag is wat de Nederlandse onderzoekgemeenschap weet over het bestaan van deze Europese voorzieningen en of ze daar gebruik van maken. Het is misschien niet verrassend dat de meerderheid weinig met dergelijke infrastructuren te maken heeft. Maar velen (83% van 626 respondenten) geven aan deze infrastructuren überhaupt niet te kennen.

Antwoordkeuzen	Reacties	
ja	4,31%	27
een beetje	12,62%	79
nee	83,07%	520
Totaal		626

Figuur 21 Bekendheid met Europese ICT-infrastructuren

Een aanzienlijk aantal respondenten heeft deze vraag beantwoord. Men mag waarschijnlijk wel aannemen dat de overige respondenten eerder bij de categorie: “niet bekend” mogen worden gerekend dan bij enige andere categorie. In ieder geval heeft ruim 83% van de respondenten hier nog nooit van gehoord. 17% weet er iets van en een veel kleiner aantal doet er ook iets mee. Omstreeks 2% *gebruikt* één of meer van de Europese ICT-voorzieningen (ICT-infrastructuren of ESFRI-faciliteiten).
Onder de categorie gebruik van “andere Europese voorzieningen”, antwoorden weliswaar 104 respondenten, maar de meeste met “geen” of “nvt”. Genoemd worden ISBE, ELIXER, BBMRI, ECMWF, CLARIN, CLARIAH, ENVRI, EBI, EuroDISH, ESO data-archief, ESA data-archief, EUROfusion,, EATRIS, EGA, ENA, Grid5000, CERN, ESA, Bibliotheken en krantendatabanken.

De vraag over wensen voor de Europese ICT-infrastructuur levert weinig *nieuwe* inzichten op: de meesten hebben geen expliciete wensen, en voor zover er wensen meer dan eens benoemd worden betreft dat opnieuw datavoorzieningen, met allerlei nuances, en supercomputers. Dat zijn natuurlijk ook wel de meest toonaangevende ICT-infrastructuren die Europa kan aan bieden (waarbij we Geánt voor het gemak maar als vanzelfsprekend beschouwen).

Het gebruik van ICT-gerelateerde voorzieningen van *buiten* Europa is beperkt, maar wel divers. Het betreft dan vooral vakgebonden databases, clusters en supercomputers. De laatste twee vaker via oudere contacten of werkverbanden dan via formele routes, zoals INCITE-calls (VS).

11 Kennis over eScience/Data Research Centra

Sinds omstreeks medio 2011, toen aan het Netherlands eScience Center (NLeSC) vorm werd gegeven, zijn er bij verschillende instellingen Data Research of Data Science instituten of centra opgezet. Bij de instellingen worden daarbij (sub-)disciplines/vakgroepen in een groter verband

samengebracht, om inhoud te geven aan met name (maar niet alleen) data gedreven of data-gedomineerd onderzoek. De vraag is hoe het staat met de bekendheid van deze centra, die de belichaming van een nieuwe trend zijn.

Van ruim 600 respondenten kent 80% noch NLeSC, noch een van de andere data science centra. Dat staat in contrast met de ruim 600 deelnemers aan het nationale escience symposium van 2015 in de Arena. Deze opkomst is, zeker in historisch perspectief, ongekend groot. Maar kennelijk is er een nog veel grotere groep die wat de communicatie betreft te ver af zit van de actieve kern van onderzoekers in het eScience en data researchdomein.

12 De Workshops

Zoals in de inleiding vermeld is een viertal workshops georganiseerd, te weten rond de thema's:

- Environment&Sustainability (28-01-2016);
- Life Sciences&eHealth (19-01-2016);
- Humanities&Social Sciences (20-01-2016);
- Physics&Beyond (25-01-2016).

Het format van de workshops bestond uit:

- Welkom door de voorzitter van ePLAN;
- Presentatie uitkomsten enquête;
- Presentatie(s) leidende personen uit het domein;
- Discussie over de uitkomsten van de enquête, aanvullingen, overige bevindingen
- Conclusies en afsluiting.

De workshops hebben duidelijke accentverschillen aan het licht gebracht tussen de prioriteiten en behoeftes in de verschillende onderzoeksrichtingen en nuttige aanvullingen opgeleverd voor de optimale ICT-infrastructuur voor de toekomst, met bijbehorende ondersteuning en diensten.
In hierna volgende vier verslagen staan daarvan de details vermeld. De uitkomsten zijn meegenomen in de conclusies en aanbevelingen.

Voor alle workshops geldt dat er grote behoefte werd geuit aan:

- Datavoorzieningen:
  - Opslag capaciteit;
  - Toegankelijkheid/diensten op FAIR grondslag;
  - Analyse tools gericht op de (hoofd-)disciplines;
- Algemene informatie- en service-voorzieningen
  - Lokale en nationale loketten voor informatie over alle resources, diensten en toegangsregelingen;
- Beleid op het gebied van Data Stewardship en Software Sustainability.

12.1 Workshop Life Sciences&eHealth

Opmerkingen direct naar aanleiding van de enquête resultaten

Men verwacht dat ongestructureerde data meer gaan groeien dan uit de enquête blijkt. Zelfs bij gebruik van Electronische Patienten Dossiers (EPD's), zullen goede taalanalyse hulpmiddelen nodig zijn om informatie in "vrije tekst velden" te kunnen gebruiken en interpreteren in onderzoek, dat doorgaans gestructureerde data vereist om vergelijkbaarheid over patiënten te verkrijgen.

Er wordt in de enquête geen onderscheid gemaakt tussen experimentele data beleid en knowledge-beleid. Voor beide is een vernieuwde infrastructuur nodig.

Bij het gebruik van nationale (of zelfs algemene instellings-) ICT-infrastructuur is het grootste probleem dat veel data niet van de beschermde omgeving van bijvoorbeeld ziekenhuizen ("van achter de firewall") naar de externe infrastructuur mag worden verplaatst. De analyse vraagt om snelle rekenvoorzieningen die binnen de firewall niet beschikbaar zijn voor data die vanuit de beschermde omgeving achter de firewall onder de huidige omstandigheden niet mogen passeren.

Onderwerpen tijdens de workshop

- Verhogen van de ontwikkel- en invoeringssnelheid nieuwe behandelperioden (30 -> 15 jaar);
- Personalised Medicine;
- Verkleinen van het gat tussen onderzoek en klinische toepassingen;
- Meer genetische en imaging methoden.

Noodzakelijk daartoe:

- Meer samenwerken, ook tussen disciplines;
- Meer standaardisatie van gegevens, want meer behoefte aan uitwisseling;
- Verbeteren datakwaliteit aan de bron;
- Data sharing op een veilige & privacy-aware manier;
- Systematisch patiënten betrekken bij het proces;
- Verbeteren data representatie voor niet-specialisten;
- Effectieve systemen voor het ondersteunen van klinische beslissingen;
- Beheer, bescherming en toegang tot grote data volumes gegenereerd door nieuwe genetische en imaging methoden;
- Het maken van sustainable services (meer dan een proof-of-concept infrastructuur);
- HPC-gebruik makkelijker (liefst onzichtbaar voor gebruiker; verbergen achter webapps);
- Analyse diensten:
  - Makkelijk zelf aanvullende mogelijkheden toevoegen
  - Reproduceerbaar
  - Deelbaar
  - Gebenchmarked
  - Gestandaardiseerd;
- (dubbel) gepseudonimiseerd koppelen van data van dezelfde individuen in verschillende registraties (Master Person Index), inclusief Person consent;
- Ondersteuning en training;
- Loketfunctie voor alle vragen, op nationaal én lokaal niveau;
- Data verzamelfunctie ter ondersteuning van mobiele Apps (eHealth domein);
- Duurzame data en metadataopslag (> 50 jaar). Longitudinaal onderzoek epidemiologie.

Top 4 prioriteiten

- Duurzaamheid en herbruikbaarheid van data volgens FAIR principes (Findable, Accessible, Interoperable, Reusable);
- Support, met loketfunctie;
- Domeinspecifieke dataservices;
- Met EUDAT een gecertificeerde secure "dropbox" for Europe (B2drop) tot stand brengen.

12.2 Workshop Humanities&Social Sciences

Vragen direct naar aanleiding van de enquête

Bij de vraag over geheugenintensief rekenwerk is de hoeveelheid geheugen niet gekwantificeerd in de vragen. Antwoord: dat klopt. Het antwoord zal passen bij de subjectieve ervaringen van de gebruiker.

Graag wil men ook analyse een analyse van de antwoorden per vakgebied. Antwoord: voor sommige vragen zullen deelanalyses in het hoofddocument worden opgenomen. Ook zullen de data na anonimisering openbaar beschikbaar komen.

Waarom zijn alleen universiteiten aangeschreven? Antwoord: aangeschreven zijn in principe alle hoogleraren, UD's en UHD's, ongeacht hun affiliatie, voor zover wij over hun adresgegevens konden beschikken. Verwezen wordt naar de discussie over de representativiteit van de enquête zoals beschreven in het analysedocument.

Onderwerpen tijdens de workshop

- "Humanities" is een erg generieke term, maar er gaat veel diversiteit/pluriformiteit achter schuil;
- Taalkunde is sterk in Digital Humanities, Geschiedenis, Letterkunde minder op de voorgrond;
- Archeologie, Musicologie, etc. speciale spelers;
- Nieuwe ontwikkelingen belichaamd door CHAT, Clariah, etc.;
- Alles sterk afhankelijk van digitale bronnen, zoals van NIOD, KB, Meertens instituut, Wikipedia;
- Onvolledigheid bronnen punt van zorg (bronnenkritiek extra belangrijk)
- Onderzoekers willen data selectie uit alle bronnen, liefst integraal;
- Wat gebeurt er met software, tools, na afloop van projecten (CATCH);
- Zorgen over innovatie waarbij zowel de informatica- als de humanities-component afzonderlijk innovatief moeten zijn, in plaats van de unieke combinatie daarvan.

Noodzakelijk daartoe:

- Nationale data-infrastructuur, inclusief beheer en services (à la DANS);
- Data voorzien van meta data en specifieke analyse tools;
- Nationale licenties op data;
- Een beveiligde omgeving voor data met copyrights, open voor onderzoekers, (i.t.t. het algemene publiek);
- Een boekenscan service;
- Kwaliteitstoetsing bij digitalisatie (data Seal of approval, ISO normen) en betere OCR;
- Aandacht voor software sustainability;
- Rekenen nog beperkt, maar "Watson"-ontwikkelingen worden gevolgd en zullen gevolgen hebben;
- Behoefte aan support (à la NLeSC), waarbij juist aandacht is voor innovatieve combinaties van ICT en onderzoek;
- Trainingen.

Top 3 prioriteiten

- Digitalisatie van bronnen;
- Beleid voor behoud van tools en software (software sustainability);
- Analyse tools voor humanities boven op een algemene infrastructuur, ondermeer wederzijds aansluiten op Clariah.

12.3 Workshop Physics&Beyond

Vragen direct naar aanleiding van de enquête

Een vraag over de termen geheugenintensief en I/O intensief. Zie antwoord bij 12.2.
Ook binnen de gemeenschap die valt onder Physics&Beyond, die meer dan gemiddeld wel op de hoogte is van de nationale ICT-Infrastructuur zijn er nog grote groepen die hier onvoldoende van weten en ook die meer ondersteuning nodig hebben voor bijvoorbeeld het gebruik van een supercomputer.

Onderwerpen tijdens de workshop

- De groep supercomputergebruikers die meer dan 1 miljoen (!) core hours op jaarbasis gebruikt is met zo'n 300, aanzienlijk;
- Voor de volgende onderzoekdomeinen is toegang tot een supercomputer essentieel: Energy Research (Materials, Chemistry, Photosynthesis, Water splitting), Heterogeneous Catalysis, Molecular Dynamics, Quantum Chemistry, Biochemistry, Climate&Water, Simulating Flows&turbulence, Astrophysics en Astronomy (data reduction&handling) en meer;
- Investeringsniveau nationale ICT-infrastructuur heeft lang geen gelijke tred gehouden met de sterk toegenomen behoeftes. Dat betreft primair de beschikbare rekenkracht, maar zeker ook de voorzieningen voor Big Data en het investeringsniveau bij de instellingen zelf;
- Parallelle verwerking is de enige oplossing voor de achterblijvende rekenkracht per processor. Maar voor systemen met 100.000 tot meer dan 1.000.000 cores vereist parallellisatie meer expertise, support en hoogwaardige kennis over algoritmen en systemen;
- Ook tussen supercomputer en (lokaal) cluster zijn voorzieningen nodig. Tier-2 niveau systemen bij instellingen of desnoods centraal gecoördineerd, HPC-cloud;
- LISA, het nationale cluster is altijd vol. Een dag wachten op de start van een job is "gewoon";
- Op Europees niveau zou in de noodzakelijke diversiteit moeten worden voorzien;
- Toegang tot Europese (PRACE) machines is erg noodzakelijk, maar dat vereist dat er ook nationaal voorzieningen zijn (toegangseis), dat de codes grote aantallen processoren kunnen gebruiken (>10.000), wat vooraf bewezen moet kunnen worden en dat er in de zeer nabije toekomst €1,5 miljoen moet worden bijgedragen;
- Internationale samenwerking én competitie zijn erg belangrijke aspecten.

Noodzakelijk daartoe:

- Nationale supercomputer maximaal een factor 10 langzamer dan de internationale top (verschil tussen 1 dag rekenen (VS) en >100 dagen onafgebroken rekenen (NL), voor één resultaat is onaanvaardbaar);
- Investeringsniveau supercomputer (hardware alleen) daarom van naar schatting €20-25 miljoen;
- Een organisatie die met de onderzoekers en hun steun kan helpen bij een evenwichtige besluitvorming over de nationale ICT-infrastructuur, rekening houden met de diepte en de breedte;
- Computational Science community building;
- Ook voor prototyping (waaronder voorbereiding voor PRACE applicaties) zijn resources nodig (à la DAS-x, LISA);
- Voorbereidingen treffen voor de data storm en bijbehorende rekenbehoefte van de nabije toekomst: 50 PB/dag (!)+ 3 Exaflop/s in 2020 tot 10 EBytes/dag en 30 Exaflop/s in 2028!;

Top 3 prioriteiten (afgeleid)

- **Een supercomputer, max. factor 10 langzamer dan de internationale top (hardware €20 à 25 miljoen) met aanvullend resources voor Big Data (Cartesius moet worden vervangen in 2018);**
- **Nationaal coördinerende organisatie met onderzoekers voor afstemming noden<-> infrastructuur;**
- **Ondersteuning à la SURFsara en NLeSC voor super-massaal parallel rekenen en meer.**

12.4 Workshop Environment&Sustainability

Vragen direct naar aanleiding van de enquête

Opgemerkt wordt dat veel mogelijkheden voor data sharing meer zijn ingericht voor peer-to-peer communicatie dan voor grote groepen.

Onderwerpen tijdens de workshop

- Klimaatcodes moeten een oceaanmodel resolutie krijgen van 1-10 km. Nu is dat nog te vaak 100 km. Daardoor valt een groot gedeelte van de warme golfstroom bijvoorbeeld maar in één gridbox.
- In het verlengde daarvan: fijnere schalen, meer processen, langere tijdschalen (meer tijdstappen), meer gedetailleerde analyses, data-assimilatie en efficiënte visualisatie;
- Data sets zijn te groot om lokaal te worden bewaard. Afgeleid probleem: data bij publicaties kunnen niet meer separaat worden aangeleverd. Men kan op zijn best toegang krijgen tot de bestanden op de locatie waarop zij zich dan bevinden;
- Analyse “on-the-fly” kan exabytes aan data storage voorkomen, maar problemen bij herhaalbaarheid en bij publicaties;
- Clouds for computing zijn geen reële optie voor complexe software systemen zoals klimaatmodellen, maar voor data sharing zou een gedifferentieerde cloud-omgeving denkbaar zijn;
- DAS-5 heeft GPU’s, LISA niet. Er is behoefte aan GPU-resources;
- Behoefte aan cursussen voor PhD’s en post-docs in Fortran, MPI, etc.
- Ondersteuning met menskracht essentieel: diensten van SURFsara en NLeSC worden genoemd;
- Satelliet-data omvangrijk: rekenkracht naar data resources brengen ipv andersom?
- Community-building is belangrijk, ook voor onderwijs, opleiding en latere kennisoverdracht naar het bedrijfsleven;
- Data en software sustainability worden steeds belangrijker.

Noodzakelijk daartoe:

- Aanzienlijk meer HPC resources;
- Inbreng organiseren van onderzoekersveld naar het nationale ICT-beleid;
- Meer ondersteuning, op technisch niveau (SURFsara) en op involvement-niveau (NLeSC);
- Organiseer/coördineer/informeer (over) opleidingszaken;

Top 4 prioriteiten

- **Nederland moet het ambitieniveau voor HPC aanzienlijk hoger stellen**
- **Er is behoefte aan een gestructureerde manier van invloed van het onderzoekersveld op het nationale ICT-beleid+ community building**
- **Gedifferentieerde data clouddiensten**
- **Ondersteuning en communicatie**

13 Afrondend

De enquête betreffende de nationale ICT-infrastructuur in brede zin –bijbehorende diensten en services, expert centra en in Europese en internationale context geeft een goed beeld van de huidige stand van zaken en toekomstwensen, maar het is slechts een snapshot. Niet alle vragen lenen zich voor beantwoording door zo'n breed publiek als "de Nederlandse onderzoekswereld". Maar toch hebben veel onderzoekers het belang ervan ingezien en de moeite genomen. Het lijkt verstandig om een dergelijke bevraging vaker te organiseren, maar niet vaker dan nodig is om te beleid rond de nationale ICT-infrastructuur te kunnen ijken en bijstellen. De frequentie en verbeteringen in de vraagstelling (die er beslist zijn) zullen onderwerp van discussie zijn.
Een bijkomend aspect van de enquête is gebleken, namelijk bewustwording bij de respondenten over het bestaan van een nationale ICT-infrastructuur en van centra en instituten die op dat vlak de onderzoekers ten dienste staan om het onderzoek te bevorderen. Daarmee wordt deze analyse afgesloten.