European Open Science Cloud

Idealen, opinies en ambities
vanuit de Nederlandse wetenschap

Uitkomsten van de ePLAN-workshop DIGITALIA op 22 december 2016 bij SURF in Utrecht.

Januari 2017

Samenvatting

De Nederlandse onderzoeksgemeenschap heeft dringend behoefte aan een aantal duurzame infrastructurele voorzieningen ten behoeve van hoogwaardig data-gedreven onderzoek. Dit blijkt uit de ePLAN-workshop DIGITALIA van december 2016, waarin een brede selectie van de Nederlandse wetenschap zich richtte op het formuleren van de Nederlandse inbreng voor de European Open Science Cloud. De belangrijkste behoeften zijn:

1. **Een nationale open science cloud, in Europees perspectief, die minimaal voldoet aan de volgende vereisten:**
    a. Laagdrempelige toegang
    b. Veilig voor privacygevoelige data
    c. Veilig tegen verlies en inbraak
    d. Transparant m.b.t. data-opslag en eenvoudig in data-transport
    e. Voorzien van hoogwaardige resources, die zowel de breedte als de diepte bedienen
    f. Goed geïntegreerd t.a.v. de verschillende types resources
    g. Snel in respons
2. **Een efficiënt en weinig belastend data- en softwarebeleid, omvattend:**
    a. FAIR data-beleid, met geschikte implementaties per discipline
    b. Passende research data management- en software sustainability-plannen per discipline
    c. Beleid gericht op data-meerwaardecreatie: metadata voor herbruikbaarheid
    d. Voorlichting en goede communicatie over meerwaarde van open data
    e. Oog voor privacy, persoonsgevoelige data en data security
3. **Ondersteuning:**
    a. Op academisch niveau (eScience) in nauwe samenwerking met individuele onderzoeksgroepen
    b. Op nationaal niveau, in de vorm van voorlichting en communicatie
    c. Op voorzieningenniveau: technische informatie
    d. Op lokaal niveau: laagdrempelige toegang tot kennis en informatie
4. **Samenwerking:**
    a. Een nationale infrastructuur in Europese context
    b. Gericht op het vereenvoudigen van internationale samenwerking
    c. Infrastructuur bouwen in co-development aanpak, gezamenlijk met onderzoekers, beleidsmakers en providers

Inhoudsopgave

1 Inleiding

1.1 Achtergrond

Dit rapport is gebaseerd op de DIGITALIA-workshop, georganiseerd door ePLAN op 22 december 2016 bij SURF in Utrecht. Deze workshop was bedoeld om actuele gegevens te verzamelen vanuit zowel de wetenschappelijke wereld als die van beleidsmakers om vanuit onderzoekersperspectief de ambities te kunnen formuleren voor de European Open Science Cloud (EOSC), uitgaande van wetenschappelijke behoeften en door wetenschappers voorziene kansen voor Nederland. In overleg met een schrijfteam dat de Nederlandse bijdrage aan de EOSC gaat formuleren, brengt ePLAN daarom dit rapport uit.

Hoewel er consensus is over de behoefte aan en de algemene kaders van een European Open Science Cloud, bestaat de behoefte om gerichter in kaart te brengen op welke manier de Nederlandse wetenschap invulling zou willen geven aan het initiatief. Om de discussies te stroomlijnen tijdens de workshop, stonden drie hoofdthema's ten aanzien van de European Open Science Cloud centraal:

- Idealen
- Opinies
- Ambities

Belangrijke vragen die binnen deze thema's zijn meegenomen, waren:

- Welke *urgentie* voelen wetenschappelijke disciplines ten aanzien van Open Science en de European Open Science Cloud, en voor een nationale cloud als infrastructuur?
- Welke *obstakels* zijn er in het wetenschappelijke proces, in relatie tot digitale infrastructuur, die een succesvolle implementatie van een open science cloud kunnen hinderen?
- Welke *prioriteiten* moeten worden benoemd waardoor de meeste wetenschappelijke vooruitgang kan worden geboekt?
- Welke *succesvolle toepassingen* van open science zijn er die verbreed en verstevigd kunnen worden door structurele nationale en internationale samenwerking?
- Op welke gebieden loopt Nederland *voorop* en zou het een leidende rol zou kunnen spelen. Hoe kun je die gebieden mobiliseren om een leidende rol waar te maken?

1.2 Workshopproces

Aan de bijeenkomst deden 26 deelnemers mee, bestaand uit onderzoekers uit verschillende wetenschappelijke disciplines (o.a., natuur- en sterrenkunde, chemie, geesteswetenschappen, sociale wetenschappen, klimaatonderzoek, informatica, gezondheidsonderzoek) en een aantal beleidsmakers. Het was hiermee een directe

confrontatie tussen onderzoekers en beleidsmakers, met het doel een zo direct mogelijke uitwisseling van gezichtspunten te bereiken, ingekaderd door zowel de “dromen” van onderzoekers als het realisme voor het haalbare.

De groep werd na presentaties van Wilco Hazeleger (voorzitter ePLAN en directeur Netherlands eScience Center), Erik Fledderus (directeur SURF) en Barend Mons (LUMC, voorzitter High Level Expert Group European Open Science Cloud) ingedeeld in drie werkgroepen die elk de drie hoofdthema’s en gerelateerde vragen hebben besproken. Notulen daarvan zijn ter plekke door de voorzitter digitaal en openbaar vastgelegd in drie werkdocumenten, die hiermee de basis vormen van dit rapport. Na afloop zijn de voorlopige uitkomsten van de drie werkgroepen plenair gepresenteerd en bediscussieerd.

1.3 Een goede start

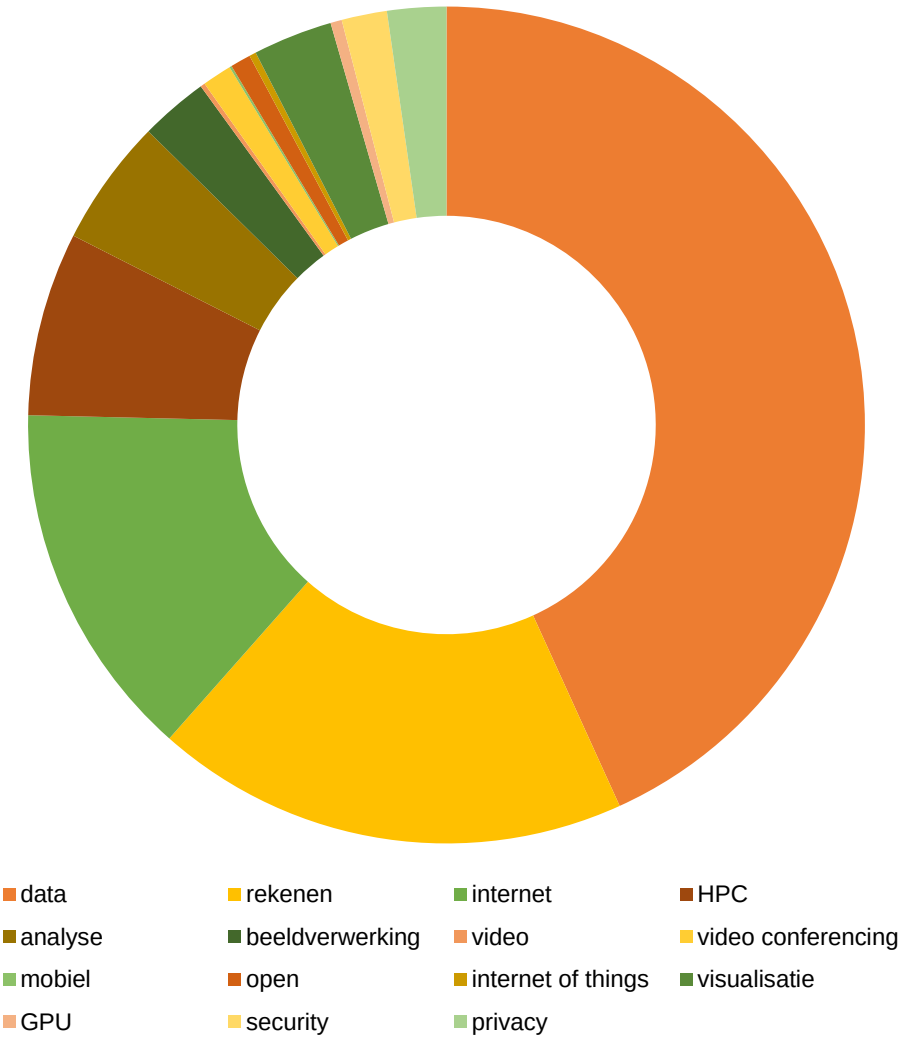
ePLAN publiceerde in 2016 de uitkomsten van een enquête over de ICT-infrastructuur, waaraan meer dan 1000 Nederlandse hoogleraren, UD’s en UHD’s hebben meegewerkt uit vrijwel alle wetenschappelijke vakgebieden. Een aantal uitkomsten worden in dit rapport herhaald en verwerkt in de resultaten, omdat ze de uitkomsten van de workshop versterken, verduidelijken of aanvullen. Zo is in de enquête onder andere expliciet de vraag gesteld of men behoefte zou hebben aan een nationale cloudvoorziening. Eerder was al vastgesteld dat vooral het gebruiksgemak de weg naar de cloud bepaalde. Het antwoord op deze vraag was bijzonder afgebakend: 81% van de wetenschappers die deze vraag beantwoordde, gaf aan behoefte te hebben aan een nationale cloudvoorziening.

Eveneens bleek uit de enquête dat er grote behoefte is aan support op alle niveaus, zowel lokaal als nationaal. De specifieke behoeftes verschillen per discipline, maar in het algemeen werd het ontbreken van support of de toegang tot kennis als obstakel gezien voor snellere ontwikkelingen binnen het vakgebied. Dit impliceert dat de discussie over een Open Science Cloud in Nederland bij voorbaat door een grote meerderheid van onderzoekers in Nederland zou worden ondersteund.

2 Idealen

Om inzicht te krijgen in de uitdagingen die onderzoekers zich in een langetermijnperspectief stellen om hun onderzoek en hun vakgebied verder te ontwikkelen, begonnen de discussies met het thema *Idealen*. Daarbij was het niet de bedoeling de dromen bij voorbaat te beperken door de beperkingen van de huidige technologie en infrastructuur, maar juist het bredere kader in beeld te krijgen. Om de idealen in het kader te plaatsen van de discussie over open science cloud-omgevingen, zijn die door de onderzoekers vertaald in termen van een ideale werkomgeving (infrastructuur).

In de eerder genoemde enquête van ePLAN, gepubliceerd in februari 2016, is het volgende beeld naar voren gekomen van de elementen die bij een ideale infrastructuur aan de orde zouden moeten komen:



Met betrekking tot de EOSC-discussie zijn tijdens de workshop de onderstaande verfijningen en aanvullingen gekomen. Onderzoekers in veel van de toepassingsdomeinen (in tegenstelling tot onderzoekers die fundamenteel data- en computeronderzoek doen), hebben vooral behoefte aan:

Algemene ICT-infrastructuur

- Tailor-made oplossingen, dat wil zeggen aanpassingen van generieke oplossingen inclusief “the last mile” (tot aan de computer van de eindgebruiker zelf toe);
- Interfaces naar de ICT-infrastructuur in hun (domein-)taal en aangepast aan de specifieke domeinbehoefte;
- Een technische omgeving die intuïtief is en context-afhankelijk;
- Single sign-on. AAI, incl. two-factor authentication en rights management;
- Verbetering van infrastructuren (inclusief sociale) met het oog op interdisciplinaire en multi-disciplinaire samenwerking;
- Een information universe (in silico), waarin curation, provenance en dergelijke, vanzelfsprekend zijn.

Het datadomein

- Betere mechanismen en afspraken ter bevordering van de betrouwbaarheid van data;
- Beschikbaar maken en ontwikkelen van betere analysemethoden om garanties te kunnen geven voor kwaliteit en integriteit. Praktisch gezegd: om “nep-data” te herkennen;
- Betere interoperabiliteit van data;
- Betere wetgeving rond data: “the right to read=the right to mine”
- Betere erkenning voor onderzoekers die het belang van (open) data en software ondersteunen en daar naar handelen;
- Mogelijk maken van de ultieme reproduceerbaarheid van onderzoeksresultaten (daar waar van toepassing);
- Reproduceerbaarheid over tien jaar van onderzoeksresultaten van heden;
- Keten van data-productie moet transparant zijn (zoals van ruwe data naar onderzoeksresultaat).

Skills

- Opleiding in de basis is nodig voor meer basaal begrip over het belang van data en software;

- Opleiding op alle niveaus is nodig voor meer kennis over data, formats, uitwisselbaarheid;
- In de opleiding hoort aandacht voor herbruikbaarheid van data en software en voor reproduceerbaarheid;
- Opleiding moet zorgen voor een cultuurverandering reeds bij de basis, met oog voor open beschikbaar stellen van data;
- Er moet meer aandacht zijn voor de skills om data te vinden;
- Kalibratie/standaardisatie bij data en data-acquisitie van groot belang, zeker bij het delen van data over grenzen van disciplines, organisaties en landen.

De veranderingen in de manieren waarop we ons onderzoek moeten inrichten worden treffend weergegeven door de uitspraak: “Vroeger was iedere chauffeur ook monteur”. In de wetenschap wordt nog niet overal erkend en herkend dat ICT zo dynamisch is en de oplossingen die ICT kan bieden zo veel complexiteit achter zich hebben, dat onderzoekers veel zaken aan experts kunnen overlaten die kennis over de actuele ICT als werkdomein hebben. De rol van het eScience Center en veel van de data science-centra van de universiteiten, als typisch intermediair tussen wetenschappelijke vraagstellingen en praktische oplossingen, is daarvan een goed voorbeeld. Bij het vervullen van deze intermediaire rol, is aanwezigheid van domein-specifieke kennis, naast kennis van ICT en in het bijzonder data, een belangrijke voorwaarde.

3 Opinies

Bij dit thema ging het om de visies over de EOSC en over de kijk op de Nederlandse functie in dat geheel.

De cloud als drager van een infrastructuur

- Het ontzorgen van de onderzoeker is van belang. Commerciële partijen (cloudaanbieders) lijken hier beter in te slagen dan de publieke. Met behoud van security én laagdrempelig.
- Het (academische) publieke bestel is groot genoeg om over een eigen cloud te beschikken. Privacy-gerelateerde problemen kunnen comfortabeler worden aangepakt, omdat er in publieke partijen meer natuurlijk vertrouwen bestaat waar het gaat om bescherming en duurzaamheid.
- Een goede balans tussen kosten en baten is belangrijk. Maar de kosten gaan ook hier voor de baten uit.
- De huidige situatie, dat zowel de nationale als de Europese ICT-infrastructuur wetenschapsbreed nauwelijks bekend is, is ongewenst.
- De bredere publieke sector moet beter aansluiten op de academische. Er zijn heel veel publieke data en organisaties die daarbij betrokken zijn, maar die hebben van FAIR-principes en dergelijke nog nooit gehoord.
- Er moet meer aandacht worden besteed aan zendingswerk om de belangen van infrastructuur en data voor het voetlicht te brengen. Er bestaat veel onbekendheid.
- De Europese Open Science Cloud kan alleen bestaan bij de gratie van aaneengesloten nationale initiatieven (clouds). Het is “slechts” de som der delen waaraan op Europees niveau meerwaarde wordt gegeven. Dus zijn nationale investeringen onontbeerlijk om de Open Science Cloud op Europees niveau een succes te maken.
- Leren van ervaringen uit het verleden betekent in deze context dat er uit Europa niet alleen iets gehaald kan worden, er moet ook op moet worden ingezet en ingebracht. Veel eerdere infraprojecten (waaronder PRACE en EGI) lopen mank door een gebrekkig bekostigingsmodel.
- Een academische cloud is succesvol als die voldoende gebalanceerd is in termen van beschikbare resources. Niet alleen de breedte bedienen (bandbreedte, clusters, storage, ease-of-use) maar ook de diepte (low latency, capability computing, big data opslag, leading edge).

De opzet van een European Science Cloud staat niet op zich zelf. Ook in het bedrijfsleven en in de publieke sector is de inzet van big data, in verschillende gradaties van openheid, en sourcing onderwerp van dagelijks debat. De komst van publieke clouds (dat zijn clouds in de *private* sector) heeft een enorme impact gehad en nog steeds op de bedrijfsvoering van

grote en MKB bedrijven. Ook daar is aan de orde: wat doe je in huis, wat besteed je uit, welke kennis heb je zelf nodig om goede beslissingen te kunnen nemen en contracten te kunnen sluiten. Aanverwante onderwerpen, zoals de Internet of Things en Block Chain brengen deze vraagstukken nog dichter bij. In de publieke sector in Nederland werkt de overheid aan richtlijnen voor het open delen van data (data.overheid.nl). Maar lijkt er lijkt nog geen brug te bestaan tussen de discussies in het bedrijfsleven, de publieke sector en de academische wereld. Zo is de term “FAIR”, die binnen het academische domein inmiddels gangbaar is, in de publieke sector nog onbekend.

Bij de Nederlandse positionering voor de European Open Science Cloud moet dan ook serieus nagedacht worden over de deelname van het bedrijfsleven en de publieke sector aan (eventueel een deel van de) activiteiten die Nederland in dat kader zal kunnen opzetten.

Het FAIR datamodel als concept

- Het is goed mogelijk een associatie te maken tussen de termen Findable-Accessible-Interoperable-Reusable die van toepassing zijn op individuele bestanden en de uitgangspunten waarop het Keurmerk Data Seal of Approval zijn gebaseerd voor data-archieven.
- Data moeten zo FAIR zijn dat ze over disciplines heen gebruikt kunnen worden.
- Eén mogelijke interpretatie van de FAIR-componenten leidt er toe dat de F, A en de I ten dienst staan van de R (“F+A+I=R”). Er wordt ook over een toevoeging gesproken: FAIRR, met de R van reproducible (bovenop reusable). Reproduceerbaarheid is heel belangrijk in bepaalde domeinen, maar is als principe domeinspecifiek.
- De I en de R moeten mogelijk domeinspecifiek geïmplementeerd moeten worden. Wel met supra-disciplinair gebruik voor ogen.
- Al deze zaken kunnen in domeinspecifieke protocollen geïmplementeerd worden.

Software

- Onderzoekssoftware en data moeten -beleidsmatig- gelijkwaardig worden behandeld.
- In de uitwerking en behandeling vragen “data” en “software” om andere oplossingen.
- Bij reproduceerbaarheid van onderzoeksresultaten is de correcte vermelding van gebruikte software, alsmede het beschikbaar houden van de betreffende software onontbeerlijk.
- Nederland met meedoen met internationale activiteiten en planningen voor uitwisseling van kennis en informatie over software: beheer, up-to-date houden, toekomstvast maken, etc.

Skills

- Opleiding, training en een permanente brugfunctie zijn nodig voor een optimale benutting van de infrastructuur;
- Opleiding: omgang met data en data analytics integraal onderdeel maken van hoger onderwijs;
- Training: “software carpentry” en vergelijkbare activiteiten nodig, gericht op de gevorderde onderzoeker; idem voor data;
- Speciale aandacht geven aan een nieuwe “professie” van data scientist;
- Inzet academic support (“eScience”) bij uitdagende, grensverleggende science projecten;
- Technische support nodig, nationaal en lokaal.

4 Ambities

Leidend voor de Nederlandse ambities in het kader van de EOSC zijn de wetenschappelijke uitdagingen die voor Nederland van belang zijn of waarin Nederland excelleert. Nederland verkeert hierbij in een gunstige uitgangspositie, omdat de ICT-infrastructuur in belangrijke mate in nationaal beheer is en binnen één organisatie afwegingen gemaakt kunnen worden tussen verschillende componenten en zorg gedragen kan worden voor onderlinge samenhang. Op weg naar een praktische cloud-achtige infrastructuur is dat een groot voordeel. Anderzijds zijn de middelen voor de academische sector om op infrastructuurgebied actief op een serieuze schaal in Europees verband mee te doen beslist onvoldoende.

ICT-infrastructuur algemeen

Nederland moet een eigen, hoogwaardige, Open Science Cloud opzetten en operationeel maken met een serieus potentieel om ook voor andere (Europese) landen interessant te zijn.

- Nederlandse “resource coins” moeten in het buitenland ingezet kunnen worden, maar de ambitie is dat buitenlandse coins op aanzienlijke schaal in Nederland besteed gaan worden.
- Nederland is organisatorisch uitstekend ingericht om hoogwaardige diensten aan het buiten land te kunnen leveren en zo ook zelf te profiteren van het bovenmodale aanbod.
- De Nederlandse kennisintensiteit en betrouwbaarheid, onze neutraliteit en connectiviteit moeten (via de coins) vermarkt kunnen worden.
- Zet de ontwikkeling van de Nederlandse Open Science Cloud zo op dat er sprake is van co-development, met de onderzoekers en de ICT-infrastructuuraanbieders als partners.
- Zorg voor een supportinfrastructuur, van hoog academisch (eScience) tot dagelijkse praktijkniveau, zodat de infrastructuur zo kosten-effectief mogelijk wordt ingezet en dat de juiste resources bij de juiste infrastructuur worden ingezet (fit-for-purpose).

Specifieke ambities op data-gebied

Nederland moet een brede hoogwaardige datainfrastructuur opzetten, onderhouden en aanbieden, die aansluit bij de huidige bijzondere expertises die in Nederland al bestaan:

- Kwaliteitstoetsingen
- Meta-data-plus, en parameters voor kwaliteit
- Data en meta-data zo inrichten dat ze machine-vindbaar en machine-interpreteerbaar zijn
- Links met de wetenschap achter de data
- Mogelijkheden voor bijzondere privacy-waarborgen die wetenschappelijk gebruik, ook op afstand, niet in de weg staan

- Mogelijkheden voor mensgericht en “eHealth”-type onderzoek
- Investeren in (zeer -) hoge snelheid(bandbeedte) en low-latency netwerkcapaciteit voor grote instrumentele wetenschappen (zoals astronomie, SKA-project) en deeltjesfysica
- Investeren in academische data-expertise en data-experts

5 Specifieke onderwerpen

In dit hoofdstuk worden de belangrijkste (aanvullend op de eerdere hoofdstukken) antwoorden weergegeven op de volgende vragen die tijdens de workshop zijn behandeld:

- Welke *urgentie* voelen wetenschappelijke disciplines ten aanzien van Open Science en de European Open Science Cloud als infrastructuur?
- Welke *obstakels* zijn er bij het wetenschappelijke proces, in relatie tot digitale infrastructuur, die een succesvolle implementatie van de EOSC kunnen hinderen?
- Welke *prioriteiten* moeten worden benoemd waardoor de meeste wetenschappelijke vooruitgang kan worden geboekt?
- Welke *succesvolle toepassingen* van open science zijn er die verbreed en verstevigd kunnen worden door structurele internationale samenwerking?
- Op welke gebieden loopt Nederland *voorop* en zou het een leidende rol zou kunnen spelen. Hoe kun je die gebieden mobiliseren om een leidende rol waar te maken?

5.1 Urgentie

Uit de eerder genoemde ePLAN enquête blijkt duidelijk dat er behoefte bestaat aan een nationale cloud-omgeving en -dienst, mits aan een aantal voorwaarden is voldaan:

- Eenvoudig toegankelijk;
- Maar wel heel veilig;
- Meerwaarde ten aanzien van privacy en security;
- Robuust en betrouwbaar;
- Met aandacht voor zowel de breedte als de diepte van het aanbod.

Bij de workshop is gebleken dat men een significante (en dus ambitieuze) Nederlandse rol in de European Open Science Cloud steunt. Mede omdat de huidige nationale voorzieningen versterking behoeven. Daarmee is meteen ook de urgentie helder.

5.2 Obstakels

De wetenschappelijke processen en aandachtsgebieden moeten leidend zijn en de diensten en voorzieningen een antwoord op de behoefte. Maar de relatieve onbekendheid met het bestaan van een nationale e-infrastructuur, laat staan een Europese, is een onbedoelde beperking van de wetenschappelijke mogelijkheden op lokaal niveau. Dat geldt evenzeer voor begrippen als data stewardship, software sustainabilty en daarmee vergelijkbare termen. Deze onbekendheid staat versnelde ontwikkelingen in de wetenschap in de weg. Dit zijn zaken waarop naar verwachting met aanzienlijk rendement kan worden geïnvesteerd.

5.3 Prioriteiten

Vanuit het wetenschappelijk veld wordt vooral gewezen op twee dingen:

- Een achterblijvend aanbod aan resources voor rekenen, analyse en data-voorzieningen;
- Een gebrek aan data-transparantie: hoe bereik je vanuit de verschillende(verwerkings- en analyse-) resources je data. Waar zijn de data als die zich in de cloud bevinden?

Daarnaast wordt vaak verwezen naar de prioriteringen in het kader van de nationale wetenschapsagenda en het daarop aansluiten van de infrastructuurbehoefte.

Vanuit het beleidsveld wordt vooral gewezen op:

- Het potentieel voor vooruitgang bij het (open) delen van data (waaronder het FAIR-principe);
- Aandacht voor “rewarding systems and incentives” als het om verantwoording gaat voor data en software (als onderdeel van het oplossen van het kip-en-ei probleem);
- Aandacht voor impact analyse tools voor data en software.

5.4 Succesvolle toepassingen

Open Science is meer een *principe* dan een toepasbaar onderdeel van onderzoek, net als FAIR meer een uitgangspunt is dat iets waar men dagelijks bij het onderzoek gebruik van maakt. Niettemin zijn er succes stories.

5.5 Koploper

Nederland loopt voorop als het gaat om:

- Geïntegreerd nationaal beleid op het gebied van de nationale e-infrastructuur;
- Beheer van wetenschappelijke data en archieven, inclusief meerwaardediensten;
- FAIR-principe is in Nederland geconcipieerd;
- Aandacht voor software in samenhang met data;
- Holistische benadering van de koppeling en interactie tussen research en infrastructuur
- Het gezamenlijk optrekken van de eScience en data science centra voor hun gemeenschappelijke belangen (ePLAN);
- Het mobiliseren van het Europese eScience en data science potentieel (PLAN-E).

De wetenschappelijk leidende domeinen moeten worden afgeleid van de nationale wetenschapsagenda en daarmee te vergelijken overzichten en initiatieven en zijn geen primaire zaak voor ePLAN.