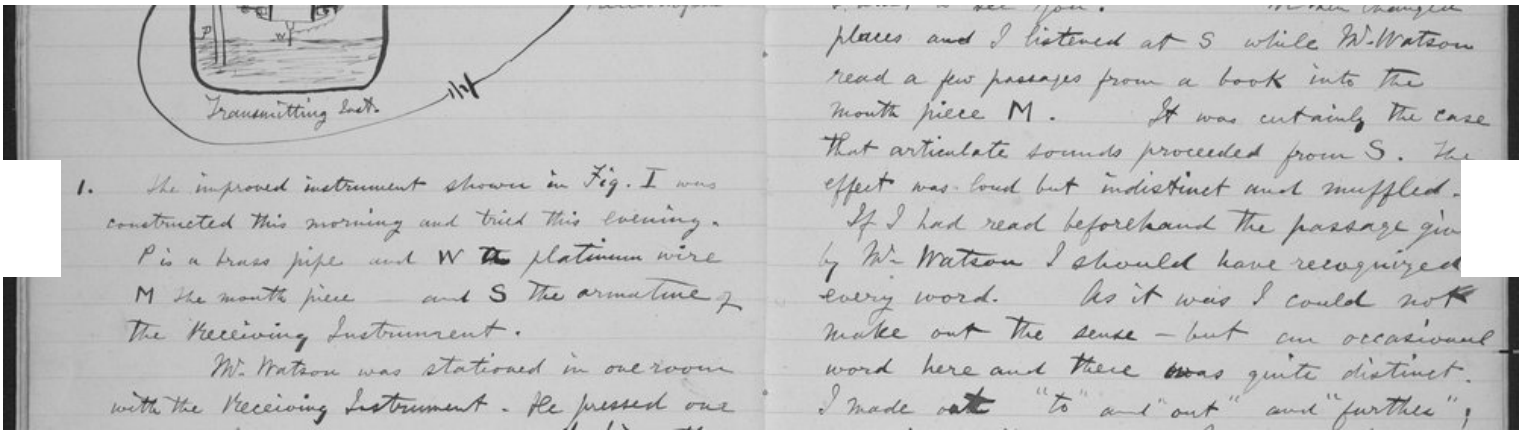


Artikel

Van labboek naar cloudservice

Gepubliceerd: 1 November 2019 13:40

Vanuit de overheid klinkt steeds harder de roep om open wetenschap, open access en open data. Op zich is het niet meer dan logisch dat de resultaten van door de overheid gefinancieerd onderzoek voor iedereen vrij beschikbaar zijn, maar de ambitieuze plannen op dit gebied veroorzaken nogal wat onrust onder onderzoekers. In dit artikel bespreken we specifiek open data en de FAIR-principes en proberen we uit te vinden hoe we van een nood een deugd kunnen maken.

Auteurs: Jasper Smits en Dries van Oosten

De onzichtbare held van de experimentele natuurwetenschap is het labboek. We kennen allemaal het verhaal over de ontdekking van supergeleiding door Heike Kamerlingh Onnes, die in zijn aantekeningenboekje de woorden “Kwik nagenoeg nul” opschreef [1]. Zonder die teruggevonden aantekeningen hadden we nu niet geweten wat het echte verhaal achter deze ontdekking was. Ook kennen we het verhaal van Dan Shechtman die bij sample nummer 1725 schreef “(*10 fold ???*)” en daarmee de eerste waarneming van een quasikristal documenteerde [2-4]. Prachtige voorbeelden van hoe we met onze data om zouden moeten gaan, want de metingen die je doet kunnen nog zo mooi zijn, zonder documentatie van de omstandigheden waaronder je ze hebt gedaan, zijn ze geen knip voor de neus waard.

Maar deze twee voorbeelden illustreren ook een belangrijke verandering van de rol van het labboek, een verandering die wordt gedreven door de enorme toename in de omvang van data die we vergaren. In het experiment van Kamerlingh Onnes moest enorm veel werk worden gedaan om het apparaat op een bepaalde temperatuur in te stellen. Per temperatuur werd met een weerstandsbrug een meting gedaan, die vervolgens in een boekje kon worden opgeschreven. Shechtman daarentegen zat een hele dag achter de transmissie-elektronenmicroscoop met een bakje samples en belichtte de ene fotografische plaat na de andere. Het labboek van Shechtman bevatte niet de data zelf, maar de metadata; de informatie die nodig is om observaties over de enorme hoeveelheid ruwe informatie op die fotografische platen om te zetten in een ontdekking.

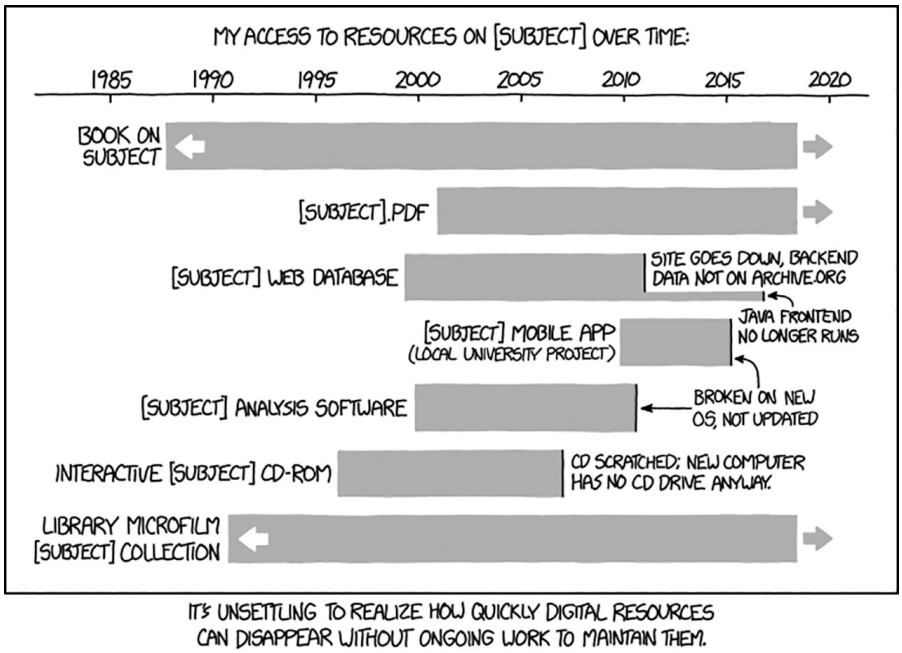
De balans tussen de hoeveelheid data die we produceren en de tijd die ons dat kost, is sinds Shechtman nog veel verder verschoven. Op CERN worden tijdens dataruns gigabytes aan ruwe data per seconde opgeslagen, maar ook een eenvoudig table-top-experiment van een enkele promovendus heeft vaak een tamelijk vlotte harde schijf nodig om de datasnelheid van een moderne digitale camera te kunnen verwerken. En al die nulletjes en eentjes zijn zinloos als je de context ervan niet kent. Metadata zijn dus belangrijker dan ooit, niet alleen voor jezelf, maar ook voor toekomstige gebruikers van de data.

Nederlands Tijdschrift voor Natuurkunde

Het Nederlands Tijdschrift voor Natuurkunde is hét Nederlandstalige vakblad voor natuurkundigen, vol nieuws uit de natuurkunde. Een greep uit de onderwerpen: onderzoek aan de frontlinie, natuurkunde en kunst, onderwijs, bijzondere loopbanen, interviews met prominente wetenschappers en een kijkje in de keuken bij bedrijven. Jaarlijks verschijnt vlak voor de zomer een extra dik themanummer. Leden van de NNV ontvangen het tijdschrift maandelijks.

Speciale editie Zeeën en oceanen

Zeeën en oceanen is het thema van onze speciale editie van 2022. Hoe worden oceanen eigenlijk onderzocht? Hoe ontstaan die zandpatronenn op de bodem toch? Hoe kunnen we zo goed mogelijk voorspellen hoe ons huidige klimaat verandert door te kijken naar vroegere klimaatveranderingen? En mochten we de huidige klimaatverandering niet voldoende kunnen beteugelen en de zeespiegel stijgt aanzienlijk, moeten we dan een dam om de Noordzee bouwen? Diezelfde Noordzee wordt nu flink gebruikt: recreatie, grote windmolenparken en



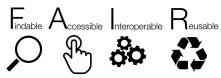
Illustratie: xkcd.com (CC BY-NC 2.5).

Daarnaast is het niet ondenkbaar dat er in die enorme hoeveelheden data resultaten verborgen liggen die er door de onderzoeker zelf niet uitgehaald zijn. Een nieuwe analyse zou die resultaten aan het licht kunnen brengen als een andere onderzoeker er met een nieuwe blik naar kijkt. Het is dan wel essentieel dat deze onderzoeker weet van het bestaan van de data en deze data kan openen, inlezen en verwerken. In sommige vakgebieden is het aanleggen van databanken met onderzoeksresultaten heel normaal. Denk daarbij bijvoorbeeld aan het Corpus Gesproken Nederlands [5], een databank voor het hedendaags Nederlands zoals dat door volwassen sprekers in Nederland en Vlaanderen wordt gesproken. Financieringsorganisaties willen dat zo veel mogelijk onderzoeksdata vrij toegankelijk en herbruikbaar worden opgeslagen en stellen daarom steeds striktere eisen aan het datamanagement van een onderzoeksproject.
Vanuit het werkveld is er echter behoorlijk wat weerstand tegen deze zogenoemde open data, om een aantal redenen. Ten eerste: principes zijn leuk, maar het kost tijd en geld om data zorgvuldig te bewaren, beschrijven en helpen te ontsluiten. Als niemand er dan ooit naar kijkt, had je met dat geld misschien beter apparatuur voor het lab kunnen kopen. Een ander probleem is dat iemand op basis van zijn analyse van jouw data onzin zou kunnen publiceren en daarmee zou suggereren dat jij erachter staat. Er zijn nu al voldoende mensen die klimaatopwarming door de mens ontkennen, die zouden de naam van een klimaatwetenschapper aardig kunnen beschadigen door zijn of haar data te ‘analyseren’. Maar het zou ook kunnen dat iemand daadwerkelijk iets uit jouw data haalt wat jij er niet uit hebt gehaald. Mooi voor de wetenschap, maar als daar een Nature- of een Science-publicatie mee gescoord wordt, is dat wel een beetje zuur voor de onderzoekers die de data hebben gegenereerd.

De FAIR-principes

De FAIR principes [6] zijn in 2014 opgesteld tijdens een workshop in het Lorentz Center in Leiden. FAIR staat voor

Findable - vindbaar
Accessible - toegankelijk
Interoperable - uitwisselbaar
Reusable - herbruikbaar



FAIR en/of open data?
Het is in deze context belangrijk om over FAIR te spreken. De FAIR-principes zijn bedacht om te zorgen dat we data op een herbruikbare manier opslaan. FAIR staat voor findable, accessible, interoperable and reusable. Je moet data kunnen vinden, vervolgens moet je toegang kunnen krijgen tot de data, je moet de bestanden in kunnen lezen en kunnen hergebruiken. Maar, waar het essentieel is voor open data dat je de FAIR-principes gebruikt bij het opslaan, hoeft FAIR niet per definitie open te betekenen. FAIR zegt namelijk dat data moeten kunnen worden gebruikt door degenen die ze zouden moeten kunnen gebruiken, terwijl open betekent dat iedereen het moeten kunnen gebruiken. Compleet open is in veel gevallen (door privacy of intellectueel eigendom) niet mogelijk, maar FAIR zegt in feite dat je, ook als data niet open zijn, je ze wel zorgvuldig op moet slaan dat je ze zonder schaamte openbaar zou kunnen maken.

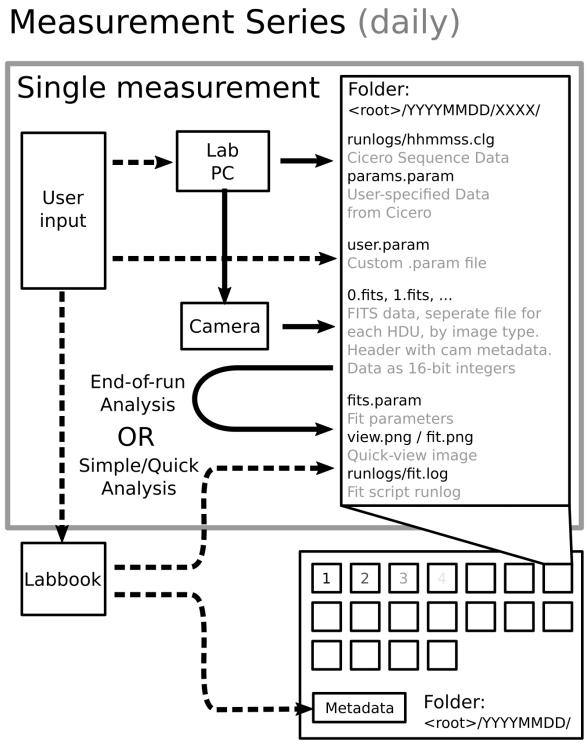
schepen die af en aan varen. Al deze activiteiten maken geluid. Hoeveel en wat doet dat met het onderwaterleven? Vragen, vragen en nog eens vragen. Wij geven antwoorden op deze en nog veel meer vragen in het themanummer Zeeën en oceanen.

Het themanummer Zeeën en oceanen is los te bestellen door een mailtje te sturen aan het NNV-bureau (**bureau@nnv.nl**), vergeet niet om een verzendadres te vermelden. Het nummer kost 10 euro, inclusief btw en verzendkosten (voor verzenden naar het buitenland rekenen we extra kosten).

Als je je data niet opslaat met het idee dat iemand anders het nog een keer moet lezen, maak je jezelf als promovendus graag wijs dat je bij het schrijven van je proefschrift nog wel weet op welke usb-stick of externe harde schijf de data staan die je hebt gebruikt om dat ene plaatje te maken uit die presentatie die je aan het begin van je promotietraject een keer hebt gegeven. Maar zoals we allemaal weten, is dat een illusie. En als promotor word je er ook niet blij van als je na het afscheid van deze promovendus wordt geconfronteerd met een stapel externe harde schijven zonder verdere toelichting. Je moet er niet aan denken dat je over drie jaar een mailtje krijgt van iemand die vraagt of je die data ook eens met een ander model zou kunnen proberen te fitten, want dan moet je die stapel harde schijven door, in de hoop dat ze het überhaupt nog doen. Of dat de computer waar de software op staat die nodig is voor de data-analyse inmiddels niet meer op het netwerk mag worden aangesloten omdat de Windowsversie waaronder deze software draait te oud is.
De FAIR-principes helpen je om dit te voorkomen, door je te dwingen om bij alle data die je opslaat te vragen: “zou iemand anders uit de voeten kunnen met deze data?”. Want die iemand anders, dat zou je zelf over een paar jaar ook kunnen zijn.

Datamanagement in een table-top-experiment

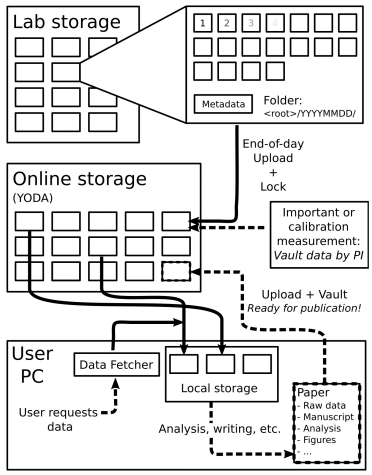
Het experiment met ultrakoude atomen in Utrecht draait volledig geautomatiseerd. Wanneer de uitlijning van de verschillende laserbundels eenmaal klopt kan elke manipulatie uitgevoerd worden door het veranderen van parameters in de aansturing. In dit experiment worden alle data automatisch opgeslagen en geüpload naar een door de universiteit beheerde cloud. Voor elke experimentele run wordt een uniek nummer gedeeld tussen alle systemen die bijdragen aan het verzamelen van data. Vervolgens worden alle data op één systeem opgeslagen en vanaf dit systeem geüpload naar deze cloud. In Utrecht wordt gebruikgemaakt van YODA, een dienst opgezet door de universiteit en gebaseerd op de **iRODS-architectuur**. Alle data worden op meerdere, geografisch gescheiden, locaties gedupliceerd.



In ons experiment worden alle data van één run opgeslagen in een container. Op gezette tijden, bijvoorbeeld aan het eind van een meetdag worden alle containers op de cloud gecontroleerd en hierna vergrendeld. Ook wordt er een checksum gegenereerd. Hierbij kan later de integriteit van de data gecontroleerd worden.
Wanneer data nodig zijn, kunnen deze opgevraagd worden en lokaal worden opgeslagen en verwerkt.

Consequenties
Hoe letterlijk je de FAIR-principes neemt, is een keuze die aan het vakgebied wordt gelaten. Daarom zijn ze door de opstellers bewust principes genoemd en geen richtlijnen. Als je ervoor kiest om de FAIR-principes fundamenteel te interpreteren, dan heeft dat vrij grote consequenties voor de dagelijkse praktijk in het lab. Met name het criterium van interoperabiliteit is een taaie. Volgens de FAIR-principes moeten de data vanzelfsprekend in een open en vrij formaat zijn opgeslagen. Maar als we eisen dat iedereen de data ook echt moet kunnen gebruiken, dan moet het ook zo zijn dat er daadwerkelijk voor iedereen een implementatie van dat formaat beschikbaar is. Met andere woorden, er moet een programma zijn, dat voor iedereen beschikbaar is, waarmee de data kunnen worden verwerkt. En als het voor iedereen beschikbaar moet zijn, dan betekent dat eigenlijk dat het een opensource-oplossing moet zijn. Maar in feite is zelfs dat niet voldoende. Want de oplossing moet ook over tien jaar nog beschikbaar zijn en tien jaar is een lange tijd voor software. Er zijn weinig softwareprojecten die over een tijdschaal van tien jaar stabiel genoeg zijn, omdat de softwaremarkt juist gedreven wordt door verandering.

Datapublicaties



Alle gegevens gerelateerd aan een artikel, zoals de ruwe data, het manuscript en de scripts gebruikt voor de analyse van deze gegevens worden in een speciale container opgeslagen. Ook deze container wordt vergrendeld en van een checksum voorzien. Een dergelijke container kan, samen met het artikel, worden gepubliceerd. Er wordt dan een digital object identifier (DOI) aan toegekend. Iedere geïnteresseerde lezer kan de data dan vinden en downloaden. Een dergelijk datapublicatie koppelen aan een artikel is een ideale manier om de data voor anderen te ontsluiten. Het artikel dient in feite als inhoudsopgave voor de dataset. Soms is het niet wenselijk om data direct te publiceren, bijvoorbeeld omdat er nog een ander manuscript in voorbereiding is die gebruikmaakt van dezelfde dataset. In dat geval kan ervoor worden gekozen om alleen de metadata te publiceren (inclusief de checksum). Als later de data dan wel openbaar worden gemaakt, kan met behulp van de checksum worden vastgesteld of de gepubliceerde dataset inderdaad dezelfde is.

De praktijk

Hoe kunnen we dit in de praktijk aanpakken? We moeten niet de illusie hebben dat we één oplossing kunnen bedenken die in alle gevallen werkt. In de praktijk zullen we ons bij ieder type meting die we doen af moeten vragen hoe we de data opslaan, hoe we de dataset structureren en welke metadata er nodig zijn om later nog iets met de data te kunnen doen. Daarbij moeten we metadata breder interpreteren dan alleen wat zoekwoorden, auteurs et cetera. Een concreet voorbeeld is de data-opslag van het Bose-Einsteincondensatie-experiment in Utrecht (zie kader Datamanagement in een table-top-experiment). In het geval van dit experiment was het relatief eenvoudig om een datamanagementsysteem op te zetten, ten eerste omdat de automatisering van het experiment sowieso moest worden gemoderniseerd, ten tweede omdat de Universiteit Utrecht een zeer uitgebreide data-opslagservice aanbiedt aan haar onderzoekers en ten derde omdat de data in dit experiment zeer uniform zijn. Dat wil zeggen, het zijn altijd plaatjes gemaakt met dezelfde camera. In andere experimenten, met name die waar de onderzoeker cruciale instellingen met de hand moet verrichten, is de opslag natuurlijk veel moeilijker. Denk daarbij bijvoorbeeld aan chemische-synthesestappen, of het polijsten of monteren van een sample. Hierbij is de documentatie van deze stappen cruciaal bij de interpretatie van de meetdata en moet deze documentatie dus ook als onderdeel van de metadata worden gezien.

Conclusie

Open data klinkt in eerste instantie iets dat heel veel werk geeft en waarvan op het eerste gezicht het praktische nut misschien niet duidelijk is. Het opstellen van een concreet systeem om data en alle relevante metadata langdurig op zo'n manier op te slaan dat een vreemde er iets mee kan doen, is enorm gecompliceerd. Maar zoals gezegd, die vreemde zijn we over een paar jaar zelf. En als we onze data nu FAIR opslaan, zal onze toekomstige zelf ons dankbaar zijn.

Over de auteurs: Jasper Smits (1991) studeerde experimentele natuurkunde in Utrecht. Momenteel doet hij daar ook zijn promotieonderzoek in de groep Nanophotonics onder promotor Peter van der Straten. Hij bestudeert hydrodynamische excitaties in Bose-Einsteincondensaten en ontwikkelt een nieuwe niet-destructieve afbeeldingsmethode van spindomeinen in Bose-Einsteincondensaten.

Dries van Oosten (1976) studeerde experimentele en theoretische natuurkunde in Utrecht. In 2004 promoveerde hij in Utrecht onder promoteren Peter van der Straten en Henk Stoof. Na postdocs bij Immanuel Bloch (toen in Mainz) en Kobus Kuipers (toen AMOLF) ontving hij in 2009 een Vidi-beurs van NWO. Sinds 2010 geeft hij op de Universteit Utrecht leiding aan een experimentele groep die onderzoek doet naar Bose-Einsteincondensatie van licht en extreem niet-lineaire optica.

Referenties

1 Dirk van Delft en Peter Kes, "Kwik nagenoeg nul", NTvN 77-04, 122-125 (2011).
2 Claud Biemans, Danny Shechtman over tegenwerking en de kracht van de Nobelprijs, NTvN 78-05, 180-182 (2012).
3 Ted Janssen, De ontdekking van quasikristallen, NTvN 77-02, 454-456 (2011).
4 D. Shechtman, The DiSCovery of QuaSiCrySTalS, Nobel Lecture 2011.
5 **http://lands.let.ru.nl/cgn/doc_Dutch/topics/project/pro_info.htm**.
6 Wilkinson et al., Scientific Data 3, 160018 (2016).