Af Asger Juel Hansen, lajh@kb.dk,
KUB Forskerservice, Det Kgl. Bibliotek

Hvorfor skal det hele være så FAIR?

FAIR står for Findable, Accessible, Interoperable, Reusable. Principperne skal fungere som en international guideline for højkvalitetsmetadata om forskningsdata.
Denne artikel er tiltænkt som en kickstart til en FAIR tankegang hos læserne.

Fremskridt i videnskaberne næres af mulighederne for at dele og tilgå videnskabelige forskningsdata. Derfor er der brug for at udvikle infrastrukturer og services, der muliggør en systemisk ændring af forskningspraksisserne hen imod Open Science og specielt eScience.

Landskabet indeholder en lang række udfordringer, er komplekst og har mange interessenter. Forskere der vil dele deres data og deres fortolkninger af disse; professionelle dataudbydere med et væld af licensbelagte services; software der tilbyder dataanalyse; behandlingstjenester; finansieringskilder både private og offentlige, med en stigende interesse i ordentlig og rettidig Data Stewardship og et Data Science fællesskab, der miner data, integrerer og analyserer outputtet i håbet om at besvare både store og små spørgsmål. Interaktionen og samarbejdet mellem disse aktører og processer er en af de største udfordringer for Open Science og eScience.

I januar 2014 mødtes en lang række aktører efter ønske fra Netherlands eScience Center og Dutch Techcentre for Life Sciences med det formål at debattere, hvordan ovenstående problemstillinger kunne løses.

Resultatet af debatten blev, at der opstod konsensus om, at selvom definition og support af et minimalt sæt af fællesskabsbestemte vejledende principper og praksisser ville resultere i umiddelbare forbedringer, var der brug for grundlæggende principper, der omfattede en bred vifte af integrerende og eksplorative typer adfærd, imens de samtidig understøttede en bred vifte af teknologibeslutninger og implementeringer.

FAIR principperne udspringer af dette ønske. I 2016 udkom artiklen 'The FAIR Guiding Principles for scientific data management and stewardship i Nature-tidsskriftet Scientific Data, med hvilket FAIR-konceptet for alvor blev søsat. På grundlag af dette har både forskningsinstitutioner og finansieringskilder styrket deres krav i forhold til data management og muligt genbrug af forskningsdata. I Europa-kommissionens Open Research Data Pilot,

blev FAIR principperne implementeret for at opfordre forskere med bevillinger fra kommissionen til at sikre, at deres data var håndteret via god data management og efterfølgende delt på forsvarlig vis.

Med enkelte undtagelser betød det, at alle projekter som var dækket af Work Programme 2017 som udgangspunkt skulle følge retningslinjerne i Open Research Data Pilot (ORDP). Efterfølgende blev piloten også udvidet til at omfatte alle områder i forhold til Horizon 2020, der dermed fordrer forskningsprojekterne til aktivt at inkorporere FAIR principperne i praksis. Dette anses som en nødvendig forudsætning for realiseringen af European Open Science Cloud (EOSC) og Den Europæiske Kommissions vision for Open Science, der i FP9 Horizon Europe (Framework Programme) har en forslået bevilling på mere end 25 milliarder euro.

Lad os få definitionerne på plads. Bare ganske kort

Det følgende er ikke den fulde beskrivelse af FAIR principperne. I den sammenhæng er artiklen i Nature eller FORCE11-websiden om FAIR et meget bedre udgangspunkt. Det følgende er i stedet en kort beskrivelse fra et simpelt synspunkt, som forhåbentlig igennem en mere jordnær tilgang vil hjælpe med forståelsen af FAIR.

Med findable menes, at data og supplerende materiale har tilstrækkelig metadata og en vedvarende identifikation. Den vedvarende identifikation (PID – Persistent Identifier) kan eksempelvis være en DOI. PID sikrer, at andre kan finde, citere og spore data. Derudover skal hvert datasæt grundigt beskrives.
Udover proveniens/tilblivelse skal metadata også indeholde informationer om licens for brug (hænger sammen med reusable princippet) og en beskrivelse af datas kontekst.

Data og metadata er læsbare og forståelige af både menneske og maskine. Derudover er data lagret i et betroet arkiv.

Accessible dækker over, at vi på ethvert tidspunkt, såfremt vi kender datasættets PID og lokation, som minimum er i stand til at læse datasættets metadata, samt hvad man skal gøre for at få den fulde adgang til datasættet. På den mere tekniske side, skal den anvendte protokol være åben, gratis og alment mulig at implementere (eksempelvis HTTP/IP, dvs. via Internettet). Til sidst skal den vedvarende identifikation og metadata altid være tilgængelige, også selvom data på et senere tidspunkt enten bliver begrænset i adgang eller slettet.

Interoperable betyder i grove træk, at metadata er formuleret i et formelt, tilgængeligt sprog som benytter vedtagne vokabularier, ontologier og thesauri som er alment tilgængelige.
I en teknisk sammenhæng skal hvert computersystem, der ønsker at udveksle data kunne forstå den syntaks samarbejdspartneren anvender. Som bruger af et system, vil man ikke umiddelbart skulle bekymre sig om dette. Såfremt data bygger på et andet datasæt, og dette datasæt er nødvendigt for at få et komplet datagrundlag – skal metadata beriges med korrekt citation og PID.

Til sidst dækker reusable over princippet om, at metadata er rigt beskrevet med relevante attributter. Data har derudover en gennemskuelig slutbrugerlicens, og igen er der fokus på proveniens.
Detaljeret information er nødvendig for brugbarheden af data, specielt i vurderingen af hvor valide data er.

Maskinlæsbarhed er et begreb som gennemsyrer FAIR principperne. Maskinlæsbarhed er gældende for alle metadata ud fra det synspunkt at størstedelen af videns-opsamlingen i den moderne verden sker gennem udvekslingen af digitale klienter og protokoller. Den grundlæggende tese er derfor, at jo mere maskinlæsbart metadata er, jo mere sandsynligt er det at websøgning i eksempelvis Google vil indeholde en henvisning til datasættet.

Hvordan passer FAIR ind i forhold til Open Data?

Selvom Open Data og FAIR data er forskellige, kan de som i eksemplet ORDP både understøtte og også overlappe hinanden. Hvor Open Data er data, som kan anvendes og deles frit til et hvert formål betyder adgangsdelen af FAIR ikke nødvendigvis fuldstændig automatisk åben adgang.

FAIR principperne tillader, at der kan opsættes begrænsninger i forhold til sensitive persondata.
Europa kommissionen skriver i deres ORDP Guidelines, at adgangen skal være 'As open as possible, as closed as necessary. I FAIR betyder Accessibility "how-to-access", det vil sige en fuldstændigt klarlagt beskrivelse af hvilke retningslinjer, der gør sig gældende, defineret i et format læsbart for menneske og maskine.

Hvad så nu?

Skiftet til FAIR data er ikke en banalitet, og er ikke et mål i sig selv. Der er i stedet tale om en løbende proces. Ud over efterlevelse af FAIR principperne, skal der foretages et paradigmeskift, som blandt andet består af Data Management Planning (styring), tilskyndelse/belønning og kulturelle ændringer. Ændringerne skal komme indefra, de skal ikke dikteres men efterleves på grund af et ønske om større synlighed for ens forskningsdata både for menneske og maskine.

Universitetsbiblioteker og dets ansatte kan spille en stor rolle i arbejdet med FAIR princippernes integration i forskningsmiljøerne. Udover at promovere

FAIR principperne til forskerne, skal vi også bestræbe os på at indflette principperne i vores egne digitale bevaringspraksisser og politikker.
Vi skal ikke kun anbefale FAIR, men være foregangsmænd og kvinder, der aktivt søger muligheden for at deltage i forskningsfællesskaber, hvor vi får muligheder for at kuratere, berige og bevare forskningsdata efter FAIR principperne. Vi skal dog handle, inden det er for sent. Allerede nu er for-profit aktører som eksempelvis Springer Nature ude med tilbud til forskere om kuratering af forskningsdata i en simpel drop-løsning, hvor forskeren kan få kurateret et datasæt på op til 50 gigabyte data til en pris på ca. 2240,- dkk. Prisen er umiddelbar overkommelig, men ser man det som endnu et niveau i prisraketten for forskningspublikationer, er det en udvikling, vi bør overveje, om vi kan acceptere. Heldigvis står vi ikke alene med ønsket om at udbrede forskningsdata og FAIR.

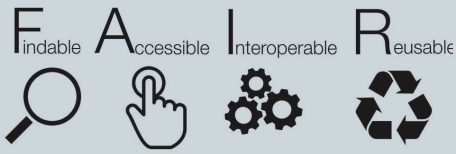
På områderne Findable og Accessible er der flere løsninger både kommercielle og ikke kommercielle. Specielt er værd at nævne to store aktører: Figshare (kommerciel), som er under implementering på DTU og The Dataverse Project (Open Source), som KU har i kikkerten (på SDU arbejders der på et eget udviklet system til forskningsdata og tilhørende metadata).

Systemerne sikrer, at deponeret forskningsdata er søgbart og tilgængeligt for discoverysystemer gennem deres grænseflader, samt at adgangen til data styres efter adgangskriterier fastsat af forskerne bag data. Om løsningen med et kommercielt system kontra et Open Source er bedre eller værre, er uden for denne artikel at bedømme. I stedet kan vi glæde os over, at der udvikles værktøjer, som sikrer større udbredelse og adgang til forskningsdata. Interessen er i hvert fald ikke faldende, hvilket også ses i at Google den 5. september 2018 annoncerede Dataset Search som tilføjelse til deres portefølje.

Områderne Interoperable og Reusable står muligvis lidt i skyggen af de to andre, men her findes også aktører som arbejder på at fremme principperne. Frictionless Data er et projekt under Open Knowledge International med fokus på at fjerne friktion fra data. Dette opnår projektet med udviklingen af specifikke værktøjer, specifikationer og best practices for hvordan data beskrives, publiceres og valideres. Specielt for dem er arbejdet med deres Data Package, et container-format for data baseret på eksisterende praksisser indenfor Open Source softwarepublicering. Data package værktøjet er specielt godt i forhold til Reusable princippet, da det indeholder krav om angivelse af licens for brug af datapakken samt versionering.

Afslutningsvis kan nævnes de forskere, som er modvillige og ikke ønsker at dele deres forskningsdata med deres kollegaer og omverdenen med begrundelsen, at de ikke vil give deres forskningsdata væk. På nuværende tidspunkt er der ikke et system til meritering af en åben forskningsdatapolitik.

Bidragsydernes krav presser dog i den rigtige retning. Det er dog utænkeligt at alle vil blive omvendt fra den protektionistiske holdning. I den sammenhæng må vi trække på skuldrene, og affinde os med at disse forskeres data vil forsvinde og gå i glemmebogen på samme måde som Pagtens Ark i Indiana Jones filmen forsvinder i et sort hul symboliseret af et stort lagerhotel, hvor intet er Findable, Accessible, Interoperable, Reusable.



Og referencerne...

http://www.datafairport.org/

Mark D. Wilkinson et al., (2016). The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data
https://ec.europa.eu/commission/sites/beta-political/files/budget-may2018-research-innovation_en.pdf

https://www.force11.org/fairprinciples
European Commission (2017), Guidelines on FAIR Data Management in Horizon 2020

https://www.springernature.com/de/authors/research-data-policy/pricing/15499842

https://www.blog.google/products/search/making-it-easier-discover-datasets

https://frictionlessdata.io/