FAIRifyHUM Leverance 2: Casebeskrivelser fra forskergrupperne

11/2 2020, Lene Offersgaard, KU

Opsummering

Målet med leverancen var gennem samarbejde med konkrete forskergrupper at afdække, hvordan de gør deres forskningsdata mere FAIR[1] og at høste erfaring med hvad der særligt udfordrer i forbindelse med at gøre forskningsdata mere FAIR. Der har været dialog og vejledning om FAIR til alle forskergrupperne.

Der har været kontakt til en større gruppe forskergrupper i Danmark med nuværende relationer til DARIAH og CLARIN, og grupper uden relationer til de to ESFRI'er. Syv forskere/forskergrupper har arbejdet med FAIRificering og deres casebeskrivelser er vedlagt som bilag. I case-beskrivelserne kan man læse, hvilke data de har arbejdet med, hvilke overvejelser de har gjort sig og hvad de finder er de største udfordringer for at gøre forskningsdata FAIR.

Herunder opsummeres erfaringerne fra de syv cases sådan som forskerne udtrykker dem i casebeskrivelserne, hvorefter der gives en afsluttende kommentar.

Erfaringer fra case 1: De vestindiske arkiver på tværs af grænser

Fokus var på at analysere data og på at gøre data mere *findable* og *reuseable*. Der er bl.a. brug for at tilføje metadata på engelsk for at gøre data søgbare og tilgængelige for brugere i Virgin Islands. En anden væsentlig udfordring er etiske spørgsmål der kan stilles, hvis man deler fotos taget af danskere af den afrocaribiske befolkning uden samtykke og i mange forskellige situationer, hvoraf nogle er i sårbare situationer. Det bør overvejes at dele datasamlingerne op, så man begrænser adgang til hvad der skønnes at være særligt følsomt materiale. Der er også brug for håndtering af copyright, da en del af materialet er beskyttet af copyright. Endelig vil det være oplagt at tilføje PID's til materialet, da de nuværende url'er ikke kan anses som persistente.

Erfaringer fra case 2: Uddannelsesdata

Fokus var på kvantitative data fra en ekstern partner og et datasæt bestående af selvgenerede kvalitative data fra interviews. De største udfordringer var spørgsmål om IPR og GDPR-forhold. Desuden er der behov for standardisering af terminologi inden for feltet. Dialogen med dataleverandøren blev igangsat men det er ikke lykkedes at komme til konsensus om eventuelt et delmængde af data kunne licenseres med henblik på deling og genbrug af data. For de indsamlede interviewdata og de tilhørende etnografiske observationsnoter, så er disse delvist publiceret I artikler og delvist svære at dokumentere konteksten for disse noter tilstrækkeligt til at andre kan genbruge data.

Erfaringer fra case 3: Fødevarer og sundhed

Fokus var her på proof-of-concept om data om forbrugeradfærd. De største udfordringer var de tekniske udfordringer med adgang til de to datasæt. Også spørgsmålene om studenters adgang til data og GDPR-forhold generelt har givet udfordringer, samt manglende konkrete værdier i en af databaserne og håndtering af NDA'er.

[1] https://www.go-fair.org/countries/guidelines-involvement-go-fair-initiative-country-level/

Der er brug for at kunne etablere et træningsdatasæt som studerende og unge forskere kan bruge til at få rutine i at bruge denne type af data. Særligt er der et stort ønske om at definere en model for FAIR udveksling af data, når der er flere aktører hvor alles IPR skal afklares og tilgodeses med en fair tilgang til brugerrettigheder og adgange til data.

Erfaringer fra case 4: Læsbarhed

Den største udfordring for at gøre data FAIR var copyright. Begge datasæt er belagt med copyright, men det ene datasæt har en licens, der tillader genbrug og videredistribution, mens det andet datasæt ikke foreløbig har kunnet deles da videredistribution ikke er tilladt. Hele opgaven med at klarlægge licensforhold, fortolke licensdokumenter og indgå i forhandlinger viste sig at være en krævende opgave. Der er brug for ekspertise vedrørende copyright og gerne midler til at frikøbe copyright-belagte data.

Det har været svært at afdække hvad værdien af at stille data til rådighed er. Ikke blot skal man vælge hvilke data der skal deles, men også afklaring af hvilke data man har lov til at dele under hvilke vilkår er svært. Også for opsummeringer og aggregerede datasæt kan det være svært at afgøre om nogle af teksterne kan føres tilbage til en oprindelig deltager i eksperimenterne. Derfor efterspørges guidelines og beskrivelser af *good practice* vedrørende deling af tekstdata.

En tredje erfaring er at i det omfang at man ikke kan dele sine forskningsdatasæt der er baseret på udvalgte dele af åbne datasæt, så tilskyndes forskerne til hver især at opbygge deres egne uddrag af tilgængelige data og dermed udfordres sammenligneligheden af resultater og på den måde kan troværdigheden af forskningsresultaterne måske anfægtes. Det er derfor vigtigt at det i fremtiden bliver lettere at indgå aftaler med teksternes rettighedshavere om tilladelser til videredistribution af forskningsdatasæt, så sammenligning af metoder og resultater på denne måde faciliteres, med klare referencer til rettighedshavere.

Erfaringer fra case 5: Keywords

Den største udfordring for at gøre data FAIR er også i dette tilfælde copyright-begrænsninger. Forhandlinger om tilladelser til brug af datasæt er komplicerede for den enkelte forsker. Selvom dataejeren i princippet tillader en vis grad af genbrug af data, så kan det være en langsommelig proces, særlig hvis dataejeren ikke er hjemhørende i EU og hvis også eksisterende kontraktforhold skal analyseres. Der er brug for ekspertise vedrørende copyright og forhandling af vilkår for genbrug og tilgængeliggørelse af data. Det manuelt skabte datasæt kan stilles til rådighed for andre, men datamængden er her så lille at mulighed for andres forskeres genbrug af data synes minimal.

Erfaringer fra case 6: Trusselsbreve

Den største udfordring er at det kun er en lille del af det samlede forskningsmateriale, der vil kunne deles fordi materialet, der indeholder truslerne er yderst følsomt materiale. For en del af materialet har det været muligt at indgå en aftale med et forlag om at tilgængeliggøre teksterne for forskere, men i den slags forhandlinger ville eksperthjælp være yderst nyttig for forskeren. For det datasæt der kan deles med andre forskere, er det ønskeligt at dele data, ikke bare hvor datasættet har en PID og kan downloades, men også via en applikation hvor forskerne kan se både de scannede trusselsbreve og søge i en version af teksterne i trusselsbrevene, med avanceret annotering. At gøre et datasæt *findable* og *accessible* med en PID og høstbare metadata, vil ikke altid være den bedste måde for forskerne at få adgang til forskningsmaterialet. Endelig er det en udfordring at få dækket omkostningerne ved metadatering og annotering af teksterne, sådan at teksterne både er veldokumenterede og overholder udbredte standarder og formater. Men en sådan indsats vil være

nødvendig for at andre forskere reelt kan genbruge teksterne. Datasættet har følgende PID: http://hdl.handle.net/20.500.12115/40

Erfaringer fra case 7: Folkeviser

En stor del af de danske folkeviser har allerede været tilgængelige på nettet i en del år efter de tidligere blev digitaliserede og manuelt annoteret i DUDS-projektet https://duds.nordisk.ku.dk/. Teksterne var også allerede opmærket i XML. Billes Visebog blev således nemt FAIRificeret ved at pakke xml-teksterne sammen til et datasæt og tilføje metadata. I dette tilfælde er der pga. den oprindelige udgivelsens alder ingen restriktioner fra udgiveren. Billes Visebog er således nu tilgængeliggjort for akademisk brug på følgende PID: http://hdl.handle.net/20.500.12115/41

Udvalgte forskergrupper

ESFRI kontakt	Forskningsdata	Forskere og institutioner	Beskrivelse
DARIAH	De vestindiske arkiver på tværs af grænser	Mette Kia Krabbe, KB og Marianne Ping Huang, AU	Materials related to the former Danish West Indies colony: describe or make available material and informations in order for other researchers to be able to investigate 1) which kinds of materials were available in the collections of The Danish Royal Library 2) what where their origins (creator) 3) what use had they been made of 4) which role had they played in and were continuously playing in the conception of the former Danish colony and in the description of Denmark as a colonial power more broadly.
None	Uddannelsesdata	Benjamin Brink Allsopp, AAU, Anders Tamborg (tidligere AAU, nu KU), Karsten Kryger Hansen, AAU	Data generated by teachers tagging lesson plans with competencies in a learning platform. The dataset focused on quantitative data from an external party, and self-generated qualitative data from interviews.
None	Fødevarer og sundhed	Bent Egberg Mikkelsen og Karsten Kryger Hansen, AAU.	Consumer behavior data on food shopping "Donate Your Food Data Lab". Proof of concept for a research infrastructure handling consumers on a voluntary basis could providing data about their shopping.
CLARIN	Læsbarhed	Frans van der Sluis, KU	Releasing datasets for training and testing textual readability and complexity models in a FAIR manner.
CLARIN	Keywords	Haakon Lund, KU	Wind Energy Journal keywords – challenges in FAIRifying. Using vocabularies or taxonomies for the wind energy sector.
CLARIN	Trusselsbreve	Tanya Karoli Christensen, KU	Tilgængeliggørelse af 120 trusselsbreve: fokus på *findable*, *accessible*, metadatering samt øvelse med interoperabilitet.

CLARIN	Folkeviser	Dorthe Dunker og Hanne Ruus, KU	FAIRIficering af folkevisebog: Jens Billes håndskrift (1555-1559). En stor del af de gamle folkeviser blev skrevet ned i 1550'erne.

Afsluttende kommentarer

Der er i en lang række tilfælde etiske udfordringer ved at gøre humanistiske forskningsdata FAIR, nogle data vil pga. GDPR og deres følsomhed ikke kunne deles. I en del situationer vil end ikke metadata kunne deles, da de også kan indeholde sensitiv information.

I de fem af de syv cases har vilkår for deling af data været en stor udfordring. Der er brug for vejledning om håndtering af copyright, forståelse af gældende kontrakter og licensforhold for data. Der kan også være brug for midler til at frikøbe copyright-belagte data. I fremtiden bør man være mere opmærksom på at afklare rettigheder til data - og muligheder for at dele data - allerede ved projekternes start. Også for opsummerede og aggregerede datasæt kan det være svært at afgøre mulighederne for at dele data. Derfor efterspørges guidelines og beskrivelser af *good practice* vedrørende deling af tekstdata, men det vil også kunne være relevant for andre typer af data. Det er også påpeget at der er brug for vejledning om forståelse af licensforhold og tildeling af licenser til datasæt.

Det generelle princip om at data gøres *findable* - med at de tilknyttes en PID[2] - har kun i begrænset omfang været muligt inden for projektets løbetid. I to cases er data blevet gjort FAIR ved at dele dem via CLARIN's dataarkiv i Danmark - clarin.dk. I tre tilfælde afventer deling af data stadig forhandling med rettighedshaverne. I et tilfælde ligger data åbent tilgængelige på url'er, men disse er pt ikke omsat til PID'er.

At gøre et datasæt *findable* og *accessible* - bl.a. med en PID og høstbare metadata - vil ikke altid være den bedste måde for forskerne at få adgang til forskningsmaterialet. Forskerne har ofte også brug for søgemuligheder ind i datasættene for at forskerne kan få et indtryk af materialet, og vurdere dets relevans. Dette kræver dog i alle tilfælde en anden service end der umiddelbart bliver tilgængeligt når data gøres FAIR. Dette aspekt er vigtigt at bemærke i alle bestræbelserne på at FAIRificere forskningsdata.

For at det giver mening for andre forskere at drage nytte af data, så skal data ofte tilføjes metadata, der er mere detaljerede, end hvis man blot selv skal håndtere data. Ofte vil det foretrækkes at metadata benytter kendte standarder, at beskrivelser tilføjes på engelsk, og at datas metadata dermed ikke kun er udtrykt på dansk. Men denne udvidelse af metadata og den dokumentation af data, der ligger i at beskrive data tager tid, og det er nødvendigt at anerkende at opgaven kræver resurser.

På nuværende tidspunkt har vi gjort den observation at det er svært at afdække hvilken merværdi, der opnås ved den ekstra indsats, som det kræver at stille data til rådighed.

Opmærksomheden på om studerende kan få adgang til data - og under hvilke begrænsninger - kan med fordel tages med i overvejelserne vedrørende adgang til data.

[2] PID: Persistent Identifier, fx. DOI.

Ligeledes er der et ønske om at definere en model for FAIR udveksling af data, når der er flere aktører, hvor alle parters IPR skal afklares og tilgodeses med en fair tilgang til brugerrettigheder og adgange til data.

Endnu en erfaring er, at hvis man ikke kan dele sine forskningsdatasæt, der er baseret på udvalgte dele af åbne datasæt, så tilskyndes forskerne til hver især at opbygge deres egne uddrag af de tilgængelige data. Dermed udfordres sammenligneligheden af resultaterne og på den måde kan troværdigheden af forskningsresultater måske anfægtes – alene fordi der er uklarheder om præcist hvilke data, der er grundlaget i det enkelte tilfælde. Det er derfor vigtigt at det i fremtiden bliver lettere at indgå aftaler med teksternes rettighedshavere om tilladelser til videredistribution af forskningsdatasæt til forskningsformål, så sammenligning af metoder og resultater på denne måde også gøres lettere.