DIGITALBEVARING.DK
- om digitalisering og digital bevaring

Forskningsdata

Artikler fra kategorien “Forskningsdata”, digitalbevaring.dk

Arkiv, artikler opdateres ikke længere

Det Kgl. Bibliotek og Rigsarkivet

Marts 2021

Indholdsfortegnelse

Bevaring af forskningsdata 1
Hvad er forskningsdata? 3
Særligt om problemer med at bevare forskningsdata på langt sigt 4
Databevaring ved Rigsarkivet 6
Fordele for forskerne ved FAIR bevaring af data 8
Tre forskere om databevaring 9
FAIR-principperne og langtidsbevaring 10
De 15 principper for FAIR data 11
Hvordan kædes FAIR-principperne sammen med langtidsbevaring? 14
Hvordan forberedes databevaring? 16
Om forskningsdata og forældelse 17
Fem gode råd om sikring af data 20
Beskrivelse af data - metadata 22
Forberedelse til langtidsbevaring 24
Fagspecifikke forskningsdata 26
Repositorier 28

Bevaring af forskningsdata

Henrik Vetter og Filip Kruse, Det Kgl. Bibliotek, december 2019

Forskningsdata kan bevares, enten fordi de har stor genanvendelsesværdi, også på langt sigt, eller fordi den enkelte forsker har et ønske om at bevare data. Data med stor genanvendelsesværdi bevares af Rigsarkivet, men i en række situationer er den enkelte forsker eller forskningsinstitution selv ansvarlig for bevaring. Denne artikel samt undersiderne giver en introduktion til emner inden for feltet "langtidsbevaring af forskningsdata".

Databevaring indeholder dels forberedelse af selve bevaringen og dels valg af egnet sted eller facilitet hertil. Hensigten med denne introduktion til emnet "langtidsbevaring af forskningsdata" er dels at gøre opmærksom på værdien af at bevare data, og dels give en oversigt over forberedelserne til at kunne gemme data og de konkrete muligheder for langtidstidsbevaring. Introduktionen er henvendt til de forskere, der selv ønsker at tage initiativ til at sikre, at data kan bevares sikkert og på en måde, som giver god mulighed for genfinding – og hermed genbrug – af data.

Hvorfor bevare forskningsdata?

I forbindelse med publicering af videnskabelige arbejder er det i dag for en række fags vedkommende et standardkrav, at forfatteren deponerer data i det tidsskrift, hvor arbejdet publiceres, eller i et anerkendt repositorie.

Det er også velkendt, at fondsfinansiering i en del tilfælde indeholder tilsvarende krav om bevaring af data. Her skal data forstås som digitale data og bevaring tilsvarende digital. Mere bredt er "research data management" i stadig stigende grad en væsentlig del af akademisk arbejde. En introduktion til digital bevaring kan findes her.

Ud over de direkte krav om databevaring kan der være en bredere samfundsmæssig interesse i at gemme og bevare forskningsdata. Således skal statslige forskningsdata – for praktiske formål data fra forskning udført ved universiteter og andre offentligt finansierede forskningsinstitutioner – anmeldes til Rigsarkivet. En del af de anmeldte data bevares derefter af Rigsarkivet.

Det juridiske grundlag for bevaring i Rigsarkivet består af:

- Arkivloven
- Arkivbekendtgørelsen
- Bekendtgørelse om anmeldelse af forskningsdata
- Bekendtgørelse om arkiveringsversioner

Hvem bevarer data – og hvilke?

Rigsarkivet bevarer de forskningsdata, som Rigsarkivet i dialog med forskerne vurderer har stor genanvendelsesværdi, både på kort og på langt sigt. Fra et forskningsperspektiv kan der være ønske om at andre og måske flere forskningsdata langtidsbevares. Hertil kommer, at der kan være forskellige syn på afgrænsningen af forskningsdata. Er der eksempelvis tale om at bevare rådata, dokumentere forskningsprocessen eller anvendte metoder og lignende, er det selvklart relevant at gemme mere end blot de endelige data, som er resultatet af forskningen.

Hvad er forskningsdata?

Henrik Vetter og Filip Kruse, Det Kgl. Bibliotek, december 2019

Forskningsdata skal forstås bredt som data, men også beskrivelser af tilblivelsen af data. Der er altså tale om dokumentation for, hvordan data er kommet til verden. Genfinding sikres gennem metadata.

Til grund for forskningsprocessen ligger materiale, der bearbejdes og analyseres videnskabeligt. I The Danish Code of Conduct for Research Integrity skelnes mellem primært materiale og data:

”Primary material is any material (e.g. biological material, notes, interviews, texts and literature, digital raw data, recordings, etc.) that forms the basis of the research. Data are detailed records of the primary materials that comprise the basis for the analysis that generates the results.”

Definitionen på data er ”records”, altså oplysninger, dokumenter, fortegnelser o.l. over de materialer, der danner grundlaget for forskningsresultaterne, herunder forskningsdata. Vi har altså et kontinuum, hvor yderpunkterne er henholdsvis forskningens slutresultater i form af data – og en række kilder til og oplysninger om disse slutresultater. De kan være baggrund, i form af tidligere forskning og data, den faglige eller samfundsmæssige kontekst for det pågældende forskningsprojekt, en konkret anledning til at udføre dette, fx en bestemt begivenhed. Sammen med dette hører også anvendte metoder til dataindsamling og analyse, teoretisk udgangspunkt osv. Kort sagt dokumentation, der giver indsigt i de nærmere omstændigheder for hvordan og hvorfor data er frembragt.

Genfinding af data

En vigtig del af forskningsprocessen er fremskaffelse af viden om hidtidig forskning inden for området. Heri indgår tidligere indsamlede og analyserede forskningsdata.

Genfinding af data foretages hyppigst ved hjælp af metadata, som er data om data. Bibliotekers informationer om en bog i et katalog er data om pågældende bog, som kan sikre, at den kan findes, og at låneren kan finde andre bøger, om samme emne, af samme forfatter osv. De biblioteksmæssige metadata ses i den bibliotekspost, der vedrører et enkelt materiale. Her angiver biblioteker proveniens, såsom trykkested, år, udgiver, forfatter osv. Der er ofte en emneklassifikation og emneord, der placerer materialet i en større forståelsesmæssig kontekst. Billedlig talt kan det sammenlignes med en placering på en hylde, hvor det enkelte materiale sættes sammen med andet materiale, der er emnemæssigt sammenfaldende, minder om eller på anden vis er relateret til.

Særligt om problemer med at bevare forskningsdata på langt sigt

Henrik Vetter og Filip Kruse, Det Kgl. Bibliotek, december 2019

Langtidsbevaring kræver opmærksomhed på formater og metadata. Valg af format har betydning for muligheden for at sikre fremtidig adgang. Metadatering er afgørende for at bevarede data kan genfindes og dermed genbruges. Man skal i forbindelse med metadatering være opmærksom på, at beskrivende metadata kan forældes.

Formater og genfinding

Når forskningsdata (og for den sags skyld andre data) skal kunne genfindes og dermed genbruges på langt sigt, er der nogle særlige problemer – ud over de mere tekniske – at være opmærksom på.

Der er særlig grund til at være opmærksom på formater. Rigsarkivet modtager data i et systemuafhængigt databaseformat, som fremgår af bekendtgørelse om arkiveringsversioner, der også definerer et særligt afleveringsformat for data skabt i de mest gængse statistikprogrammer som SAS, Stata og SPSS. Hvis der er tale om video, lyd og billeder, skal data migreres til JPEG-2000, TIFF, MP3 og/eller WAVE eller MPEG. Man må forvente, at disse formater giver tilpas sikkerhed for tilgængelighed i forhold til langtidsbevaring. Derfor er det en del af forberedelsen til langtidsbevaring at sikre et relevant format, evt. gennem migrering. Der henvises til afsnittet om forskningsdata og forældelse for en yderligere beskrivelse.

Metadata og genfinding

Genbrug af data forudsætter selvsagt, at data kan genfindes. Men data om data, metadata, kan ændre eller eventuelt helt miste betydning over tid. Eksempelvis finder Zavalina og Zavalin (2018), der analyserer 400.000 autoritetsdata – biblioteksdata om emne, forfatter, titel osv. i vedtagne og anerkendte termer – at anvendelsen ændres inden for en periode på bare 22 måneder, hvilket efter forfatternes opfattelse påvirker funktionaliteten i forhold til brug. Metadata, der beskriver forskningsdata, kan være yderligere komplekse, og man må forvente, at de problemer, som Zavalina og Zavalin (2018) beskriver, ikke bliver mindre, når mængden af metadata vokser.

Forældelse af metadata

Der er mindst ét yderligere problem i forhold genfinding af data. De beskrivende metadata fx emneord vil naturligt forældes. Der kan endda være tilfældet, at en hel forskningsretning forsvinder. For eksempel blev alkymi i ældre tid set som en videnskab. Det gør alkymi ikke længere. Men metadata, der engang ville have kunnet beskrive dennes mulige data, fx emneord om anvendelsen af salamandre, vil ikke længere kunne være indgang til genfinding. Der kan også være tale om at nogle funktioner helt forsvinder, hvorfor emnebeskrivelse af data gennem funktionen bliver intetsigende.

Tag som et eksempel det danske ord ”drager”, der ligesom det engelske ord ”porter,” refererer til en person, der (traditionelt ved jernbanestationer) er beskæftiget med at bære bagage. I takt med, at funktionen forsvinder, er det naturligt, at ordet forsvinder fra sædvanlig sprogbrug, hvorfor sandsynligheden for at finde emnedata beskrevet ved ordet ”drager” eller ”porter” falder.

Der er naturligvis ikke noget nyt i, at sprogbrug ændres over tid. I en del tilfælde klarer bibliotekers klassifikationssystemer i vidt omfang denne type problem, fordi man i klassifikationssystemer giver en ret præcis klassifikation af emne samt en angivelse af tidligere anvendte betegnelser og tidspunkt for revision. Tilsvarende gælder for emne-thesaurusser i artikelbaser. Når der i særlig grad kan opstå genfindingsproblemer i forhold til forskningsdata skyldes det, at emnebeskrivelsen ofte hviler på ophavspersonen, hvorfor emnebeskrivelsen kan blive noget tilfældig og derfor vanskelig at opdatere i forhold til ændringer af opfattelser, sprogbrug, etc.

Der er ikke nogen let standardløsning på hvad der udgør en god beskrivelse. Men der er måske en bedre chance for, at en mere overordnet emnebeskrivelse bevarer sin relevans i forhold til en mere snæver beskrivelse. Man kan også lade sig inspirere af emne-thesaurusser og eventuelt have en række sideordnede beskrivelser af emnet for eksempel ved brug af synonymer. Endelig kan man læne sig op ad FAIR-principperne for at sikre den bedst mulige chance for genfinding.

Læs mere

Zavalina, O.L. & Zavalin, V. (2018), Evaluation of Metadata Change in Authority Data over Time: An Effect of a Standard Evolution, paper, ASIS&T Annual Meeting
https://doi.org/10.1002/pra2.2018.14505501064

Databevaring ved Rigsarkivet

Mette Hald-Andersen, Rigsarkivet, maj 2020

Hvorfor skal nogle forskningsdata bevares på lang sigt?

Forskningsdata repræsenterer en stor værdi, da de dels kan bruges i samtiden til at verificere forskningsresultater og dels i fremtiden kan anvendes i ny forskning af forskellig art. Rigsarkivet har gennem flere år arbejdet med at definere hvilke unikke forskningsdata, som kan siges at have værdi for fremtiden, og som derfor skal bevares på struktureret vis og med en dokumentation, som sikrer genanvendelsesværdien – altså FAIR.

Det har vist sig, at det er nemmere at definere grupper af data, som man ikke behøver at langtidsbevare fremfor at lave en positivliste over data, som skal bevares. Når det er nødvendigt at vælge hvilke forskningsdata, der skal bevares, skyldes det, at der er væsentlige udgifter forbundet med at gøre data FAIR samt at sikre, at de forbliver FAIR over tid.

Hvilke forskningsdata bør bevares for fremtiden?

Det er nødvendigt at have et godt kendskab til indholdet af det enkelte datasæt for at kunne udpege hvilke data, der skal bevares for fremtiden, og altså til mere end verifikation af det enkelte forskningsprojekts resultater. Rigsarkivet beder derfor om en række oplysninger, når bevaringsværdien af datasættet skal vurderes, herunder forskerens egen vurdering af bevaringsværdien.

Nogle datasæt skal dog ikke bevares for fremtiden. Det gælder for forskningsdatasæt som:

- alene baserer sig på udtræk fra administrative registre som fx Landspatientregistret. Det skyldes, at registrene allerede langtidsbevares i Rigsarkivet, så udtræk kan rekvireres herfra eller fra den styrelse, der opsamler data.
- er publiceret i en artikel el.lign. i sin helhed. Disse datasæt sikres for fremtiden via pligtafleveringsloven og skal ikke "dobbeltarkiveres".
- som er skabt gennem eksperimenter eller simulationer, der kan gentages. Data fra hyppigt gentagne eksperimenter og simulationer kan dokumenteres via forskningsrapporter o.lign., og viden om dem kan bevares på denne vis. De kan så om

nødvendigt gentages. Det gælder fx iterationer på økonomiske data, kemiske og elektrofysiske forsøg.
- er produceret ved forskning under ph.d.-niveau. Betragtningen her er, at betydningsfuld forskning under ph.d.-niveau i langt de fleste tilfælde vil føre til et ph.d.-projekt, hvorfra data så kan bevares.

Hvilke rammer gælder?

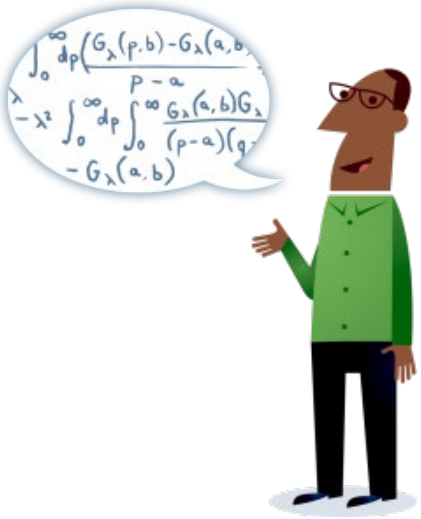
D. 1. maj 2020 trådte en ny bekendtgørelse i kraft for anmeldelse af digitale forskningsdata skabt af statslige myndigheder, hvor ovenstående principper indgår. Tilsvarende findes der bekendtgørelser, som fastsætter bestemmelser for forskningsdata produceret i regionalt eller kommunalt regi.

Find bekendtgørelserne for alle myndigheder på Rigsarkivets hjemmeside.

Fordele for forskerne ved FAIR bevaring af data

Gertrud Stougård Thomsen, AU Library, Det Kgl. Bibliotek, december 2019

En ting er, at der i stigende grad stilles krav til forskerne fra bevillingsgivere og forlag om opbevaring af data. Men hvad er gevinsten for den enkelte forsker?



Undersøgelser har vist, at god praksis for opbevaring af forskningsdata har en række fordele:

Fordele for forskeren

- Større chance for at tiltrække nye forskningsmidler, når man kan demonstrere god data management praksis.
- Større troværdighed, når forskningsresultater kan reproduceres.
- Større synlighed og flere citationer, hvis data kan genfindes og genbruges.
- Sikring af data mod at gå tabt og mod forældelse.
- Bedre muligheder for selv at genbruge data i ny forskning.
- Åbner op for nye samarbejdsmuligheder med andre forskere eller erhvervsliv.
- Mulighed for nye medforfatterskaber.
- Mulighed for at licensiere data, evt. med en embargoperiode.
- Forskeren lever op til samfundets ønsker om at drage nytte af forskningsmidler.

Fordele for samfundet

- Deling og genbrug af ellers utilgængelige og uoverskueligt store datamængder kan lede til banebrydende ny forskning.
- Bedre udnyttelse af ressourcer, større videnskabelig produktivitet og effektivitet.
- Imødegåelse af reproducerbarhedskrisen. Øget transparens og forskningsintegritet.

Referencer

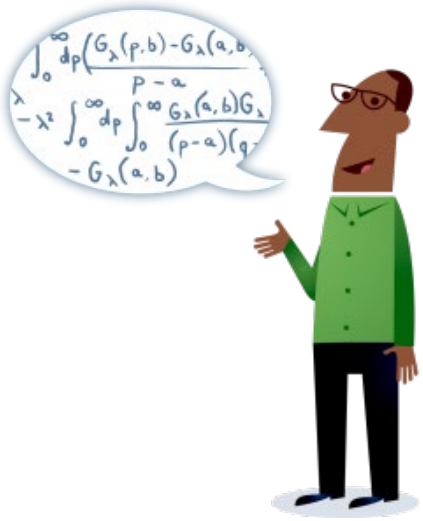
Benefits Summary for a Data Archive. Charles Beagrie Ltd and CESSDA 2017. Retrieved from http://dx.doi.org/10.18448/16.0010

Martone, M. E., Garcia-Castro, A., & VandenBos, G. R. (2018). Data sharing in psychology. American Psychologist, 73(2), 111–125. https://doi.org/10.1037/amp0000242

Tre forskere om databevaring

Asger Væring Larsen, Syddansk Universitetsbibliotek, december 2019

Det følgende afsnit er baseret på tre interviews med forskere fra Syddansk Universitet og Danmarks Tekniske Universitet.



Nikola Vasiljevic (Lektor ved Institut for Vindenergi, DTU)
Nikola oplever et skift i forskernes indstilling. Han har selv oplevet, at synlighed og antal citationer er øget, efter kun et års tilgængeliggørelse af data. Nikola har været med til at levere data til DTU's data repository, og disse data er efterfølgende blevet set omkring 500 gange. Instituttet har nu sikret sig interne midler til at gøre 16 store datasæt tilgængelige via en egen platform.

Nikola overvejer desuden at igangsætte et dataredningsprojekt, der skal sikre ældre datasæt. Der er ikke tale om store mængder, men om et par hundrede GB. Han arbejder på en strategi, der bl.a. skal definere procedurer: Hvilket lagringsmedie skal bruges, osv.

Instituttet har selv et system til aktive data, men er usikre på, hvad skal de stille op med deres data på lang sigt. De har data, der er underlagt restriktioner ifølge aftaler med private firmaer, men som jo også skal gemmes på lang sigt. Nikola Vasiljevic råder til at huske at få lavet aftaler med firmaerne, om at der eventuelt kan være en embargo på data – at de fx kan åbnes efter 10 år.

Anna Thit Johnsen (Lektor ved Institut for Psykologi, SDU)
"Databevaring bør tænkes ind fra starten. I nogle projekter giver det ikke mening at bevare data for eftertiden, men overvej det, og hvis data er velegnede, så planlæg med det fra begyndelsen. Dermed skal man ikke bruge mange resurser på at ordne data inden aflevering."

Jørgen T. Lauridsen (Professor ved Institut for Virksomhedsledelse og Økonomi, SDU)
"Forskere bør se sig selv som medborgere med et ansvar for at dele deres data og publikationer."

FAIR-principperne og langtidsbevaring

Lea Sztuk Haahr, Rigsarkivet, december 2019

FAIR principperne er virkelig et fænomen som har fået manges øjne op for, at data kan få et meget længere liv ved at kunne findes og anvendes, hvis man følger nogle enkle principper. FAIR principperne er en række anbefalinger, der især handler om at berige sine data med metadata. Jo mere forskere gør deres data FAIR, i jo højere grad bliver det muligt at genfinde og genanvende dem i nye sammenhænge, også på langt sigt.

FAIR-principperne blev første gang beskrevet i 2016 i denne artikel i Nature. Oprindeligt var FAIR principperne udviklet til at handle om det it-tekniske til støtte for øget anvendelse af forskningsdata, men principperne har fået en langt bredere appel.

FAIR dækker over 15 principper som er inddelt i fire følgende kategorier: Findable, Accessible, Interoperable, Reusable.

En af de mest hyppige fordomme om FAIR er, at FAIR = Open Data. Det er ikke korrekt. FAIR står for, at data er så åbne som muligt, men så lukkede (beskyttede) som nødvendigt. Der er dermed taget hensyn til GDPR, sensitive data, ophavsret mv.

Det er især vigtigt at kunne skelne mellem data og metadata i FAIR regi. Hvor data er det, der genereres i en dataindsamling i et forskningsprojekt, er metadata beskrivende data om data. Metadata er ikke følsomme data, idet de ikke indeholder rådata, men alene beskriver dem. Metadata kan beskrive indsamlingsmetode, tid og sted og hvor mange respondenter/forsøg, der er lavet. I artiklen og i FAIR-principperne bruges ofte ordet: (meta)data. Ordet dækker over både metadata og data, og synliggør vigtigheden af, at begge komponenter ofte er lige vigtige og skal prioriteres lige højt.

De 15 principper for FAIR data

Lea Sztuk Haahr, Rigsarkivet, december 2019

FAIR dækker over 15 principper som er inddelt i fire følgende kategorier: Findable, Accessible, Interoperable, Reusable.

Findable

For at data kan genbruges er første skridt på vejen, at det kan findes. FAIR arbejder med fire principper, som kan øge sandsynligheden for, at data kan findes.

F1. (Meta)data are assigned a globally unique and persistent identifier
En persistent identifier (PID) er en digital henvisning som aldrig vil forsvinde, og altid vil henvise til metadata og til datas placering. Vi kender alle situationen med døde links på internettet, men med en PID vil datareferencen altid henvise til den rette placering for data. Der findes forskellige udbydere af PID-services. I Danmark tilbyder DeiC DataCite Danmark – en service for tildeling af PID (i form af DOI, Digital Object Identifier).

F2. Data are described with rich metadata (defined by R1 below)
Grundtanken er, at data skal være mulige at finde og anvende. Derfor skal data være rigt beskrevet, så enhver anden kan bruge data uden yderligere forklaringer. Data uden metadata er i mange tilfælde kun brugbart for forskeren selv, og dette kan endda blive svært efter nogle år – simpelthen fordi man glemmer detaljerne, når man ikke arbejder med det.

F3. Metadata clearly and explicitly include the identifier of the data they describe
Hvis metadata er rigt beskrevne, som F2 forlanger, vil det være katastrofalt, hvis ikke man kan finde det tilhørende data bagefter. Derfor er det vigtigt, at metadata altid indeholder et objekt som kan henvise til data.

F4. (Meta)data are registered or indexed in a searchable resource
Hvis en forsker har opfyldt de tre foregående principper, men ikke har gjort (meta)data søgbart er der ingen der kan få glæde af data. Ingen kan vide, hvad der ligger på hans harddisk, og derfor vil arbejdet med metadata kun give mening for hans eget videre arbejde med data. Derfor skal metadata gøres tilgængelige som indekseret/søgbar ressource på internettet.

Accessible

Ovenstående muliggør at data er brugbart og kan findes. At få adgang til data skal have lige så stort fokus på brugervenlighed.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
A1 henviser til at når man har fundet linket til (meta)data, så skal linket fungere universelt, og ikke give yderligere problemer at få adgang til data. Er det sensitive data, er det skal adgangen til data kontrolleres på en veldokumenteret og overskuelig måde.

A1.1 The protocol is open, free, and universally implementable
Ingen betalingsmur skal forhindre adgangen til metadata. Nogle dataudbydere har en forretningsmæssig tilgang til data og ønsker derfor markedsvilkår. FAIR principperne derimod har fokus på, at alle brugere skal have adgang til metadata, så de kan afgøre, om data kan bruges igen til deres forskningsprojekt.

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
Som skrevet i indledningen er FAIR data ikke lig med open data. Hvis der er behov for at lave en autorisation af brugeren, er det i tråd med FAIR principperne. Proceduren skal blot være gennemskuelig og gerne maskinlæsbar.

A2. Metadata are accessible, even when the data are no longer available
Hvis et datasæt er blevet slettet, skal metadata stadig være tilgængelig. Hvis metadata er beskrevet i detaljer, vil de i sig selv være værdifulde ift. fx. at gentage studiet.

Interoperable
(Meta)data skal gerne være kompatibelt med andre (meta)data forstået på den måde, at en forsker på kort tid skal kunne sammenligne data med noget andet data og vurdere, om data er brugbart i hans arbejdsområde.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
Afhængig af forskningsområde er det forskellige ting der lægges vægt på i en dataindsamling. Dog er de fleste dataindsamlinger stadig overordnet set ens. Der beskrives bl.a. tid, sted, emnet/emnerne afgrænses og metoden beskrives. FAIR lægger vægt på, at metadata i detaljer beskriver, hvordan data er indsamlet, således at en anden forsker, også en uden for forskningsområdet, kan genbruge data. Dette kræver, at beskrivelserne er veludførte.

I2. (Meta)data use vocabularies that follow FAIR principles
Beskrivende metadata som tid, sted, emne mm. kan indgå som elementer i en Controlled Vocabulary, fordi det er metadata, som ideelt har en fast definition, så forskeren, bibliotekaren, arkivaren og efterfølgende forskere ved præcis, hvad elementet dækker over. En Controlled Vocabulary med definitioner skal have en PID (persistent identifier) som beskrevet tidligere, så listen altid kan findes. På den måde fremtidssikrer man data.

I3. (Meta)data include qualified references to other (meta)data
Hvis data bygger videre på eksisterende viden eller andre datasæt, bør dette refereres i (meta)data. Alt information om datasættet, der gør lettere for en anden bruger at anvende – eller blot for forskeren selv, bør inkluderes i metadata.

Reusable
Genbrug af data kan være svært, men behøver ikke at være det. Det vigtige er, at man ved præcis hvordan dataindsamlingen er gennemført, og det er detaljeret beskrevet.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
Her er vi nede på det mest detaljerede beskrivelsesniveau, som for forskeren selv kan være grundlæggende eller selvforklarende viden, men for en anden forsker kan være vitale oplysninger. Beskrivelse af indsamlingstidspunktet, hvilke metoder blev brugt fx software, kodeskemaer mm.

R1.1. (Meta)data are released with a clear and accessible data usage license
Data og metadata skal indeholde en licens, som beskriver adgang til og brug af data. Er data åbent og kan data blive delt med alle, eller kan man ansøge personligt om adgang? Som forsker bør du afklare med dig selv, hvilken licens du ønsker sat på dit data.

R1.2. (Meta)data are associated with detailed provenance
(Meta)data skal indeholde en beskrivelse af tilblivelsen af data, altså workflowet før, under og efter dataindsamlingen. Her gives en detaljeret beskrivelse af workflowet, som kan være vigtigt for at kunne genskabe eller genbruge data

R1.3. (Meta)data meet domain-relevant community standards
Hvis der findes en metadatastandard inden for forskningsområdet, som er anerkendt og brugt, bør metadata som minimum opfylde denne. På den måde sikrer man sig, at data fra samme forskningsområder er struktureret på en ensartet måde og metadata ligner hinanden, så det er nemt og overskueligt at sammenligne (meta)data.

Hvordan kædes FAIR-principperne sammen med langtidsbevaring?

Lea Sztuk Haahr, Rigsarkivet, december 2019

At langtidsbevare forskningsdata i henhold til FAIR-principperne er et stort og omfattende arbejde, som kræver løbende vedligeholdelse af både formater, standarder og metadata. Det kræver også meget ekspertviden som løbende skal udvikles, og holdes opdateret. Til gengæld vil man som forsker i fremtiden kunne finde og få adgang til data som bliver indleveret til arkivering i dag, og data vil kunne læses og forstås.

At langtidsbevare digitalt skabte data er en disciplin som bevaringsinstitutionerne i Danmark har beskæftiget sig med siden begyndelsen af 1970'erne.

Ønsker man at lave en FAIR langtidsbevaring, er der vigtige dimensioner at tage højde for. Man forpligter sig til en større og mere omfattende opgave. Hvis data kan findes (Findable) nytter det fx ikke, at det findes i formater, som ikke er brugbare længere, det klassiske eksempel er interviewudskrifter eller feltnoter gemt som WordPerfect filer. Ligeledes vil data som kan findes og åbnes i brugbare formater være nyttesløse, hvis ikke der medfølger metadata til at tolke data, som fx oversættelse af numeriske koder til forståelige kategorier, 1=nej, 2=ja.

Langtidsbevaring kræver, at forskerne og bevaringsinstitutionerne går meget dybt ind i metadataopbygning og forpligter sig på at vedligeholde samlingen af data, ikke kun for dataformater, men også for metadata.

En bevaringsinstitution skal, hvis den ønsker at lave FAIR langtidsbevaring, sikre kendskab til områdespecifikke metadatastandarder og have mulighed for at opdatere disse løbende. En opdatering er nødvendig, fordi forskningsmetoder, teknologier, sprog mv. udvikles over tid.

Mere viden om FAIR

Ønskes der mere viden om FAIR vil det anbefales at besøge følgende hjemmesider:

Go FAIR (https://www.go-fair.org/) – en bottom up organisation som har målet at øge kendskabet til og brugen af FAIR principperne.

FAIR projektet (https://vidensportal.deic.dk/da/FAIR) – et dansk projekt fra 2018 som kortlagde FAIR i dansk forskning, og som har mange praktiske input og forklaringer:

FAIRytale (https://zenodo.org/record/2248200#.XOUAcIem2Ul) – en leverance fra ovenstående projekt. Et eventyr om FAIR principperne, som tydeliggør og viser hvordan FAIR principperne altid giver en happy ending.

Research Data Alliance (RDA) – Global alliance (https://www.rd-alliance.org/), der bygger broer til støtte for øget anvendelse af forskningsdata. RDA har fokus på FAIR bl.a. i kraft af en arbejdsgruppe, der udvikler en FAIR Maturity Model. Pr. 1. april 2019 findes en dansk RDA node, der skal arbejder for, at dansk forskning får udbytte af RDA. FAIR langtidsbevaring kunne gøres til et fokuspunkt for nodens aktiviteter.

Hvordan forberedes databevaring?

Henrik Vetter og Filip Kruse, Det Kgl. Bibliotek, december 2019

Databevaring indeholder både forberedelse af data og beslutning om, hvor data og databeskrivelse (metadata) skal bevares. Grunden hertil er, at de forskellige repositorier har forskellige muligheder i forhold til fx formater. Dele af forberedelsen hænger således sammen med, hvor man vælger at bevare data, hvis de ikke bliver bevaret i Rigsarkivet.

Billedlig talt svarer bevaring til at pakke data i en kasse og stille den et sted. Kassen har en bestemt størrelse, den skal mærkes med indhold, altså emner, ophav, mv. Selve bevaringen er at placere kassen på et sikkert lager. Her skal kassen passe til hylderne i lageret, mærkningen af kassen skal være synlig i forhold til lagerets specifikke indretning, etc.

I afsnittene herunder er der forskellige råd til forberedelse, men de skal ses i sammenhæng med det endelige valg af repositorie og de krav, dette kan stille.

Om forskningsdata og forældelse

Henrik Vetter og Filip Kruse, Det Kgl. Bibliotek, december 2019

Er dine forskningsdata i fare for at blive forældede? Er dine forskningsdata gemt i åbne, frie og udbredte formater, der er bredt understøttede af både software og hardware? Bliver dine forskningsdata ikke gemt i Rigsarkivet? Er dine forskningsdata FAIR? Kan du svare nej til et eller flere af spørgsmålene, så læs endelig videre. Har du kun lidt tid, så overvej med disse fem tips, hvordan du bedre kan sikre dine forskningsdata og værne dem mod forældelse.

Vi tager det i dag ofte for givet, at data vi producerer er tilgængelige, gerne i skyen, og for altid. Også skønt historien er rig på eksempler på det modsatte og viser, hvordan det kan være tilfældigheder, som eksempelvis fundet af Rosettastenen, der har gjort det muligt for senere generationer at forstå gamle datakilder. Digitale data adskiller sig på det punkt ikke fra analoge kilder, og der er udfordringer, som kan føre til en øget risiko for informationstab over tid, hvis ikke de adresseres tidligt og løbende i bevaringsprocessen.

En sådan udfordring er risikoen for, at data forældes og ikke længere kan læses og forstås. I arkiveringsjargon kaldes dette for logisk forældelse. Det betyder, at data, hvor fejlfrit det end står skrevet med 0- og 1-taller, ikke længere er muligt logisk at læse og forstå, fordi senere generationer har mistet evnen til at læse de binære værdier på den rigtige måde. Dette problem gælder ikke mindst, og måske endda i højere grad, for forskningsdata, fordi forskningsdata i sin natur skabes i en kontekst, hvor udvikling og foranderlighed ikke alene er vilkår, men noget, der aktivt tilstræbes, og data ofte er knyttet til anvendelsen af en bestemt teknologi.

Her introducerer vi dig til emnet og giver en række gode råd til at reducere risici for forældelse af dine forskningsdata, så også senere generationer kan drage nytte af ældre viden.

Når vi i det følgende omtaler forskningsdata, menes data, som er skabt i forbindelse med forskning med anvendelse af en videnskabelig metode. Det kan derfor være data af vidt forskellig oprindelse, med forskellige datatyper og brug. Forskningsdata kan derfor i denne sammenhæng både være tekst, billeder, lyde, statistiske data, data fra spørgeskemaundersøgelser, medicinske billeddata, eksperimentelle måledata, m.fl.

Hvad er logisk forældelse?

For at forstå, hvad logisk forældelse er, starter vi ved begyndelsen af det, vi kunne kalde en fortolkningskæde. På det første niveau af denne kæde står alt data skrevet i en lang sekvens af binære tal, enten 0 eller 1. Denne sekvens af binære tal udgør en bitstrøm. Bitstrømmen er i sig

selv ikke forståelig for et menneske, førend vi anvender en metode til at dele bitstrømmen i bidder og etablere en ramme for, hvordan de opdelte binære tal kan oversættes til et tegn, vi som mennesker kan forstå, fx et bogstav.

Dette er, hvad næste trin i fortolkningskæden gør, nemlig opdelingen og afkodningen af den binære strøm til et tegnsæt. Et tegnsæt rummer de tal, tegn og bogstaver som vi mennesker anvender i vores alfabeter. Der findes mange sådanne "tegnsætsindkodere" eller "tekstkoder", f.eks. ASCII, UTF-8 og ISO 8858 til de fleste vestlige tegnsæt, samt tegnsætsindkodere til f.eks. kinesiske eller japanske skrifttegn. Fælles for alle tegnsæt er en forståelse af, at bitstrømmen skal læses og fortolkes i henhold til en standard. Sker det, og har vi adgang til det rigtige tegnsæt, kan vi med vores computere åbne og læse bitstrømmen som det, vi kender, som en tekstfil.

Hvis vores fortolkningskæde stoppede her, ville spørgsmålet om logisk forældelse være begrænset til en søgen efter det tegnsæt, som skal anvendes til at oversætte bitstrømmen til de korrekte tegn. Logisk forældelse ville i denne sammenhæng betyde, at vi ikke længere har adgang til den rigtige tegnsætsindkoder og derfor ikke kan få mening ud af den lange sekvens af binære tal.

Desværre, eller heldigvis, er data mere righoldige end simple tekstfiler og indgår i til stadighed mere komplekse strukturer. Det gør arbejdet med at forstå bitstrømmen sværere, men til gengæld kan vi glæde os over at kunne bearbejde komplekse data som at analysere 3-dimensionelle hjerneskanningsbilleder og arbejde med flerdimensionelle datasæt.

Dette fører os til næste trin i fortolkningskæden, nemlig organiseringen af bits (eller tegn) i strukturer, vi kan kalde for formater. Her ligger der igen en fælles forståelse til grund for, hvordan den binære data skal læses og forstås på en måde, så programmer, udover simple teksteditorer, kan læse og repræsentere data på en meningsskabende måde. Nu handler logisk forældelse om, hvordan formatet kan læses og forstås, så den information, bitstrømmen i sidste ende repræsenterer, kan vises korrekt og meningsfuldt for en bruger af programmet.

Logisk forældelse er altså udtryk for det at miste evnen til at kunne læse binær data på en meningsfuld måde, så data kan forstås og fortolkes. Hvad enten ens bevaringsstrategi baserer sig på en migrerings- eller emuleringstankegang, se fx bevaringsmetoder, skal formatet i begge tilfælde kunne læses og forstås, inden det forsøges bevaret på den ene eller anden måde.

Årsager til logisk forældelse

Der er flere grunde til logisk forældelse, fx:

- Et formats udbredelse er afgørende for dets overlevelse og fortsatte understøttelse. Hvis udbredelsen når under en kritisk størrelse, reduceres bevæggrundene for fortsat understøttelse. I kombination med hyppige udviklinger af mere moderne formater, kan et formats risiko for forældelse accelereres.
- Den hastighed, programmer og formater udvikles med for at tilbyde nye funktionaliteter til brugere, kan gøre det svært og bekosteligt at sikre bagudkompatibilitet med ældre versioner. I sidste ende er risikoen, at ældre data i ældre formater, ikke længere understøttes af tilgængelige programmer. For forskningsområder med stor udviklingshastighed, hvor analyseværktøjer og understøttende formater udvikles hyppigt, er risikoen større.
- At data er lagret i et lukket, proprietært format med utilgængelige eller utilstrækkelige beskrivelser af formatet. Proprietær betyder blot, at formatet er ejet af en privat

virksomhed (eller organisation) med de rettigheder og den beskyttelse det giver ejeren. Dette kan skyldes, at ejeren har en konkurrencefordel og en økonomisk interesse i at hemmeligholde specifikationen. Uden specifikation og beskrivelse er det en svær og dyr affære at forstå et format.

- På et mere kommercielt plan, kan overtagelser af konkurrerende virksomheder føre til, at nye ejere bevidst beslutter at markedsføre og understøtte ét program og format frem for et eller flere konkurrerende og nyligt erhvervet format. Dette sigter mod en bevidst begrænset understøttelse og udbredelse, som igen kan føre til at et format forældes.

Hvordan bevarer jeg mine forskningsdata?

Der er ikke en enkelt løsning på, hvad du skal gøre for at sikre dig, at dine forskningsdata fortsat kan tilgås, læses og genbruges, også om lang tid, hvis du ikke har mulighed for at aflevere dem til Rigsarkivet, der overtager opgaven for dig. Generelt er der udbredt enighed om en række tiltag, som kan medvirke til at reducere risici for forældelse. Det, vi kan være medvirkende til ved at vælge bevaringsegnede formater, er ikke at eliminere problemet, men at forlænge den tid det tager, før det bliver et problem.

Se de fem gode råd til dig, der overvejer et format til dine forskningsdata, som kan medvirke til at reducere risici for forældelse.

Fem gode råd om sikring af data

Henrik Vetter og Filip Kruse, Det Kgl. Bibliotek, december 2019

Her finder du fem gode råd til, hvordan du sikrer dine forskningsdata bedst muligt.

Tip 1: Gør dine data FAIR

At gøre dine data FAIR bidrager til at sikre dem mod logisk forældelse. Jf. FAIR-principperne og langtidsbevaring. I udgangspunktet er valget af format et område, som vedrører "A"-et i FAIR, nemlig at data skal være tilgængelig (eng. accessible) for mennesker og maskiner (programmer), men formatet skal også gerne tillade, at tilstrækkelig rig metadata kan beskrive data, jf. "F"-et (findable) i FAIR. For eksempel skal statistiske datasæt gerne bevares i et format, der giver mulighed for bl.a. at beskrive variabelnavne og labels, og have koder for manglende værdier og nøglevariable, m.fl.

Tip 2: Brug åbne og frie formater

Brug for så vidt muligt formater, der er åbne og frie. Formater som er åbne og frie, og helst med internationale standardiserede specifikationer, har en forventet lavere risiko for formatforældelse og er bedre sikret mod kommercielle interesser, der kan påvirke understøttelsen af særligt ældre formater negativt. I forskningskredse er det ikke altid muligt at anvende åbne og frie formater formater, eksempelvis fordi der kan være en teknologisk binding til maskinel, som producenter har en økonomisk interesse i at hemmeligholde virkemåden af ved ikke at tilgængeliggøre specifikationen af formatet. I så fald anbefales det at eksportere data til et beslægtet åbent format og bevare dette sammen med data i originalformatet. Arbejder du i Word, kan du overveje en kopi i OpenOffice format, arbejder du i SPSS, kan du overveje at eksportere dine data til R eller SDMX.

Tip 3: Brug formater med stor udbredelse

Vælg formater, der er velkendte og udbredte over formater, som er smalle og nicheprægede. Jo større udbredelse, desto større sandsynlighed for, at der findes fællesskaber og bevæggrunde til fortsat understøttelse. Det kan betyde, at de nyeste og mest moderne formater, ud fra en ren bevaringsmæssig synsvinkel ikke er de mest oplagte formater til langtidsbevaring. For bevaringsformater i arkivverdenen vægter det ofte højt, at der er sparsomme og få opdateringer til formatet over en årrække. Der kan være forskningsområder, som kun har mulighed for at anvende små og snævre formater eller sågar egenudviklede formater. I det tilfælde bør der som minimum eksistere en righoldig og fyldestgørende dokumentation og beskrivelse af formatet,

som kan bevares sammen med data. Hvis det er muligt, kan du eksportere data til et udbredt (og gerne åbent) format og gemme denne kopi sammen med data i originalformatet.

Tip 4: Brug formater der er bredt understøttede

Formater, der kan læses af flere forskellige programmer, proprietære som åbne, på forskellige platforme med forskellige operativsystemer, er bevaringsmæssigt mindre risikable end formater, der er tæt knyttet til specifikke programmer og arkitekturer. Er det ikke muligt at vælge sådanne formater, skal du som minimum notere dig så meget som muligt om, hvilke operativsystemer, programmer, platforme m.v., der skal til for at åbne og læse formatet og dets indhold.

Tip 5: Test dine data

Selv et format, som er egnet til langtidsbevaring, kan være ubrugeligt, hvis data er korrupt eller på anden vis ikke valide. Derfor er det vigtigt, at dine data kan testes, verificeres og sikres integritet over tid. Med testes forstås, at formatet og dets dataindhold kan testes for, om det overholder den korrekte notation (syntaks) og/eller om indholdet er meningsskabende. Integritet kan fx kontrolleres ved at registrere filers checksummer over tid på faste tidspunkter og kontrollere, at de ikke ændrer sig over tid.

Undersøg, om der findes værktøjer eller programmer/scripts, der kan teste og verificere formatet. Hvis ikke, kan du være nødsaget til at lave egne tests, som minimum at sikre, at filen kan åbnes og læses af et givet program og, at filen er i det format, filen angiver. Det engelske nationalarkivs PRONOM database indeholder for en stor mængde formater både eksterne signaturer (eksempelvis filendelser) og interne signaturer (eksempelvis en fastdefineret bestemt værdi på et bestemt sted i filen), som kan bruges til at tjekke, om filen er, hvad den giver sig ud for at være. Filidentifikationsværktøjer som DROID og FITS gør bl.a. brug af PRONOM databasen og kan frit anvendes.

Beskrivelse af data - metadata

Henrik Vetter og Filip Kruse, Det Kgl. Bibliotek, december 2019

Metadata er information om data. For at sikre de største chancer for genfinding anbefales det at følge den accepterede standard for bevaringsmetadata. I et langtidsperspektiv er det absolut en fordel, hvis metadata er digitale og kan processeres maskinelt og også gerne opdateres.

Data, der beskriver og giver information om data, er centrale for langtidsbevaring og genfinding af forskningsdata. Digitale metadata er data, der kan processeres maskinelt, i modsætning til metadata i form af fx håndskrevne noter, smalfilm og båndoptagelser.

En metadatastandard

PREMIS (Preservation Metadata: Implementation Strategies) og PREMIS Data Dictionary er ikke en vedtagen standard, men er p.t. den accepterede standard for bevaringsmetadata. Se mere herom i Premis Data Dictionary og DPC Handbook.

PREMIS' standarden er bygget op om fem semantiske elementer. "Semantisk" er her en typologisering ud fra indholdskarakteristika: Enhed, objekt, begivenhed, aktør og rettighed. De fem typologier beskrives som:

- **Enhed** (intellectual entity): samling af indhold, fx som bog. Dette er først og fremmest relateret til objekt, hvorimod de fire øvrige er indbyrdes forbundne.
- **Objekt** (object): afgrænset enhed, der rummer information, fx en pdf-fil.
- Begivenhed (event): en handling i forbindelse med bevaring, fx indlæggelse af en pdf-fil i et repositorie.
- **Aktør** (agent): en aktør (person, institution, organisation), der forbindes med event'en, fx den, der lægger pdf-filen ind.
- **Rettigheder** (rights): tilladelse forbundet med objektet, fx til kopiering i forbindelse med bevaring.

Disse fem elementer angiver, hvad der i bevaringsmæssig sammenhæng bør være metadata om og dermed videre, hvad der betydningsmæssigt skal ajourføres således, at bevaring kan sikre genfinding og dermed muligt genbrug.

Datadokumentation hører sammen med metadata

De fem typologier er, som det ses, kontekstuafhængige. Som oftest er kontekstuel information om data afgørende for genfinding af data – især andres – og dermed potentiel genbrug. Her er

datadokumentation afgørende. Den kan for eksempel være beskrivelse af undersøgelsesdesign og anvendt software og dokumentationen giver således indblik i datas karakter, hvordan data blev skabt og bearbejdet. Datadokumentation er dermed også kilde til betydningen af emneord anvendt som metadata. Videre er den kilde til vurdering af autenticitet, er det originaldata, reviderede eller aktualiserede data, af hvem, hvornår osv. Dette fremgår ikke uden videre af metadata.

Selvom PREMIS er den gældende standard, må det dog anbefales at tilføje yderligere beskrivelse til denne. I forhold til dokumentation er der ikke entydige retningslinjer. Men den kontekstuelle dokumentation af data er en del af arbejdet med metadata og formentlig nøglen til, at fremtidig forskning vil finde og bruge bevarede data. Så dokumentation og metadata hører sammen, men metadata er kun meningsfulde hvis sættet af metadata er mindre og mere simpelt end sættet af forskningsdata.

Forberedelse til langtidsbevaring

Henrik Vetter, Det Kgl. Bibliotek, December 2019

Det er i mange tilfælde overkommeligt at forberede data til langtidsbevaring. Til helt særlige behov findes der særlige produkter, som er specifikt rettet mod de problemer — for eksempel opdatering af metadata — der opstår i forbindelse med langtidsbevaring.

Langtidsbevaring af forskningsdata kan betragtes som en både generel og fagspecifik opgave. Som nævnt skal data forberedes for bevaring. Der er problemer, som for eksempel ændret sprogbrug og opdatering af emneord, som må forventes at være fælles for forskellige fag. Men der kan også være væsentlige forskelligheder. Nogle fag kan generere meget store mængder af data, fysiske materialer kan spille en stor rolle osv. Det er en konkret vurdering, om der er fagspecifikke krav til bevaring af data og dermed behov for helt særlige produkter og faciliteter.

I mange tilfælde er det ret overkommeligt at forberede data til langtidsbevaring. Danmarks Tekniske Universitet, der tilbyder forskerne datalagring, har en beskrivelse, som ofte vil kunne række i en forberedelsesfase. Når det drejer sig om langtidsbevaring af data, er genfinding i særlig grad afhængig af gode metadata og god datadokumentation og vigtigheden heraf vil vokse med tidshorisonten for bevaring. I den sammenhæng kan man for eksempel bruge FAIR-principperne som retningslinje.

Hvis man er tilfreds med den metadatering, som vejledningen fra Danmarks Tekniske Universitet anviser, er næste skridt i databevaringsprocessen at finde et repositorie.

Hvis der er særlige behov til databeskrivelsen, som ikke tilfredsstilles af vejledningen fra Danmarks Tekniske Universitet, fx på grund af specifikke forhold ved data, eller særlige ønsker til forberedelse af data til lagring (for eksempel for at tilgodese opdatering af metadata), kan man overveje at bruge værktøjer fra en af de forskellige platforme, der er til rådighed for denne type opgave.

Særlige værktøjer til langtidsbevaring

Archivematica indeholder værktøjer til brug for sikring af langtidsbevaring af data. Det er en platform, der bygger på standardprodukter, og som opfylder krav om at være interoperationel. Platformen baserer sig på open source produkter, hvorfor den enkelte bruger har mulighed for at modificere platformen til eget brug. I forhold til formater er platformen meget fleksibel. Den

har gode søgemuligheder og kan anvendes (integreres) med tredjeparts produkter. Den er en samling af værktøjer og ikke et repositorie i sig selv.

RODA er en platform der har mange af de samme funktioner som Archivematica, og RODA baserer sig også på open-source og standardprodukter. RODA har et særligt fokus på at sikre autenticitet gennem vedligeholdelse af metadata.

Det gælder om begge førnævnte platforme, at brugen af dem kræver en del arbejde. Men der findes en række produkter, der kombinerer repositorier med værktøjer, der er mere tilgængelige. Se evt. Amorim et al (2017), hvor en række repositorier, der kræver begrænset indsats af brugeren, sammenlignes. Disse repositorier udmærker sig i højere grad ved forholdsvis let anvendelighed, snarere end ved specifik relevans i forhold til langtidsbevaring af data.

Reference

Amorim, R.C., Castro, J.A., Rocha da Silva, J. et al.: A comparison of research data management platforms: architecture, flexible metadata and interoperability, Universal Access in the Information Society, 16(4): 851-862, 2017. https://doi.org/10.1007/s10209-016-0475-y

Fagspecifikke forskningsdata

Henrik Vetter, Det Kgl. Bibliotek, december 2019

Forskellige videnskabelige retninger har forskellige traditioner, og det afspejler sig i forskningsdata. Med hensyn til langtidsbevaring er der især grund til at være opmærksom på, at nogle fag kan have data i form af for eksempel billeder og lyd. Andre fag er karakteriseret ved, at der produceres en meget stor mængde forskningsdata. Disse forhold spiller en rolle, når man skal løse opgaven med langtidsbevaring af forskningsdata.

Humanistiske og samfundsvidenskabelige forskningsdata

Nogen humanistisk og i højere grad samfundsvidenskabelig forskning er rettet mod data forstået som tal, eller genstandsfelter, der kan beskrives udtømmende gennem tal. Hvor talopgørelser ikke er meningsfulde, for eksempel studier af byrum og arkitektur, kan billeder være en erstatning. Hvor billeder og lyd giver fyldestgørende beskrivelser, angiver Rigsarkivet formater (som nævnt JPEG-2000, TIFF, MP3 og/eller WAVE eller MPEG formater), til at klare opgaven. I det omfang Rigsarkivet tilføjer nye formater må man forvente, at disse sikrer langtidsbevaring.

Data som beskrevet ovenfor, er der ikke problemer med at bevare. Men det efterlader analogt materiale med dertil hørende forskningsdata, hvor nytteværdien vil være afhængig af genstandenes fysiske fremtoning, for eksempel tekstur i klæde. Ved digitalisering af analogt materiale anbefales det, at man fra start benytter et af de ovennævnte formater.

Naturvidenskabelige forskningsdata

Inden for naturvidenskab er det særligt relevant at være opmærksom på, om man genererer store mængder eksperimentelle data. Joint European Torus er et eksempel. JET er et forskningsanlæg for forsøg inden for fusionsenergi. Mens JET er i brug skabes data i forbindelse med hvert forsøg, og der kan for det enkelte være tale om meget store datamængder. Det er relevant at bevare data ud over det enkelte eksperiments tidshorisont og sådan set også længere end JET's fysiske levetid.

Eksemplet viser to overvejelser, som er vigtige i forhold til naturvidenskab. Erfaringer fra JET viser, at når der genereres meget store datamængder, er det væsentligt, at bevarings- og arkiveringsstrategi fra starten er indlejret i projektet. Spørgsmålet om databevaring er altså også et spørgsmål om organisation. JET's første arkiv dækker en periode fra år 1983 til år 2000, hvor data migreredes til et nyt arkiv. I tilfældet med JET viste det sig, at udskiftning af

arkivmedier (maskiner) var et mindre problem. Derimod har forældelse af filformater vist sig at være et problem. For yderligere beskrivelse henvises læseren til Layne et al (2012).

Layne et al (2012) identificerer tre særlige opmærksomhedspunkter. Langtidsbevaring af store datamængder er et organisationsspørgsmål. Arkivmedier forældes og langtidsbevaring og adgang kræver derfor planlagt udskiftning af arkivmedier. Formater til bevaring skal være velbeskrevne og dokumenterede og denne information skal opbevares centralt i organisationen.

Biomedicinske forskningsdata

Der er inden for det biomedicinske område særlig interesse for dataopbevaring, formentlig fordi området er udviklet til at være datatungt. Navale og McAuliffe (2018) giver den meget specifikke anbefaling, at man gennem Open Archival Information System modellen sikrer autenticitet og præcision i forhold til opbevaring af biomedicinske data. OAIS-modellen fungerer sammen med for eksempel Archivematica. Der findes en generel introduktion til best practices for bevaring – men ikke med særligt fokus på langtidsbevaring – af biomedicinske data.

Sundhedsvidenskabelige forskningsdata

Sundhedsvidenskabelige forskningsdata stammer fra forskningsaktiviteter.

Referencer

Layne, R., Capel, A., Cook, N., & Wheatley, M. (2012) Long term preservation of scientific data: Lessons from jet and other domains, Fusion Engineering and Design, 87(12), pp. 2209 – 2212. https://doi.org/10.1016/j.fusengdes.2012.07.004

Navale, V. & McAuliffe, M. (2018), Long-term preservation of biomedical research data [version 1; peer review: 4 approved, 1 approved with reservations], F1000Research, 7:1353. https://doi.org/10.12688/f1000research.16015.1

Repositorier

Henrik Vetter, Det Kgl. Bibliotek, december 2019

Bevaring af forskningsdata – når data er gjort klar med metadatering, mv. – sker i et repositorie, hvis ikke i Rigsarkivet. Der findes en lang række repositorier, så det skulle være muligt at finde et velegnet. Ved langtidsbevaring skal man have særlig opmærksomhed på, om det er muligt at migrere data fra et repositorie til et andet. Det er også ønskværdigt, at repositoriet understøtter forskellige formater.

Når man har forberedt sine data, er der flere steder, hvor man som forsker kan finde arkiver til sine egne data og søge videre efter andres data. Det mest omfattende er r3data.org. Det er et tyskbaseret, globalt register over repositorier til permanent bevaring af data, dækkende de fleste akademiske fagområder. Der kan søges fagligt afgrænset, efter typer af indhold og efter land. Der er kort beskrivelse med emneord, oversigt over muligheder for adgang til data, upload mm. og links til andre arkiver inden for fagområdet.

Tidsskriftet Nature vedligeholder en fagopdelt liste over repositorier anbefalet af Scientific Data (Springer).

Fra DeiC's hjemmeside er der adgang til en liste over værktøjer relateret til FAIR data, herunder repositorier, vejledninger mm.

Kriterier for valg af repositorie

Danmarks Tekniske Universitet anbefaler, at et repositorie skal opfylde disse betingelser:

- Være anerkendt i forskningsverdenen
- Have politik for adgang til data og betingelser for brug
- Være funderet i en organisation, gerne universitet eller bibliotek
- Understøtte de almindelige standarder for metadata
- Tilbyde langtidsholdbare og unikke ID'er, f.eks. DOI'er
- Tilbyde standardlicensvilkår for datasæt
- Være certificeret eller anvende særlige arkivstandarder

Blandt andet disse repositorier lever op til kriterierne:

- Dryad
- Giga'nDB

- GigaScience
- GitLab
- Protein Data Bank
- Figshare
- Zenodo

I forhold til langtidsbevaring af forskningsdata er der yderligere to vigtige kriterier:

- Tillade migrering til andet repositorie
- Understøtte flere dataformater og -typer

Migrering af data er vigtigt, fx hvis institutionen ændrer i anbefalinger af valg af repositorie eller opretter sit eget.