



**Roskilde
University**

FIF - Fælles Infrastruktur for Forskningsdata

Rapport fra fase I

Larsen, Asger Væring; Zurcher, Sacha; Andersen, Mikkel Hvidtfeldt; Thestrup, Jesper Boserup; Ærendahl Mikkelsen, Søren; Sandfær, Mogens; Conrad, Anders

Publication date:
2014

Document Version
Også kaldet Forlagets PDF

Citation for published version (APA):

Larsen, A. V., Zurcher, S., Andersen, M. H., Thestrup, J. B., Ærendahl Mikkelsen, S., Sandfær, M., & Conrad, A. (2014). *FIF - Fælles Infrastruktur for Forskningsdata: Rapport fra fase I*. DEFF.
<http://projekter.kulturstyrelsen.dk/sites/default/files/documents/FIF%20afsluttende%20rapport%2020140718.pdf>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact rucforsk@kb.dk providing details, and we will remove access to the work immediately and investigate your claim.

DEFF

FIF - Fælles Infrastruktur for Forskningsdata

Rapport fra fase I.

Asger Væring Larsen, Sacha Zurcher, Mikkel Hvidtfeldt Andersen, Jesper Boserup
Thestrup, Søren Ærendahl Mikkelsen, Mogens Sandfær, Anders Conrad
11-07-2014

Indhold

Resumé.....	1
Kapitel 1: Indledning	2
Kapitel 2: Data Management Planning (før forskningsprojektet)	6
A. Data Management Planning	6
B. Test af DCC's DMPonline	6
C. Funktionalitet i DMPonline	6
D. Software.....	7
E. Fremtidsønsker.....	7
F. Opsummering.....	8
Kapitel 3: Managing active data (under forskningsprojektet)	9
A. Generelt om modtagelse, forvaltning og bevaring af forskningsdata.....	9
B. Forvaltning af aktive forskningsdata	9
C. Brug og videre bearbejdelse af aktive forskningsdata	10
D. Opsummering	11
Kapitel 4: Data selection and handover (umiddelbart efter forskningsprojektets afslutning)	12
A. Data repositories	12
B. Testcase: Curate / Hydradata	13
C. Andre deponeringstjenester.....	21
D. Opsummering	22
Kapitel 5: Data repositories (efter forskningsprojektet, langtidsbevaring)	24
E. Opsummering.....	25
Kapitel 6: Data Catalogues – om publicering, identifikation, genfinding, citering og meritering af forskningsdata.....	26
A. Indledning	26
B. Publicering af data	27
C. Identifikation af data	29
D. Genfinding og genbrug af data	31
E. Citering og meritering af data	32
Kapitel 7: Guidance, training and support (værktøjer til understøttelse og uddannelse)	34
A. Vurderingskriterier	35
B. Værktøjer rettet mod planlægningsfasen	36
C. Værktøjer rettet mod den udførende fase.....	37

D. Værktøjer rettet mod publicering og langtidsbevaring.....	38
E. Værktøjer rettet mod vejledning, træning og support	38
F. Opsummering.....	39
Bilag 1: Studieture til Holland og UK, november-december 2013.....	41
Studietur Edinburgh og Oxford	43
Annual Workshop of the International Association of Scientific and Technological Libraries (IATUL).....	46
Bilag 2: Vurdering af informations-, trænings- og planlægningsværktøjer for data management	48
Bilag 3: DEFF-ansøgningen <i>Data management i praksis</i>	51

Resumé

I forbindelse med DEFF-projektet "Fælles Infrastruktur for Forskningsdata" er et par softwareløsninger blevet installeret og evalueret. Samtidig er forskellige trænings- og supportværktøjer blevet evalueret med udgangspunkt i Research Data Management livscyklussen.

DMPonline er blevet installeret. Softwaren vil relativt simpelt kunne tilrettes danske behov. Tjenesten vil kunne køre som en national skytjeneste. I forhold til at etablere tjenesten er det en stor udfordring at sikre, at alle interessenter har udviklet deres egen politik, da disse er udgangspunkt for DMPonline's funktioner.

Håndteringen af data i løbet af forskningsprocessen vurderes at være så komplekst et område, at service her bedst kan udvikles i tæt samarbejde med konkrete forskningsprojekter.

En institutional repository-løsning, Curate, er blevet installeret og testet. Det skete for at vurdere anvendelsesmulighederne under Data Selection and Handover samt i tilknytning til langtidsopbevaring i Data repositories. Curate vurderes at kunne tilrettes danske forhold relativt let og vil kunne fungere som både national og lokal tjeneste. I forbindelse med langtidsopbevaring kræves mere af et givent system, som det eksempelvis er implementeret i Det Nationale Bitmagasin.

Det er nødvendigt, at de anvendte systemer giver mulighed for at dele datasæt. Datasæt kan tilknyttes permanente identifikationer, som eksempelvis DOI, og andre forskere kan sikres styret adgang til data. Der redegøres for brugen af data repositories og brugen af *data catalogues* og discovery tjenester.

DMPTool og DMPonline vurderes som velegnede i forbindelse med support. Disse to softwareløsninger er designet efter de krav, der stilles til ansøgeren, hvilket vil gøre det let for vejledende personale at hjælpe forskeren. MANTRA og RDMRose er, i denne sammenhæng, oplagte værktøjer i forhold til at opbygge supportforløb.

Kapitel 1: Indledning

Denne rapport er blevet udarbejdet i forbindelse med projektet "Fælles Infrastruktur for Forskningsdata". Projektet blev startet i foråret 2013 og afsluttedes i juli 2014. Projektet bag rapporten var DEFF-finansieret, med deltagere fra Roskilde Universitetsbibliotek, Statsbiblioteket, Det Kongelige Bibliotek, Danmarks Tekniske Informationscenter og Danish e-Infrastructure Cooperation. Copenhagen Business School deltog i projektets tidlige periode, men måtte udtræde pga. langtidssygdom hos en nøgleperson.

Projektet var oprindeligt tænkt i 2 faser, med nærværende projekt (fase 1) som forprojekt til et større implementeringsprojekt (fase 2). Der var i ansøgningen opstillet 3 leverancer til fase 1:

L1.1	Testområde	Etablering af fælles testområde og – systemer til løbende evaluering.
L1.2	Behovsopgørelse	Sammenfatning af krav og behov fra tidligere initiativer. Evaluering af relevante services og systemer. Evaluering af eksisterende ressourcer indenfor træning, vejledning og god praksis. Indstilling af evaluerede løsninger til dansk lokalisering og implementering. Beskrivelse af evt. supplerende ressourcer. Kravspecifikation.
L1.3	Projektinitieringsdokument	Projektinitieringsdokumentet fastlægger endeligt projektets økonomi og en detaljeret plan for anskaffelse af projektets leverancer.

L1.1, Testområde, anses for opfyldt i form af den skytjeneste som DelC har stillet til rådighed, som har omfattet virtuelle maskiner, storage og netværksadgang. Dette testområde er blevet benyttet til installering og evaluering af både DMPOnline (kapitel 2) og Curate (kapitel 4).

L1.2, Behovsopgørelse, anses for opfyldt gennem nærværende rapport.

L1.3, Projektinitieringsdokumentet, som i den oprindelige ansøgning var tænkt som en plan for fase 2 af projektet, anses for opfyldt i form af DEFF-ansøgningen *Data management i praksis*, som erstatter fase 2 af FIF-projektet. Denne ansøgning vedhæftes derfor rapporten som bilag 3.

Oprindeligt var det planlagt at inkludere Dataflow projektet, en repositorielløsning fra University of Oxford, i evalueringen, men i løbet af projektet tegnede der sig et billede af, at konkrete repositorielløsninger (f.eks. Easy) i høj grad leverede specialiserede services baseret på udbygninger af standardkomponenter som Fedora Commons eller Dspace, og at Dataflow suite i højere grad var en speciel løsning. Sustainability afhænger af, at institutioner, der skal implementere services, bruger en teknologistak, som passer ind i deres generelle arkitektur og strategi. På den baggrund blev det besluttet i stedet at fokusere på Curate, som bygger på Hydra

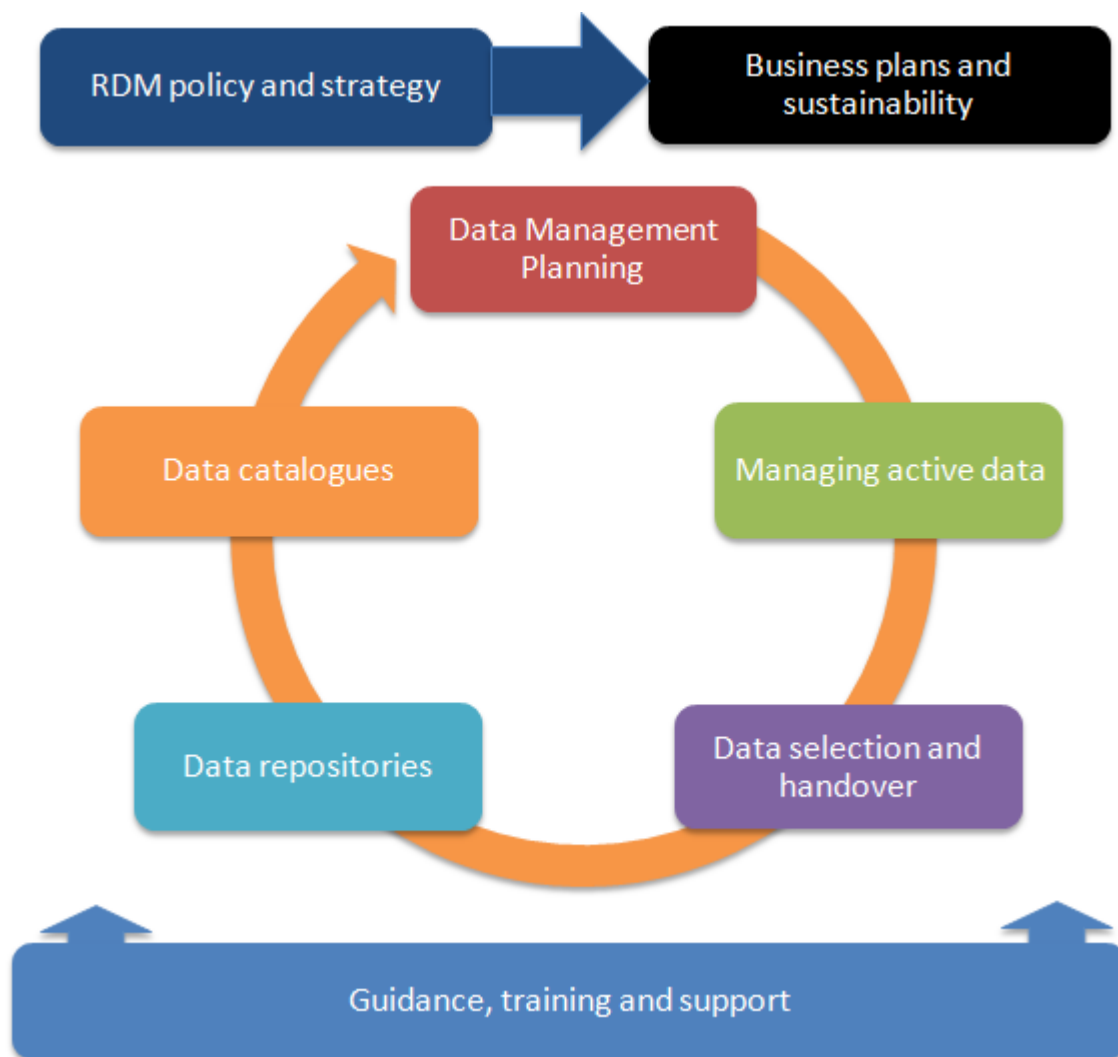
løsningsen, der udvikles af et dynamisk og voksende community, og som allerede er kendt og i brug på Det Kongelige Bibliotek.

Sideløbende med projektet gennemførte både DeIC og DEFF forskellige strategiprocesser og, for DEFF's vedkommende, en organisatorisk omlægning. I den proces blev fase 2 aflyst og erstattet af en anden selvstændig projektansøgning. I samarbejde med DEFF besluttedes det, at videreføre FIF projektets fase 1 for at sikre viden om Data management, der kan bruges af andre. Denne proces afsluttedes med udarbejdelsen af nærværende rapport. Rapportens forfattere håber at kunne bidrage til DeIC's strategiproces, der forventes afsluttet i efteråret 2014 og samtidigt kunne bidrage med viden i forbindelse med udarbejdelsen af projekter, der indsendes til DEFF i forbindelse med DEFF's projektkald, der forventes udsendt i august 2014.

Projektet tager sit afsæt i en stigende bevidsthed om vigtigheden af at opsamle og videreudnytte forskningsdata og grå litteratur. På verdensplan har mange institutioner allerede arbejdet med udformning af relevante services, og der kan høstes af andre landes erfaringer og løsninger. Især Storbritannien, Holland, Australien og USA er langt fremme; Digital Curation Centre, University of Edinburgh og University of Oxford er fremtrædende eksponenter.

Danske institutioner er bevidste om behovet, men der er et væsentligt, udækket behov for normer og løsninger. Udfordringens størrelse og facetter peger desuden på hensigtsmæssigheden af en effektiv institutionel arbejdsdeling. Erfaringer fra sektorens tidlige initiativer, inkl. DEFF projektet Forvaltning af Forskningsdata og DeIC's igangværende strategiproces, har bidraget til at kortlægge krav og behov for håndteringen af danske forskningsdata.

Visse typer af data inden for sundhed og samfundsvidenskab langtidsbevares af Dansk Data Arkiv, men funktionen understøtter ikke den tidlige del af forskningsprocessen, og data er ikke alle i dag tilgængelige online. Samlet set er mængden af data, som opsamles, sikres og gøres online tilgængelig, fortsat lille.

Den benyttede Data management livscyklus model

Kilde: Digital Curation Centre: How to Develop RDM Services - a guide for HEIs, <http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services>.

I forbindelse med projektets undersøgelse, har projektgruppen brugt ovenstående Data Management Livscyklus model. Der findes mange forskellige modeller, der inddeler datas livscyklus i flere delfaser.¹ Vi har valgt denne model, da det i forhold til vores undersøgelse af forskellige systemer og services var simplere at gennemføre undersøgelsen, og DCC er normsættende for god praksis indenfor området i Storbritannien.

Modellen viser, at forskningsdata gennemløber 5 faser fra forskerens indledende overvejelser og planlægning (data management planning), forskernes håndtering af data i løbet af forskningsprocessen (managing active data), udvælgelsen af data der skal gemmes (data selection

¹ Se f.eks. CEOS Working Group on Information System and Services, *Data Lifecycle Models and Concepts*, version 1.2 (<http://www.ceos.org/images/DSIG/Data%20Lifecycle%20Models%20and%20Concepts%20v13.docx>, 16/7-14) for 55 sådanne.

and handover), lagring af data (data repositories) og deling af gemt data (data catalogues). Ved siden af dette illustrerer modellen, at der er behov for støtte og undervisning i hele forløbet.

På den baggrund er rapporten opdelt i følgende afsnit baseret på de fem faser og en tværgående undersøgelse af undervisningsmateriale:

- Data management planning, den første fase (før forskningsprojektet)
- Managing active data (under forskningsprojektet)
- Data selection and handover (umiddelbart efter forskningsprojektets afslutning)
- Data repositories (efter forskningsprojektet, langtidsbevaring)
- Data catalogues (genfinding, genbrug)
- Guidance, training and support

Kapitel 2: Data Management Planning (før forskningsprojektet)

A. Data Management Planning

I USA, England, Australien og Holland stiller forskningsfonde ofte krav om, at der udarbejdes en data management plan som led i ansøgningsprocessen.

De forskellige fonde har ofte forskellige krav til indholdet af en data management plan, hvilket afspejler de forskellige videnskabelige traditioner indenfor forskningsfelterne. Forskerens forskningsinstitution kan også stille krav til indholdet af en data management plan, med udgangspunkt i den forskningsdata politik samt de forskningsdata services, som institutionen tilbyder.

En data management plan vil som minimum indeholde en definition af forskningsprojektets data, hvordan data opbevares undervejs og efter projektet, samt hvem der må tilgå data.

I England og USA har man udviklet IT-værktøjer, som guider forskerne igennem de forskellige krav til data management-planer. DMPonline og DMPTool er værktøjer der bygger på, at forskeren først vælger ønsket forskningsfond samt forskningsinstitution, hvorefter værktøjet opbygger en skabelon med de krav den pågældende forskningsfond og forskningsinstitution stiller. Hvert krav resulterer i et eller flere spørgsmål, som forskeren skal besvare. Via værktøjerne kan forskeren samarbejde med kollegaer, arbejde videre med planerne efterhånden som projektet skrider frem, samt eksportere data management-planen.

B. Test af DCC's DMPonline

Det engelske "Digital Curation Centre" (DCC) (<http://www.dcc.ac.uk/>) har udviklet værktøjet DMPonline (<https://DMPonline.dcc.ac.uk>). DCC samlede kravene til indholdet i ansøgningerne fra Storbritanniens forskningsråd, fonde med videre i en database og kombinerede dette med de krav universiteterne stiller til forskere. Det skabte et redskab, der i høj grad er tilpasset specifikke faglige og institutionelle krav. Softwarens udvikling er blandt andet beskrevet på både DCC og DMPonline's hjemmesider samt af Martin Donnelly i artiklen "Data management plans and planning" side 83-104 i Graham Pryor (Red.): Managing Research Data; Facet Publishing, 2012.

DMPonline er frigivet som et open source projekt i GitHub. Dels på grund af DMPonline's tilgængelighed men også på grund af relationer til folkene bag DCC, så har FIF projektet valgt at download DMPonline softwaren og installere den i en DEIC etableret skyløsning².

Det er den nye version 4 af DMPonline softwaren som er testet. Version 4 fremstår væsentlig forbedret i forhold til version 3 mht. brugervenligheden. Specielt er skabelonerne forenklet med færre spørgsmål, men også layout og usability er forbedret, så det er relativt simpelt for brugerne at få overblik over værktøjets formål og funktionalitet.

C. Funktionalitet i DMPonline

DMPonline har følgende end-user funktionalitet:

Brugeroprettelse: Brugeren opretter sig i systemet med email, password, fornavn og efternavn. I oprettelsesprocessen afsendes en email med et link, som brugeren skal aktivere for at bekræfte sin identitet.

Se tidligere planer: Brugeren kan se en liste over tidligere oprettede planer og redigere videre i disse.

² I denne proces har DMPonline udviklingskoordinatorer Marta Ribeiro fra DCC været en stor hjælp.

Opret ny plan: Brugeren kan oprette en ny plan. I oprettelsesprocessen skal brugeren vælge den fond der ansøges samt brugeren forskningsinstitution.

Del plan: Brugeren kan invitere andre til at se og rette i en plan

Eksporter plan: En plan kan eksporteres til følgende formater: HTML, XML, Text eller Json

Følgende skal administreres af serviceudbyderen:

Liste over uddannelsesinstitutioner og fonde: I testen er der ikke fundet noget administrativt værktøj til administration af uddannelsesinstitutioner og fonde. Det er dog muligt at ændre og tilføje institutioner og fonde via datafiler, som importeres til databasen via Ruby on Rails frameworket.

Fondes og forskningsinstitutioners datamanagement skabeloner: I testen er der ikke fundet noget administrativt værktøj til administration af datamanagement skabeloner. En fonds datamanagement skabelon kan nedbrydes i en række delspørgsmål og hjælpetekster inden for forskellige kategorier. Alt data som udgør en skabelon ligger i datamodellen spredt over en håndfuld tabeller. Det er muligt at ændre og tilføje datamanagement planer via datafiler, som importeres i databasen via Ruby on Rails frameworket.

Grænseflade tekster og logo'er: Alle faste tekster i brugergrænsefladen er placeret i en konfigurationsfil. Det er derfor nemt at ændre tekster og logo'er i brugergrænsefladen

D. Software

Software er skrevet til Ruby on Rails. Ud over installation af Ruby on Rails, så kræver løsningen en MySQL database samt en mailserver. Mailserveren anvendes i forbindelse med brugeroprettelsesprocessen, hvor der sendes et aktiveringslink til brugerens konto.

GitHub open source-projektet³ bærer præg af at være relativt nyetableret. Dokumentation er stort set fraværende. Dette var specielt problematisk i forbindelse med installationsprocessen, hvor det tog nogen tid at udregne, hvordan konfiguration og indlæsning af data i databasen skulle foretages.

Der mangler også dokumentation på, hvordan nye uddannelsesinstitutioner, fonde og data management-skabeloner oprettes. Fra DCC's side er de opmærksomme på den manglende dokumentation, og de planlægger at arbejde med dette i nær fremtid.

Testen viste, at DMPonline kan håndtere dansk tegnsæt, og at det er relativt nemt at ændre i data management-skabelonerne via de medfølgende datafiler. Endeligt viste testen, at grænseflade teksterne er isoleret i en tekstfil, og ændring i denne slår igennem til grænsefladen.

Da DMPonline samtidig er et relativt simpelt system, så forventes det, at en Ruby on Rails programmør nemt vil kunne tilpasse løsningen til danske forhold. En tilpasning bør kunne indrettes, således at migrering til fremtidige DMPonline versioner kan foretages.

E. Fremtidsønsker

Udarbejdelsen af data management-planer og dermed brugen af værktøjer som DMPonline, sker primært i de indledende faser af et forskningsprojekt. Der lægges godt nok op til, at forskeren kan arbejde inkrementelt på planerne, efterhånden som projektet skrider frem. Men der er på den anden side ikke noget, som tilskynder forskeren til at gøre dette. Dette er en skam, da erfaringer

³ https://github.com/DigitalCurationCentre/DMPonline_v4

viser, at strukturering og dokumentation af forskningsdata med fordel kan ske i starten og undervejs i et forskningsprojekt, frem for ved afslutning af et forskningsprojekt.

Et fremtidsønske til den videre udvikling af DMPonline er derfor en tættere integration til de efterfølgende processer og værktøjer i forskningsprojektet. Fx via links til data repositorier og publiceringsplatform, men også via udtræk og genbrug af forskningsdatadokumentation på tværs af systemerne. Fx dokumentation af dataformater, sammenhæng mellem data og adgang til data. En sådan integration vil gøre data management-planerne aktuelle igennem hele forskningsprojektet. Desuden har det den fordel, at forskeren primært vedligeholder dokumentation vedr. forskningsdata et sted.

Et andet fremtidsønske til DMPonline er Single Sign On mod andre identitets føderationer, som fx WAYF, hvorved forskeren kan logge på DMPonline via sin lokale forskningsinstitution.

F. Opsummering

Værktøjet DMPonline lægger op til en national service, og der vil med en relativ lille indsats kunne etableres et nationalt data management planning service med udgangspunkt i DMPonline. Tilpasning til danske forhold via open source softwaren fra GitHub vil højst tage et par måneder for en erfaren Ruby on Rails programmør. Softwaren vil kunne driftes i en hvilken som helst skyløsning.

Den primære udfordring ved etablering af et nationalt DMP tool er af politisk og organisatorisk karakter. Værktøjet kommer med en default skabelon til data management krav. Men implementering af DMPonline i den danske kontekst vil først give mening når forskningsfondene og forskningsinstitutionerne har udarbejdet forskningsdata politikker, som kan omsættes til krav i DMPonline. Der skal etableres en organisation som kan vedligeholde, drifte og markedsføre DMPonline.

Kapitel 3: Managing active data (under forskningsprojektet)

A. Generelt om modtagelse, forvaltning og bevaring af forskningsdata

Der må lægges til grund at forskellige videnskabelige discipliner og forskningsmiljøer har meget forskellige tilgange og behov, både i forhold til skabelse og brug af data i forskningsprocessen. Der er stor forskel på fx. et astronomisk eller meteorologisk projekt der opsamler data fra en satellit, et medicinsk laboratorieforsøg, en spørgeskemaundersøgelse og et projekt som laver annotationer til et sprogligt eller litterært korpus.

Nogle af deltagerne i projektgruppen har i relaterede projekter haft til opgave at interviewe humanistiske forskere om deres tilgang til brug og deling af forskningsdata⁴. Vi har kunnet konstatere at der selv mellem humanistiske forskere er et spænd mellem fx at bygge og metadatere et korpus af kulturarvskilder til mere analytiske og fortolkende annotationer som ligger tættere på den endelige forskningspublikation.

Der skal således ikke i FIF-projektets begrænsede rammer gøres noget forsøg på en fuldstændig udredning af håndteringen af de data som genereres i forskningsprojekter. I stedet vil vi, i dette og i de 2 efterfølgende kapitler, eksemplificere nogle enkelte løsningsmuligheder for håndtering af forskningsdata, fra projekt til langtidsbevaring, som baserer sig dels på erfaringer som allerede er indhøstet på deltagerinstitutionerne, dels enkelte afprøvninger undervejs, og dels noget af den viden vi har indhøstet på studieture til Holland og England⁵.

Den benyttede livscyklusmodel skelner mellem data der opsamles/opbygges under forskningsprojektet, data som deponeres til et arkiv efter projektafslutning samt langtidsbevaring. Det er underforstået at der er store forskelle i både formål og behov i forhold til at opsamle og eventuelt dele forskellige datasæt, samt i hvilken grad det er formålstjenligt og økonomisk muligt at langtidsbevare datasæt. Disse forskellige forhold vil ikke blive yderligere belyst i denne sammenhæng hvor vi alene peger på nogle teknologiske og organisatoriske rammer for nogle mulige løsninger.

B. Forvaltning af aktive forskningsdata

Uanset om man er laboratorieforsker som opsamler data direkte fra instrumenter og måleudstyr, eller humanist som annoterer tekster eller videoklip, er det vores indtryk at forskerne ønsker at en eventuel forvaltning af data under forskningsprocessen skal ske så smertefrit som muligt, og uden ekstra tidsforbrug. Gevinsterne for forskeren ved en professionel dataforvaltning i forskningsperioden kan være at der er sikkerhed for backup, beskyttelse af fortrolighed og mulighed for samarbejde i teams og i øvrigt en kontrolleret deling af forskningsdata, i princippet i samme øjeblik de er blevet skabt.

Ønsket om en smertefri eller sømløs overlevering af aktive data til forvaltning, betyder i en eller anden forstand at forskeren gør hvad han/hun i øvrigt ville gøre, og at noget teknologi sørger for

⁴ Der er både tale om dels Knowledge Exchange's undersøgelse af incitamenter for at dele forskningsdata, dels om DeIC's afdækning af de faglige miljøers behov og præferencer i forbindelse med en national strategi for forskningsdata. Begge disse undersøgelser spænder bredt over mange fagområder, mens de interviews som er blevet gennemført af hhv. Statsbiblioteket og Det Kongelige Bibliotek, har rettet sig mod humaniora.

⁵ Jf. bilag 23423894 : studieture i UK og Holland.

indhøstning, deponering eller lignende. I nogle af de humanistiske eksempler, foregår skabelsen af data direkte via værktøjer i en forskningsinfrastruktur og forvaltes og deles dermed direkte og fuldautomatisk⁶. I andre tilfælde kunne man forestille sig at data fra instrumenter blev deponeret direkte i et repositorie eller en database og efterfølgende står til rådighed for bearbejdelse og analyse i forskernes værktøjer.

Et ofte fremført ønske er at kunne lægge sine forskningsdata i en DropBox el. lign., altså i en mappe på sin computer, hvorved filerne automatisk gemmes på en server og efterfølgende kan ses på flere enheder og evt. af flere brugere efter nærmere tilladelser. Nogle humanistiske forskere i Knowledge Exchange-undersøgelsen (se note 4 ovenfor) sagde at hvis de kan dele uden at gøre ekstra arbejde, så deler de, og ellers ikke.

Vi så på vores studietur til UK eksempler på at universiteter via deres egen IT-afdeling indkøbte storage og helt rutinemæssigt stiller et netværksdrev til rådighed for alle forskere til brug for forskningsdata⁷. En sådan løsning kan give en sikker forvaltning af filerne, men risikerer at give institutionelle begrænsninger i forhold til deling og samarbejde på tværs da den typisk vil være bundet op på institutionens egen interne brugerbase og brugervalidering. En DropBox eller lignende skyløsning synes at være et simpelt svar på dette problem, men resulterer til gengæld i at data ligger uden for universitetets egen forvaltning.

I en dansk sammenhæng vil det være interessant at gøre erfaringer med DelCBox, men det er ikke noget projektgruppen har haft mulighed for at arbejde med.

C. Brug og videre bearbejdelse af aktive forskningsdata

Man kan sige at opsamling af forskningsdata via netværks- eller skydrev som udgangspunkt resulterer i et råt og ukurateret datasæt. Hvis der er tale om data der skal gemmes og eventuelt langtidsbevares efter forskningsprojektet, er der imidlertid en udfordring med at få dem forsynet med tilstrækkelige metadata til at de senere vil kunne identificeres, fremsøges og genbruges. Det er næppe sandsynligt ud fra de simple eksempler med virtuelle drev vi har set i projektet, at man vil ønske eller have råd til at arkivere eller langtidsbevare alle sådanne opsamlede datasæt.

Der er i givet fald brug for at få etableret en proces til at udvælge og automatisk metadatere indsamlede datasæt. Skønt sådanne metoder formodes at eksistere i forhold til opsamling af laboratedata som skabes af apparatur som kan udvikles til at levere de nødvendige metadata, så har vi ikke set eksempler på dette i projektets begrænsede arbejde med menneskeskabte data på en computer.

Vi er stødt på teknologier som integrerer skytjenester (Dropbox, Google Drive og Box) med almindelige institutional repository-funktioner. I Curate som vi har afprøvet (se nedenfor) er det således muligt at hente filer ind fra et skydrev. Her er der dog tale om en manuel kuratering hvor der indtastes de nødvendige metadata omkring datasættet hvorefter filerne udvælges og

⁶ Det gælder fx. annoteringer til radioprogrammer som foretages i grænsefladen LARM.fm og som herefter gemmes i en database og gøres synlige for andre forskere (<http://larm.fm/>).

⁷ Fx. University of Edinburgh som stiller 0,5 Tb gratis til rådighed for alle aktive forskere; skal man bruge mere plads, koster det penge.

uploades til repositoret. Dette kan næppe tilgodeses den use case som blev antydnet ovenfor, hvor data fra et skydrev direkte bliver høstet og delt.

Rent teknisk integrerer Curate med de forskellige skytjenester ved hjælp af et sæt drivere som arbejder op mod de pågældende tjenesters API'er. Disse drivere ligger i et generel Ruby on Rails⁸-bibliotek, og vil derfor kunne bruges i andre Ruby on Rails-applikationer, fx. hvis man ville lave en automatisk indhøstning af data fra et skydrev som Dropbox eller DelCBox til et repositorie. Vi vurderer det som relativt overkommeligt rent teknisk at etablere en sådan løsning, men der vil være betydelige spørgsmål omkring udvælgelse, adgangskontrol, rettigheder og metadatering som vil skulle afklares.

D. Opsummering

Forskerne ønsker at en eventuel forvaltning af data under forskningsprocessen skal ske så smertefrit som muligt, som en integreret del af forskningsprocessen. En simpel løsning for dette synes at være til stede i form af virtuelle fildrev som placeres i skyen. Dette kan ske enten på institutionelt drevne filservere eller på kommercielle tjenester som til stadighed udvider deres tilbud. I en vis forstand kan data siges at være under forvaltning, alene ved at blive placeret på et sådant skydrev som bliver drevet af en professionel organisation.

Data vil i denne tilstand dog ikke let kunne findes og genbruges af andre, dette kræver en grad af kuratering (udvælgelse, metadatering, evt. formatkonverteringer osv.). Disse spørgsmål synes i en vis grad at hænge tæt sammen med selve den videnskabelige praksis de skal understøtte. Selv om en indhøstning i et repositorie formentlig vil være relativt enkel at etablere rent teknisk, vil kurateringen afvige stærkt fra fag til fag. Projektgruppen anbefaler dette område som egnet til mindre, eksperimenterende projekter sammen med udvalgte forskerteams.

⁸ Ruby on Rails er navnet på det programmeringssprog som bruges af Curate og af det rammesystem, Hydra, som Curate er baseret på.

Kapitel 4: Data selection and handover (umiddelbart efter forskningsprojektets afslutning)

A. Data repositories

Data repositorier er rammen om hvad man kunne kalde en kurateret deponering af forskningsdata til arkivering og/eller langtidsbevaring efter et projekts afslutning. Her er der tale om at datasæt bliver udvalgt med henblik på bevaring, ud af en måske meget større mængde af data som har eksisteret undervejs i projektet. Alternativet vil ofte være at data forsvinder når projektets servere eller storage-områder nedlægges efter projektafslutning. Der er en glidende overgang mellem deponering og arkivering/langtidsbevaring som behandles i næste kapitel. I dette kapitel vil vi derfor fokusere på selve deponeringen.

Udvælgelsen af data til arkivering kan ske ud fra en række forskellige kriterier, fx. behovet for at kunne linke fra en artikel til bagvedliggende datasæt, muligheden for at efterprøve beregninger, konklusioner eller fortolkninger og relevansen for senere genbrug i andre projekter. Der er typisk tale om at forskeren selv eller en bibliotekar tilføjer beskrivende metadata, eventuelt en identifier (fx DOI), og uploader en eller flere filer for det pågældende datasæt til et repository.

De fleste repositorier bygger på en Open Access-filosofi, men der kan være juridiske (typisk ophavsretslige eller persondatarelaterede) eller karieremæssige (ønsket om at komme først med at udnytte datamaterialet) årsager til at nogle datasæt holdes skjult, eventuelt med en embargoperiode i forbindelse med publicering af resultaterne. Igennem repositoryets søgegrænseflade vil arkiverede datasæt kunne findes frem, evalueres og hentes ned af andre hvis de er frit tilgængelige. Der kan være tale om at man ved deponering skal godkende en deponeringslicens som nærmere angiver parternes rettigheder og betingelserne for repositoryets forvaltning og tilgængeliggørelse af datasættet. Ligesom slutbrugere i visse tilfælde skal godkende en slutbrugerlicens, fx. fra Creative Commons-familien.

Repositorier oprettes og drives rundt omkring af forskellige aktører, organiseret efter forskellige principper og institutionelle rammer. Projektgruppen har grundlæggende identificeret tre typer datarepositories:

- repositorier som drives på nationalt plan og stilles til rådighed for den pågældende nations forskningsinstitutioner
- institutionelle repositorier som drives af en enkelt institution, typisk et universitet, for den pågældende institutions forskere
- fagspecifikke repositorier som kan gå på tværs af både kontinenter, landegrænser og institutioner, men som til gengæld kan være relativt snævre rent indholdsmæssigt.

Et kendt eksempel på et datarepository er det hollandske *EASY*⁹ som drives af DANS, som er en national institution der opsamler og så vidt muligt tilgængeliggør forskningsdatasæt. EASY er baseret på en front-office / back-office model som fordeler rollerne mellem forskellige aktører. Det giver mulighed for variation i front-office tilbuddene som bliver varetaget af lokale universitetsbiblioteker, og som især går ud på direkte at supportere de forskellige forskergrupper.

⁹ <https://easy.dans.knaw.nl/ui/home>

Back-office er de tekniske data management-tjenester som drives af DANS og 3TU.Datacentrum for hele spektret af nationale og disciplinære brugere. Bag tjenestelaget ligger teknisk infrastruktur, storage osv. som også drives på nationalt plan, bl.a. af SURF.

Det tekniske system bag EASY er Fedora Commons repositoriesoftware¹⁰ som også bruges i Danmark, og som bliver yderligere omtalt nedenfor. DANS ønsker at sikre tilgængeligheden af data mange år frem i tiden og arbejder med en overkommelig model for certificering der hedder Data Seal of Approval¹¹.

Mens EASY er et eksempel på en udfoldet national infrastruktur, så pegede vores erfaringer fra UK¹² på at man her i højere grad så de enkelte institutioner som ansvarlige for at opfange, arkivere og tilgængeliggøre forskningsdata. Her bruger man især universiteternes institutional repositories og fremhæver at når man allerede har etableret en interaktion med forskerne, som jo deponerer deres publikationer, så er skridtet ikke langt for at få dem til at deponere datasættene i samme ombæring. Der vil nedenfor blive gennemgået et praktisk forsøg med at deponere datasæt til en institutional repository-løsning.

Ud over hhv. nationale og institutionelle repositorieløsninger, er det velkendt at man inden for en række fagområder og specifikke forskningsfelter foretrækker at deponere til fagspecifikke repositorer som eventuelt knyttes til en værktøjskasse eller et Virtuel Research Environment. Inden for humaniora, kender vi i Danmark til forskningsinfrastrukturer som LARM og Clarin.dk der i en vis forstand kobler datarepository med forskningsværktøjer. Mere internationale initiativer er det europæiske Clarin.eu og et globalt fagrepository som det sociallingvistiske *Talkbank*¹³.

B. Testcase: Curate / Hydramata

Den valgte case vil demonstrere et eksempel på deponering af forskningsdatasæt til en institutional repository-løsning. Vi har valgt at teste det Hydra¹⁴-baserede Curate-system som vi har installeret på DeIC's server. Curate er et open source system¹⁵ som er udviklet som et fællesprojekt af en række amerikanske universitetsbiblioteker¹⁶. Det fælles system er p.t. under udvikling og baserer sig på funktioner fra flere af projektparternes eksisterende løsninger, herunder Penn State's *Scholarsphere* og Notre Dame's *CurateND*.

Vi har måttet teste en ikke-frigivet version 0.6.6 som ikke er driftsegnet, men som det har været muligt at installere på DeIC's skytjeneste¹⁷ og afprøve de implementerede funktioner på. Det

¹⁰ <http://www.fedoracommons.org/>

¹¹ www.datasealofapproval.org

¹² Bl.a. deltagelse i workshoppen *Research Data Management: Finding our Role* på University of Oxford, den 6. december 2013.

¹³ <http://talkbank.org/>

¹⁴ Hydra er et open source rammesystem for udvikling af digital biblioteksapplikationer i Ruby on Rails programmeringssproget. Hydra baserer sig på Fedora repositoryet (se note 8), søgemaskinen Solr og discovery-værktøjet Blacklight. Læs mere på <http://projecthydra.org/>.

¹⁵ Koden kan hentes fra <https://github.com/projecthydra-labs/curate>.

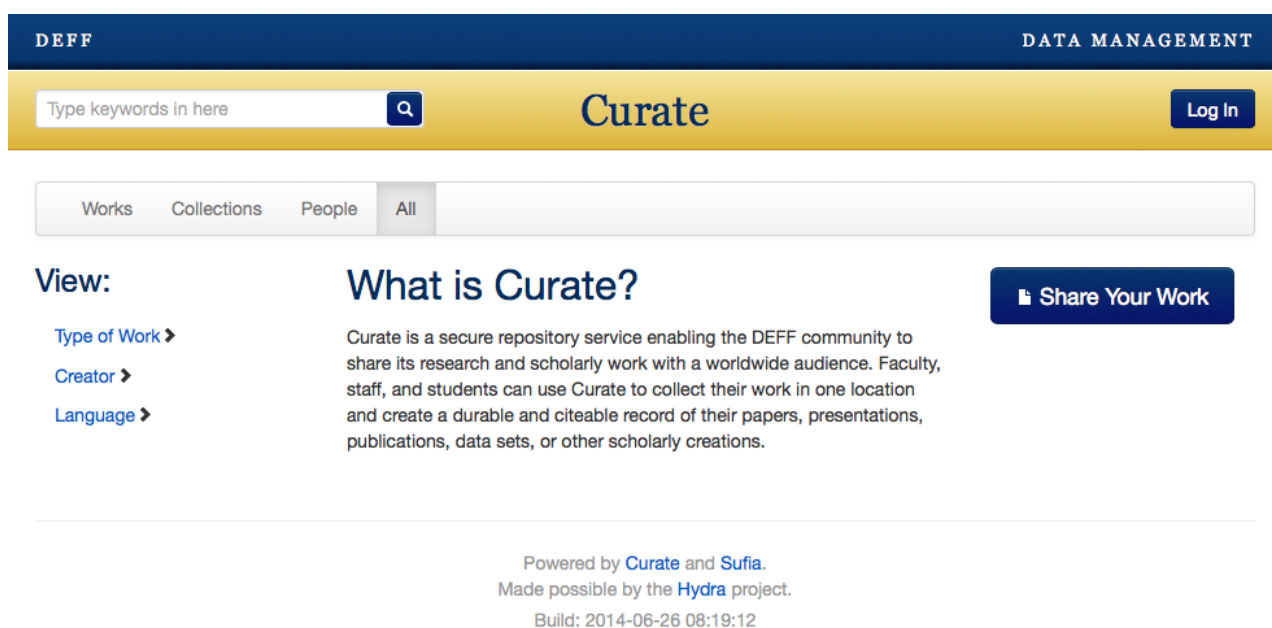
¹⁶ Projektet kaldes nu ofte for Hydramata, deltagere i projektet er universiteterne Indiana, Notre Dame, Northwestern, Penn State, Cincinnati og Virginia.

¹⁷ Tjenesten kører i skrivende stund på hydramata.deic.dk:3000

testede system er amerikansk-sproget og navngivet Curate, dette har ikke været forsøgt ændret, hvorimod vi har ændret enkelte ting så som institutionsnavn idet visse ledetekster og overskrifter ligger i separate sprogfiler. Vi er blevet oplyst om at en kommende frigivet version vil være fuldt forberedt til oversættelse til andre sprog. I testen har vi brugt Curate "out-of-the-box", uden ændringer i selve koden af nogen art.

Webapplikation og deponeringsworkflow

Curate har tre grundlæggende typer af objekter: *Works*, *Collections* og *People*. Dette fremgår af topmenuen på forsiden og alle søgeresultatsider hvor topmenuknapperne bruges til at styre visningen af resultater:



Startsiden er i bund og grund en søgeside hvor man kan søge i alle de dele af repositoret man har adgang til. For at kunne deponere data, må man oprette sig som bruger og logge ind. I testen er der tale om en lokal brugerbase, og man optræder efter at have oprettet sig, som "Person" i systemet. Det fremgår af Curate-projektets README-fil¹⁸ at der kan integreres med identity-services via CAS, LDAP og Shibboleth. Her i Danmark ville Curate således kunne integreres med WAYF.dk. Det fremgår ligeledes af README-filen at der kan etableres adgang via Omniauth til at etablere adgang via forskellige adgangskontrolsystemer (fx Facebook og Twitter som dog næppe er relevant i akademisk sammenhæng).

I testsystemet skal man, for at oprette sig som bruger, godkende et sæt *Terms of Use*. Vi har ikke konstateret hvorvidt dette logges eller om det blot implicit antages at være godkendt, i og med at registreringen som bruger er blevet gennemført. Vi har konstateret at det ikke er muligt at få adgang til systemet som bruger uden at godkende disse betingelser for anvendelsen.

¹⁸ <https://github.com/projecthydra-labs/curate/blob/develop/README.md>

For at kunne uploade et datasæt skal man klikke på *Share Your Work*-knappen fra forsiden - eller på +-knappen fra en hvilken som helst resultatside og bliver, via en login-side, sendt til følgende side hvor man kan vælge mellem de typer af materialer der p.t. understøttes af Curate. Vi formoder at det ville være muligt at indskrænke dette til *Dataset* hvis man vil bruge Curate i en snævrere sammenhæng end et institutional repository, men vi har ikke undersøgt denne mulighed nærmere.

DEFF DATA MANAGEMENT

Type keywords in here

Curate

What are you uploading?

Before we can begin we need to know a little about what you're uploading.

Article

Deposit or reference a preprint or published article.

Dataset

One or more files related to your research.

Document

Deposit any text-based document (other than an article).

ETD

Deposit a senior thesis, master's thesis, or dissertation.

Generic Work

Image

Any image file: art, photograph, poster, etc.

Powered by [Curate](#) and [Sufia](#).
Made possible by the [Hydra](#) project.
Build: 2014-06-26 08:19:12

Klikker man på *Dataset*, kommer der en formular op som rummer 4 dele:

- Indtast metadata om datasættet
- Angiv sti til datasættets fil(er)
- Angiv licens/rettigheder for anvendelse
- Godkendelse af deponeringslicens

De enkelte dele af proceduren vises og gennemgås nedenfor.



Curate

 Anders Conrad  

Describe Your Dataset

The more descriptive information you provide the better we can serve your needs.

Please consider releasing your dataset as an **Open Access** work.

Required Information

* Title

Contributor

Anders Conrad 



Description Please keep your description to 300 words or less.

Content Editors

Who should be able to make changes to this work?

* Editor Individual At least one person should be able to edit this work

Anders Conrad 



Group Add groups to allow others to collaboratively edit this work.



Additional Information

Subject



Publisher



Bibliographic citation



Source



Language



Metadata er et relativt enkelt sæt beskrivende metadata som gemmes i en datastream i Fedora. Det er muligt at oprette grupper af brugere og angive at en gruppe må redigere en fælles resurse.

Næste led i processen er at udpege filer til upload eller som ekstern reference via en URL. På inddateringsbilledet er det kun muligt at vælge den ene af disse muligheder; når man efterfølgende redigerer det oprettede objekt, er det muligt at have begge dele. Der er også en knap til at hente filer fra en af tjenesterne Dropbox, Google Drive og Box. For at hente filer via en af disse skytjenester, skal brugerens ID valideres via den pågældende tjenes API. Vi har ikke set dette virke, og vi er fra Curates udviklingsteam blevet informeret om at denne funktionalitet endnu ikke er færdigimplementeret.

Add Your Content

The subject of your dataset can either be a file or a URL.

Upload a file A PDF is preferred.

Ingen fil valgt.

OR

External link

Please click the button below to upload remote files.

Assign a Digital Object Identifier (DOI)

A DOI is a permanent link to your dataset. It's an easy way for other people to cite your work.

Want more information on DOIs? Here's a [brief summary](#) and the [DOI FAQ](#).

☐ Yes, I already have one that I want to use:

☐ Not now...but maybe later.

Access Rights Who should be able to view or download this content?

Please note, making something visible to the world (i.e. marking this as **Open Access**) may be viewed as publishing which could impact your ability to:

- Patent your work
- Publish your work in a journal

Check out [SHERPA/RoMEO](#) for more information about publisher copyright policies.

☐ **Open Access** Visible to the world.

☐ **Open Access with Embargo** Treated as **Private** until then it is **Open Access**.

☐ **DEFF** Visible to all DEFF users.

☒ **Private** Only visible to you and your delegates.

Det fremgår af koden at Curate har integreret en generel Hydra-funktion til at karakterisere filer ved ingest (ved hjælp af FITS). Vi kan se i koden at resultatet af karakteriseringen gemmes i en separat Fedora datastream der hedder FileContentDatastream. Vi har ikke i praksis verificeret dette eller nærmere kunnet vurdere hvorledes disse informationer bruges af Curate. Det Kongelige Bibliotek har i anden sammenhæng brugt det samme programbibliotek med succes.

Så vidt vi har erfaret, er det kun muligt at uploade en fil ad gangen, og det kan jo være en begrænsning som i givet fald kan kræve at en større samling filer bliver samlet i en zip- eller tarfil¹⁹ inden deponering. En hurtig inspektion af kildekoden tyder på at den bagvedliggende datamodel faktisk tillader datasæt med mange filer. Det kan derfor tænkes at upload af multi-fil datasæt vil blive muligt i en senere version - eller at en sådan funktion vil kunne tilføjes efter behov.

¹⁹ tar og zip er eksempler på almindeligt brugte komprimerings- og arkiveringsformater

Når man uploader filer, er der en ledetekst der siger "A PDF is preferred". Vi formoder at dette er et levn fra Notre Dames kørende CurateND-system. Så vidt vi kan vurdere, er der snarere tale om en politik end om en teknisk begrænsning. Vi er ikke i praksis stødt på nogen begrænsninger i forhold til at uploade forskellige filtyper.

Curate-softwarepakken kan bygges på to måder: med en automatisk tildeling af DOI-er til nyoprettede objekter via en ekstern service - eller uden denne funktion. Vi har bygget vores testversion uden tildeling af DOI-er, og der kommer i dette tilfælde en mulighed op for brugeren til manuelt at tildele objektet en DOI.

Curate er bygget op med fuldstændig kontrol for den deponerende part over hvad materialet omfatter, og på hvilken måde det må anvendes. Man bliver således bedt om at angive rettigheder og slutbrugerlicens for benyttelsen af objektet. Dette kan gå fra privat, hvor kun brugeren selv må tilgå det, over adgang til brugere som er logget ind på repositoriet, til open access, herunder *Private* i en nærmere angivet embargoperiode. Vælger man embargo, vil objektet i embargoperioden blive behandlet som privat, herefter overgår det til Open Access. Vi formoder at dette vil kunne imødekomme forskere som gerne vil dele deres data, men først efter at de selv har fået deres artikler godkendt og publiceret. Det er tilsyneladende muligt at opretholde forskellige adgangsbetingelser til hhv. metadata og filindhold, fx. med *Open Access* metadata for en *Private* fil.

Man kan vælge mellem en række slubrugertilicenser, fra "All rights reserved", en række CC-licenser, sluttende med CC0 for et helt frit dokument. De valgte CC-licenser er amerikanske, vi formoder at denne pull-down menu er konfigurerbar til fx. danske CC-licenser.

Choose a License for your Content

What do you want others to be able to do with your work?

[Here's some help](#) if you don't know which license to choose.

* Content License

All rights reserved

Curate Content Submission Terms

I am submitting my work for inclusion in the Curate repository maintained by the University Libraries of the DEFF. I acknowledge that publication of the work may implicate my legal rights with respect to the work and its contents, including my ability to publish the work in other venues. I UNDERSTAND AND AGREE THAT BY SUBMITTING MY CONTENT FOR INCLUSION IN THE CURATE REPOSITORY, I AGREE TO THE FOLLOWING TERMS:

I hereby grant the DEFF permission to reproduce, edit, publish, disseminate, publicly display, or publicly perform, in whole or in part, the work in any medium at the University's discretion (including but not limited to public websites). I agree that the University can keep more than one copy of the submission for the purpose of preservation and security. I agree that the University can preserve the submission by migrating or translating it to a new format or medium as needed in the future. I also agree that the metadata attached to the item can be reviewed and altered by the University to aid in preservation and discovery. I understand that I may be allowed the opportunity to select the intended audience for the materials that I submit, and I agree that I am fully responsible for any claims and all responsibility for materials submitted. The Curate service is offered as-is with no warranties, express or implied. The University may suspend or terminate Curate, or remove any content within the system, at any time for any reason in the University's sole discretion.

I warrant that the submitted material is original to me and that I have power to make this agreement. I also warrant that the submission does not, to the best of my knowledge, infringe upon anyone's copyright. I also warrant that if the work has been previously published elsewhere in whole or in part, that I have obtained the permission of the copyright owner to grant Curate the rights required by this license. I also guarantee that I do not have any other publication agreements that involve this material or substantial parts of it that conflict with my submission of materials for dissemination in Curate.

I represent that any materials that are used in this submission that I do not hold copyright for, I have obtained permission to use and display the materials according to the rights required by this license, and that third-party owned materials are clearly identified and credited within the content of the submission. If it is determined that permissions of use have not properly been obtained, I acknowledge that the University may remove or restrict access to items as necessary.

I understand that I may request the University to remove my submitted materials from the Curate repository; however, I acknowledge that the University cannot control or retract works that may have been accessed by third parties prior to my request for removal. If the submission is removed I agree that the item can be replaced by a page with a statement declaring that the item was removed by my request.

☐ I have read and accept the contributor license agreement

Create Dataset

Cancel

Endelig skal man, når man deponerer, i et afkrydsningsfelt markere at man har læst og godkendt deponeringslicensen. Ligesom med de generelle betingelser for brug, har vi ikke undersøgt om denne godkendelse logges eksplicit. Vi har derimod konstateret at det ikke er muligt at deponere datasæt eller andre resurser uden at godkende deponeringslicensen. Vi formoder at selve teksten til licensen kan findes i koden og tilrettes efter behov, men har ikke undersøgt dette nærmere.

Nu kan man trykke på "Create Dataset" hvorved objektet bliver bygget/persisteret i Fedora med sine metadata, og filer uploades.

Lydfiler (Dataset)



Attributes

Attribute Name	Values
Contributors	Anders Conrad
Subject	Computermusik
Access Rights	Open Access with Embargo
Embargo Release Date	2014-06-28
Content License	Attribution-NonCommercial 3.0 United States

Files

File	Filename	Date Uploaded	Visibility	Actions
 SFDi.zip	SFDi.zip	2014-06-27	Open Access with Embargo	Edit Rollback Delete Download

External Links

This Dataset has no external links associated with it. You can add one using the "Add an External Link" button below.

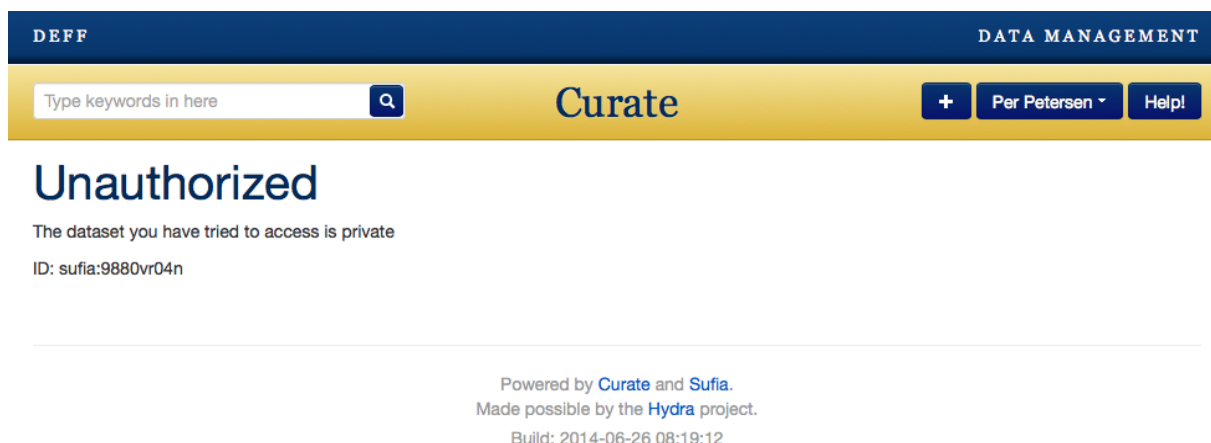
Related Works

Add works that are related to this Dataset.

[Update Related Works](#)[Edit This Dataset](#)[Delete This Dataset](#)[Attach a File](#)[Add an External Link](#)

I det efterfølgende skærbillede som svarer til den "landing page" man får ved at klikke ind på objektet fra en søgning, har man som *depositor* mulighed for, dels umiddelbart at redigere filindholdet, dels at knytte det netop skabte datasæt sammen med andre resurser med knappen "Update Related Works". Denne funktion er aktiv inden for repositoriet, men kan tilsyneladende ikke linke ud, her skal man bruge "External Links". Dette må anses at være en oplagt mulighed for at linke en artikel i repositoriet til det bagvedliggende datasæt. Related Works funktionen giver mulighed for type-ahead søgning på titler i repositoriet, og der knyttes relationer begge veje.

Referencer er synlige, også til objekter som bruger ikke har adgang til. Når man klikker på et link til en beskyttet resurse, får man:



Som det fremgår, er der efter deponering mulighed for den deponerende person for at redigere i såvel metadata som dataindhold. Vi formoder at det underliggende Fedora-system gemmer de forskellige versioner af objektet, men har ikke fundet nogen tegn på at versionering indgår i Curates brugergrænseflade. Efter vores erfaring, kan man alene tilgå den senest gemte version, det fremgår ikke om der findes eventuelt tidligere versioner, og URI'en til objektet ændrer sig ikke når der gemmes nye versioner.

Vi har ikke set det demonstreret i den testede udviklingsversion, men ved fra anden side at en integration med ORCID forsker-id tjenesten er på trapperne.

C. Andre deponeringstjenester

Vi har vist Curate som en slags prototype på hvad et institutionelt repositorie vil kunne tilbyde forskere og forskningsinstitutioner med hensyn til deponering efter endt forskningsprojekt. Vi har ikke inden for rammerne af projektet haft muligheder for yderligere test, men vil alligevel rette opmærksomheden på nogle alternative muligheder.

Dspace²⁰ er en meget benyttet open source institutional repository softwarepakke som, ligesom Fedora, vedligeholdes og distribueres af non-profit organisationen Duraspace²¹. Edinburgh University's EDINA²²-repositorie er et eksempel på en løsning der er baseret på en tilrettet Dspace-version. Islandora er en anden open source institutional repository-løsning som, ligesom Hydra og Curate, baserer sig på Fedora. Også inden for Islandora-verdenen ser der ud til at være bevægelse omkring håndteringen af forskningsdata. Vi er ikke, ud fra en hurtig undersøgelse på nettet, i stand til at fastslå i hvor høj grad sådanne funktioner er integreret i selve grundprodukterne, eller om funktioner der retter sig mod deponering af forskningsdata, er applikationer eller tilrettelser som brugerinstitutionerne selv implementerer.

²⁰ <http://www.dspace.org/>

²¹ <http://www.duraspace.org/>

²² <http://edina.ac.uk/>

Et alternativ til institutionelle repositorier, der ofte ligesom Curate dækker en vifte af materialetyper, kan være mere frittilgængelige repositorier i skyen, som til gengæld kun kan bruges til forskningsdata.

Harvard Dataverse Network²³ er et eksempel på en sådan løsning som tilbydes til videnskabsfolk i hele verden. Løsningen understøtter bl.a. versionering af datasæt, metadatering og tildeling af persistente id'er, citation af datasæt²⁴ og fleksibel adgangsstyring. Harvards Dataverse Network baserer sig på et open source-projekt hvorfra koden frit kan hentes og installeres lokalt. Det Kongelige Bibliotek har igennem flere år brugt denne software til at understøtte forvaltning, publikation og citation af egne datasæt for forskere fra Københavns Universitet under titlen *KUBIS Databox*.²⁵

En anden generel skyløsning er Dryad Digital Repository²⁶ som tilbyder forskere frit at deponere og dele videnskabelige datasæt. Dryad er en non-profit organisation hvor organisationer som universiteter og forlag betaler en vis pris for at være medlem hvorimod tjenesten er gratis at benytte for forskere. Noget tilsvarende gælder tilsyneladende Figshare²⁷ som dog drives af en kommerciel virksomhed.

D. Opsummering

Vi har vist at det kan lade sig gøre at deponere udvalgte forskningsdatasæt, organiseret som enkle filer, ved hjælp af Curate-softwarepakken som stadig er under udvikling. Der er tale om et meget simpelt og tilsyneladende rent forskerdrevent workflow hvor forskeren deponerer og efterfølgende kan redigere og evt. slette sine datasæt, både hvad angår metadata, filindhold, referencer og adgangsforhold. Tjenesten vil typisk basere sig på en validering af brugeren som tilhørende det akademiske samfund, fx. via Wayfd.dk, men har ikke umiddelbart indbygget yderligere kontrol eller godkendelse af indholdet.

Curates potentiale som deponeringssystem er en velfungerende og enkel grafisk brugergrænseflade, sammenkoblet med at alt dataindhold gemmes i et Fedora-repositorie. Da der er tale om open source software, vil det være overkommeligt at ændre applikationen, fx ved at tilføje mere komplekse workflows, begrænse materialetyper, tilføje versionering, tillade flere filer pr. objekt eller andet. Det vil dog også være muligt ved hjælp af Hydra rammesystemet at genbruge nogle af Curates grundkomponenter til en helt ny applikation, målrettet fx. national deponering af forskningsdata.

Fra et teknologisk synspunkt er vi ikke stødt på afgørende forskelle mellem institutionelle og nationale repositorier, jvf. det nationale hollandske DANS EASY der, ligesom Curate, er Fedora-baseret. I tilfældet Curate, ville man godt kunne forestille sig flere forskellige lokale varianter af applikationen kørende op imod et samlet nationalt Fedora repository. Det Kongelige Bibliotek har

²³ <http://thedata.harvard.edu/dvn/>

²⁴ I det Kongelige Biblioteks installation baseret på Handle systemet. Se også kapitel 6.

²⁵ Se <https://data.kb.dk/dvn/>

²⁶ <http://datadryad.org/>. Dryad drives angiveligt af North Carolina State University.

²⁷ <http://figshare.com/>

demonstreret at softwaren fra Harvard Dataverse kan deployeres til en institutionel løsning, og den ville ligeså kunne deployeres nationalt. Derimod skiller de mere generelle skytjenester sig ud med et langt bredere sigte, både i forhold til organisationer og faglige discipliner.

Kapitel 5: Data repositories (efter forskningsprojektet, langtidsbevaring)

Mens vejen fra deponering af aktive forskningsdata og til arkivering i et repository kun har kunnet antydes i denne rapport, er der tilsyneladende kort fra hvad vi har kaldt den kuraterede deponering efter projektet og til arkivering og eventuelt langtidsbevaring i et repository. I Curate-eksemplet fungerer Fedora som dette backend-repository.

Det diskuteres lidt i repositoriesammenhæng hvad begrebet bevaring dækker over. For nogen vil det være opfyldt hvis et datasæt bliver forsynet med beskrivende metadata og placeret i et repository som drives af en organisation med i hvert fald nogle få års fremtidig drift sikret. Hvis alternativet er at data slettes ved forskningsprojektets afslutning, så er alene deponering til et repository et udtryk for bevaring. For andre betyder bevaring certificerede trusted repositories som opfylder en række policy-krav, forankres i stabile og langtidsholdbare organisationer. Bevaring kobles her ofte med specialiserede værktøjer, specielle metadata og funktioner der kan sikre dataintegritet, flere kopier, formatmigreringer, bevaringsplaner osv.

Curate-eksemplet som blev gennemgået ovenfor, og øvrige institutional repository-løsninger, opfylder umiddelbart det lidt enklere krav om at data bliver arkiveret i et repository som drives af en troværdig organisation. De vil typisk være stabile og kunne sikre data i en årrække fremover under normale driftsforhold. Der vil være backup og almindelig best practice system management for den bagvedliggende infrastruktur, og data vil kunne være tilgængelige i præcis samme form som de blev deponeret, i en årrække frem.

I Danmark er Statsbiblioteket, Statens Arkiver og Det Kongelige Bibliotek gået sammen om at skabe Det Nationale Bitmagasin²⁸ som en ramme om en højkvalitetsbevaring af de bitstrømme som er kernen i alle data. Vi må formode at nogle, men ikke alle forskningsdata vil have en værdi som berettiger til at blive bevaret i en løsning som denne. Der må dog rimeligvis være mulighed for at graduere bevaringshorisont og graden af aktiv indsats, afhængigt af de enkelte forskningsprojekter, deres relevans for fremtiden og de økonomiske muligheder.

Her vil vi pege på muligheden for at integrere en Hydra-baseret repositoryløsning med det nationale bitmagasin. En sådan integration er blevet udviklet af Det Kongelige Bibliotek som led i et igangværende digital infrastruktur-projekt. Funktionaliteten styres fra et skærm billede i bibliotekets forvaltningsapplikation hvor en kurator kan definere bevaringsniveau for et givet objekt. Forvaltningssystemet sender herefter objektets data til et separat integrationsmodul som pakker filerne for objektet i WARC-format og deponerer pakken i bitmagasinet. Det vil ligeledes blive muligt at importere objekter tilbage fra bitmagasinet til forvaltningssystemet.

På nuværende tidspunkt er der integreret værktøjer til checksum-beregning og karakterisering i Hydra applikationen. På længere sigt er det planen at integrere flere bevaringsrelaterede værktøjer i Hydra på en måde så disse bliver stillet til rådighed for alle Hydra-applikationer. Det vil således være muligt at udvide fx. Curate med de bevaringsfunktioner som udvikles i Hydra-projektet i øvrigt, herunder på Det Kongelige Bibliotek.

²⁸ <http://digitalbevaring.dk/det-nationale-bitmagasin/>

E. Opsummering

Vi har, ud fra en generel institutional repository-applikation (Curate) sandsynliggjort at der med relativt enkle midler kan skabes repositorier til en velordnet metadatering og arkivering af forskningsdata, både på institutionelt og på nationalt plan. Ud fra almindelig best practice for it-drift vil en sådan repositorielløsning kunne anses for at være en robust arkiv over de deponerede data. Der er intet til hinder for at de bevaringsværktøjer som tænkes integreret med Hydra ikke også vil have værdi i forhold til fx et institutionelt eller nationalt Fedora-repositorie. Noget lignende må formodes at gøre sig gældende for nogle af de øvrige repositorielløsninger vi kort har nævnt i kapitlet ovenfor.

Begrænsningerne i en sådan løsning vil træde frem hvis man tænker bevaring på meget lang sigt, eller af videnskabelige data af særlig værdi som kræver specielle bevaringsfunktioner gennemført. Her er der behov for en mere kompleks infrastruktur som fx. i Det Nationale Bitmagasin. Det Kongelige Bibliotek har i anden sammenhæng vist at man kan integrere en Hydra-baseret løsning som Curate med bitmagasinet. Tilsvarende løsninger må formodes at kunne etableres fra andre open source repositorie-software pakker, afhængigt af hvorledes både koden og community omkring softwareproduktet er organiseret.

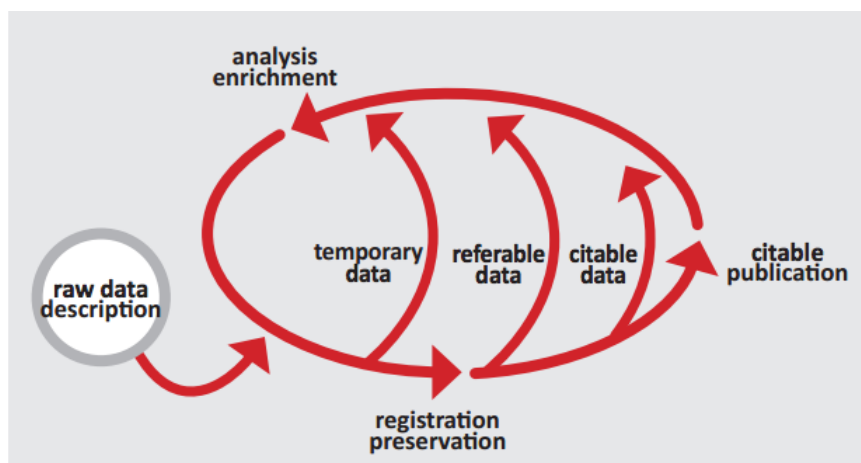
Det anses for tvivlsomt om langtidsbevaring af aktive forskningsdata som høstes direkte og automatisk (behandlet i kap. 4) vil kunne foretages uden nogen form for udvælgelse og kuratering i øvrigt. Med tiden kunne man forestille sig en eventuel national infrastruktur udvikle en tilstrækkelig sofistikeret workflow-håndtering til at dele af en sådan kuratering vil kunne foregå automatisk. En undtagelse ville være datasæt hvor det forlods er besluttet at alle data skal langtidsbevares i fx Det Nationale Bitmagasin, et sådant scenarie ville formentlig relativt enkelt kunne etableres allerede i dag.

Kapitel 6: Data Catalogues - om publicering, identifikation, genfinding, citering og meritering af forskningsdata

A. Indledning

Dette kapitel omhandler det sidste trin i den cyklus, der er valgt som struktur for rapporten: "... → *Data management planning* → *Managing active data* → *Data selection and handover* → *Data repositories* → *Data catalogues* → ..."

Mens denne data management cyklus er en udmærket abstrakt struktur til at disponere et overblikssnotat, er det vigtigt samtidig at notere sig, at man i den "virkelige forskningsverden" sjældent vil opleve at man fuldt og helt gennemløber denne cyklus med dens fem successive trin i fast rækkefølge. I overensstemmelse med den valgte cyklus behandles identifikation af data (DOI mv.) som en del af dette sidste trin, mens det ofte vil være tilfældet, at et datasæt-DOI udtages meget tidligere i processen, og at datasættets komponenter og dets beskrivende metadata gradvist vokser og udvikles i løbet af forskningsprojektet. Der er i realiteten ofte tale om mange rekursive iterationer, således som figuren nedenfor illustrerer. Mange pile der går bagud i diagrammet og mange små eller større rekursive cykler.



"Data catalogues" er her et teknisk stikord for et bredere sæt af udfordringer i forbindelse med at stille forskningsdata til rådighed for andre og mere præcist at sikre, at data kan:

- Findes eller opdages på enkel vis (kataloger, søgemaskiner, tidsskrifter mv.)
- Refereres til på en præcis måde (identificering – DOI'er)
- Medvirke til meritering af de aktører, der har etableret datasættet og stillet dette til rådighed for andre

Kapitlet er således disponeret i følgende underafsnit – hver med et kortfattet overblik over de mest iøjnefaldende tendenser på området:

- Publicering af data
- Identifikation af data
- Genfinding og genbrug af data
- Citering og meritering af data

B. Publicering af data

Der er primært tre måder til at publicere sine data på:

- At publicere datasættet selvstændigt via en af de fremspirende data journals
- At publicere datasættet som en del af publiceringen af projektets resultater i et tidsskrift el.
- At publicere datasættet i et repository, der lægger vægt på formidling, ikke kun på bevaring

”Data journals” er en kanal i vækst – om end fra et lavt udgangspunkt. Eksempler er tidsskrifterne: Geoscience Data Journal, Earth System Science Data, Ecological Archives Data Papers, Hindawi Dataset Papers in Agriculture/Biology/Chemistry/Ecology/Geosciences/ Materials Science/Medicine/Nanotechnology/Neuroscience/Pharmacology/Physics, Journal of Chemical and Engineering Data, GigaScience, Journal of Physical and Chemical Research Data, Biodiversity Data Journal, F1000 Research, CODATA's Data Science Journal, Journal of Open Archaeology Data. Som det fremgår, er STM-området (ikke overraskende) bedst forsynet med tilbud på dette område, mens humaniora og bløde samfundsfag (endnu?) ikke benytter denne tilgang – med ærkæologi som den lige så typiske undtagelse.

Et prægnant eksempel på et sådant tidsskrift er Nature Publishing's ”Scientific Data”, der blev lanceret sidste år under stor opmærksomhed (<http://www.nature.com/sdata/>):



Det australske data management initiativ ANDS vedligeholder en nyttig guide: ANDS Data Journal Guide (<http://ands.org.au/guides/data-journals.html>)

Mange "klassiske" tidsskrifter giver mulighed for at beskrive og linke til datasæt som en del af (eller bilag til) artiklen, der beskriver forskningsresultaterne i.e. de konklusioner, der kan udledes af de genererede og analyserede data. Dette er ofte et tilbud til forfatterne, men i stigende grad eskaleres dette til et krav – en forudsætning for artiklens peer-review og eventuelle publicering. Fremtrædende eksempler på dette er Nature Publishing's Policy on availability of data and materials:

*"An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in a Nature journal is that **authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications**. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must **also** be disclosed in the submitted manuscript, including details of how readers can obtain materials and information."*

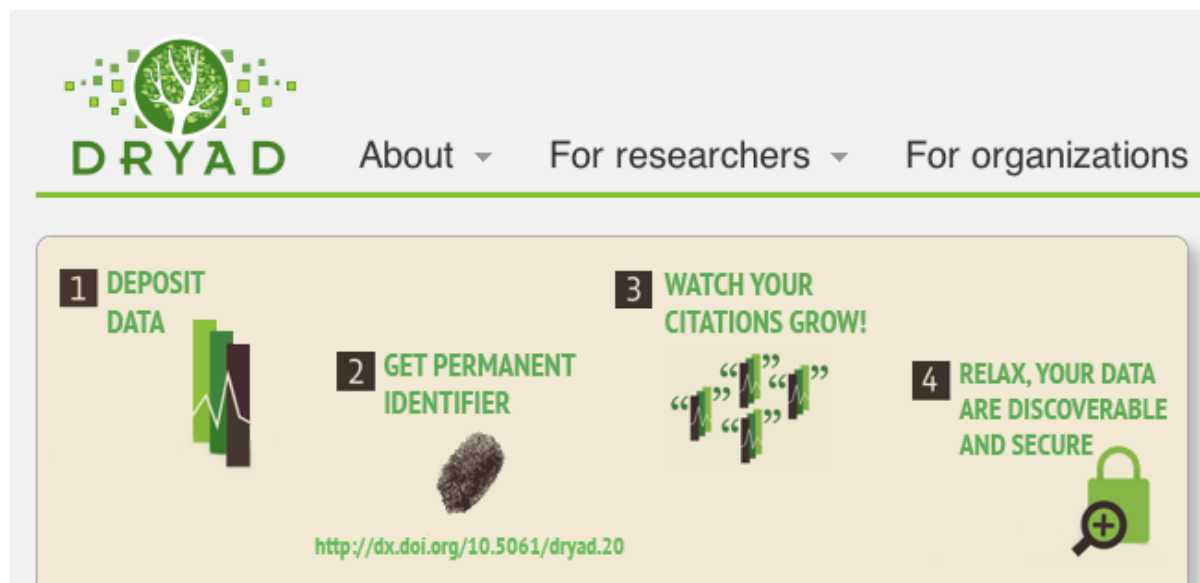


Samt verdens største tidsskrift PLOS One's Policy on sharing of data, materials, and software:

"PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception. When submitting a manuscript online, authors must provide a Data Availability Statement describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the final article. Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from

articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication."

Data repositories bruges også som en publiceringskanal i de tilfælde, hvor det valgte repository sørger for DOI-identificering af datasættene og aktivt søger at formidle datasættene – f.eks. gennem samarbejde med nettets søgemaskiner og relevante fora for videnskabelig diskussion.

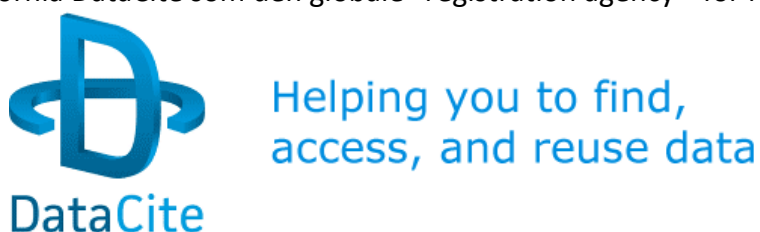


DRYAD (a nonprofit repository for data underlying the international scientific and medical literature) udgør et eksempel på en vellykket symbiose mellem tidsskrifter og repositories. DRYAD fungerer som fælles data repository for pt. 73 life science tidsskrifter, der har integreret deres manuscript submission systems med DRYAD's data submissions system, hvorved arbejdet lettes for forskeren, der således får lettere ved at leve op til policies som de ovennævnte fra PLOS og Nature.

C. Identifikation af data

Som det allerede fremgår ovenfor er korrekt, præcis og permanent identifikation af data et væsentligt omdrejningspunkt for alle tiltag til deling af data, hvad enten det drejer sig om at arkivere, publicere, finde, citere eller genbruge data.

Digital Object Identifier (DOI) systemet har for længst etableret sig som best practice for den digitale forskningslitteratur, hvor forlagene i fælleskab driver CrossRef som "registration agency" for deres tidsskriftartikler mv. Inspireret af dette, men koblet med et ønske om at holde DOI-systemet for forskningsdata under forskningsmiljøernes egen kontrol, lancerede en gruppe europæiske tekniske universiteter og biblioteker i samarbejde med bl.a. British Library, ANDS og University of California DataCite som den globale "registration agency" for forskningsdata.



DataCite har siden lanceringen i 2009 (med DEFF's støtte og deltagelse) udviklet sig i retning af global dækning og deltagelse og har et godt samarbejde med CrossRef under den fælles paraply for alle registration agencies, "The International DOI Foundation". En DOI resolveres (omsættes til

en virksom URL) således altid på samme måde - uanset om den en er udstedt i regi af forlagenes CrossRef-samarbejde eller forskningsmiljøernes DataCite-samarbejde.



DRYAD, nævnt ovenfor, er en eksponent for samarbejdet mellem (typisk kommercielle) forlag og data repositories under kontrol af det videnskabelige samfund – på mere end én måde. Ud over ovennævnte symbiose gælder det, at artiklerne tildeles DOI i CrossRef regi og datasættene i DataCite regi.

Ud over ønsket om at holde dele af denne centrale infrastruktur under det videnskabelige samfunds kontrol, er en anden baggrund for DataCite ønsket om holde driftsudgifterne nede på et minimum. Ikke mindst udgiften pr. DOI, da der vil blive tale om et meget stort antal DOI'er, der skal virke over meget lang tid. Derfor er DataCite's model for omkostnings-dækning baseret på nationale medlemskaber med fast kontingent, uanset antallet af registrerede DOI'er. Det nationale medlemskab og udgifter deles dernæst mellem landets deltagende institutioner eller datacentre. DTU Bibliotek har siden grundlæggelsen koordineret og administreret Danmarks medlemskab, se <http://www.datacite.dtu.dk/>.

DataCites infrastruktur udbyder tre primære services

- Registrering af metadata og tilknytning af DOI ("DOI minting") via API eller web-UI
"DataCite requires only a small number of descriptive elements in order to register a resource for a DOI: the creator(s), title, publisher and publication year"
- Global søgemaskine for forskningsdata på basis af de registrerede metadata
<http://search.datacite.org/ui>
- Data tjeneste (OAI provider) der frit stiller de registrede metadata til rådighed for indeksering i andre søgemaskiner, således at datasættene kan findes mange steder og på mange måder.

D. Genfinding og genbrug af data

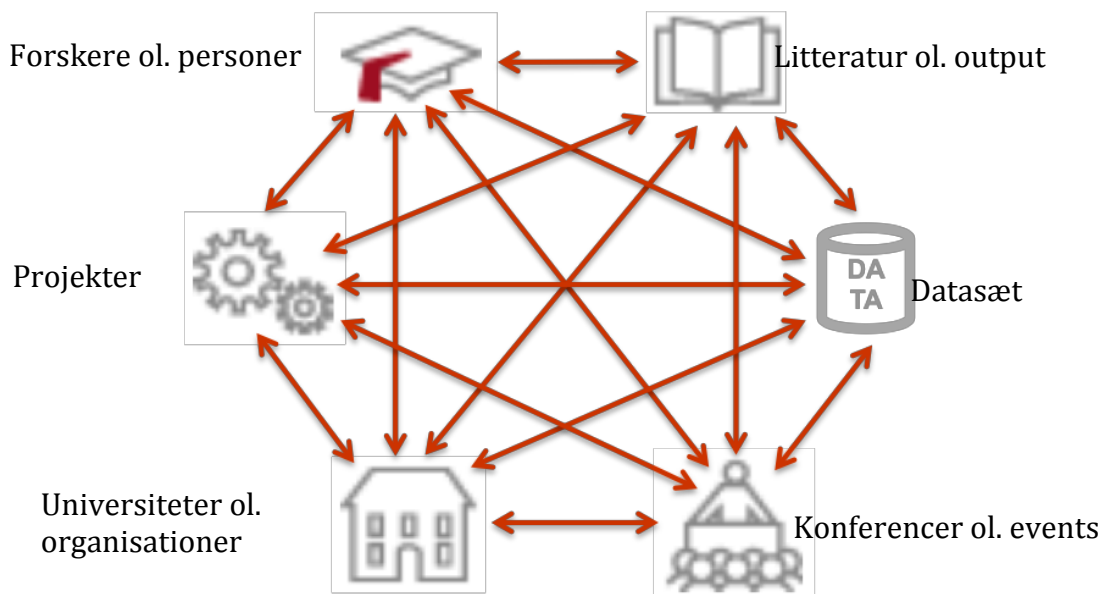
At finde (eller på anden måde opdage) et relevant datasæt er naturligvis første betingelse for at kunne drage nytte af det – herunder evt. at genbruge det i en eller anden udstrækning.

I en række tilfælde vil de forskellige tidsskriftrelaterede tiltag nævnt ovenfor under B. være en optimal løsning. De sætter datasættene direkte i kontekst med de videnskabelige artikler, der beskriver baggrund og metodik for forsøg og resulterende datasæt sammen med de observationer og konklusioner, de har givet anledning til. Fra et snævert datagenbrugs-synspunkt kan artiklen betragtes som et fyldigt sæt deskriptive metadata for selve datasættet. I tillæg gælder det, at tidsskriftartikler er velrepræsenterede og velindekserede i utallige søgemaskiner og bibliografiske databaser.

I andre tilfælde kan man ikke forlade sig på at datasættene er beskrevet i eller genstand for en tidsskriftartikel. Her må man benytte databaser og søgemaskiner, der direkte åbner for søgning i indholdet af relevante data repositories. Da der i realiteten kun søges i metadata om datasættene (ikke datasættene som sådan) bruger man termen "data catalogue" for den søgemæssigt interessante metadatadel af repositoryet, der foruden tekst rummer links til selve datasættene. Dette er en analogi med "library catalogue" der ligeledes er en metadatasamling med links til den fulde tekst, der som regel håndteres af andre systemer. I fortsat analogi kan man tale om datakataloger, der ikke er en fast bestanddel af et enkelt repository, men snarere udmærker sig ved at rumme søgbar metadata på tværs af mange repositories. DataCites globale søgetjeneste eller datakatalog er et prægnant eksempel på dette.

Afhængig af brugssituation og behov kan det være relevant at søge globalt i alle tænkelige datakataloger, nationalt i alle danske datakataloger eller lokalt i universitetets eget datakatalog. For en global søgning vil man typisk vende sig til Google, der også dækker DataCites datakatalog, der til gengæld kan anvendes af dem, der er villig til at ofre noget i bredden af søgningen for til gengæld at få færre resultater med en højere anvendelighed.

På nationalt og lokalt plan er vi i Danmark i en særlig gunstig situation grundet Den Danske Forskningsdatabase (DDF) og det netværk af lokale forskningsdatabaser, der er vokset frem på universiteter mv. og som bygger på DDF's datamodel og suite af dataformater – og som løbende opdaterer den nationale søgetjeneste og portal gennem automatiserede dataoverførsler.



Som det fremgår, omfatter datamodellen dels en række objekter i forskningslandskabet og dels (og vigtigst) relationerne mellem disse objekter. Dette sikrer at konteksten er præcist struktureret og dokumenteret, hvorved metadateringsopgaven for det enkelte objekt lettes betydeligt.

For et datasæt er der således tale om følgende (potentielle) relationer til:

- Den litteratur, der baserer sig på datasættet
- Den eller de forskere mv., der har ansvar for/bidraget til datasættet
- Det projekt, der har givet anledning til/finansieret datasættets dannelse
- Den organisation, der ejer eller huser datasættet
- Den conference el., hvor datasættet er blevet præsenteret

De fleste lokale forskningsdatabaser er i Danmark af mærket PURE, der i den seneste version er klar til at modtage beskrivelser af datasæt. Forventeligt vil dette i de kommende år blive udnyttet til at generere lokale datakataloger, der så kan høstes og synliggøres i den nationale forskningsdatabase – inklusive de ovennævnte relationer, der sikrer den afgørende kontekstualisering af datasættet. Herved bliver det enkelt at finde (relevante) danske datasæt – hvor de lagres og langtidsbevares er en anden sag (se kapitel 5).

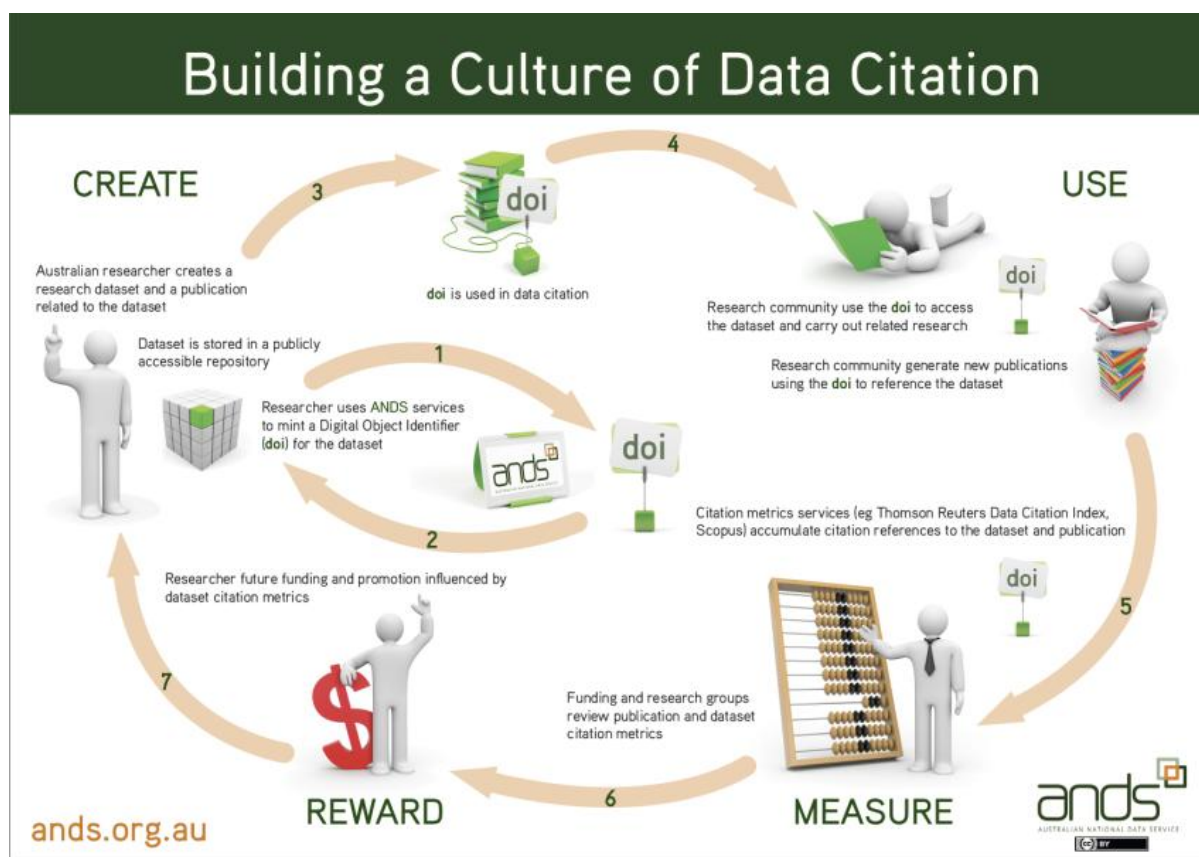
E. Citering og meritering af data

Mens alle i princippet er enige om, at transparens omkring og deling af forskningsdata generelt er "best practice", er realiseringen af denne åbenhed en noget trægere affære. Dette skyldes (for de fleste fagområder) fraværet af motiverende faktorer, der giver anledning til at igangsætte det merarbejde, det er at dokumentere og dele sine data.

I forskningsverdenen er citering af litteraturen den helt store kilde til anerkendelse og dermed faglig succes. Der arbejdes derfor verden rundt på at udbrede denne gode praksis og motiverende faktor til datasæt også. Fremtrædende eksempler herpå er:

- Force11 ("A growing community that aims to transform and improve the future of research communications and e-scholarship") har publiceret en Joint Declaration of Data Citation Principles, se <https://www.force11.org/datacitation>.
- Thomson-Reuters har lanceret et Data Citation Index som en del af Web of Science platformen, se <http://thomsonreuters.com/data-citation-index/>.
- RDA, Research Data Alliance, har som en af sine tidlige aktiviteter startet en Working Group on Data Citation, se <https://rd-alliance.org/group/data-citation-wg.html>.

Ligeledes tilbyder ANDS fra sit eksemplariske web site såvel en Guide to Data Citation som et Data Citation Toolkit, se <http://ands.org.au/cite-data/index.html>. Nedenstående illustration stammer herfra, se http://ands.org.au/guides/data_citation_poster.pdf.



Kapitel 7: Guidance, training and support (værktøjer til understøttelse og uddannelse)

Data management understøttes i forgangslande af en række informations-, trænings- og planlægningsværktøjer - eksempelvis læringsplatforme, tutorials, skabeloner, vejledninger, checklister eller andre digitale formidlingsservices. Følgeligt præsenteres et udvalg af disse, herefter benævnt værktøjer.

Til trods for bredden i værktøjernes funktionaliteter og formål har vi forsøgt at sammenligne dem skematisk (jf. bilag 2). Det er intentionen af den strukturerede sammenholdning skal kunne bidrage til vurdering af hvilke(t) materiale(r) der synes oplagte at afprøve, implementere eller tilrette i et konkret data management supportforløb. Eksempelvis kan man med skemaet relativt nemt danne sig et overblik over hvilke målgrupper eller licensbetingelser der karakteriserer værktøjerne.

Værktøjerne er analyseret i en periode af 3 måneder fra slutningen af 2013 til begyndelsen af 2014. Da værktøjer er under konstant opdatering/forandring, giver vores sammenligning selvsagt kun et øjebliksbillede af det, der var tilgængeligt i undersøgelsesperioden. Det vil derfor være en væsentlig opgave, at vurdere nytilkommen funktionalitet og formål såfremt det ønskes at arbejde indgående med dem.

De undersøgte værktøjer er udvalgt på baggrund af hvad der synes at være state of the art i guidance, training og support fra UK, USA, og Holland, hvor data management generelt er institutionaliseret i højere grad end i Danmark. I DEFF rapporten Forvaltning af forskningsdata²⁹ er især disse landes tilgang til data management blevet fremhævet som inspirerende. Man kan over en bred kam betegne disse producenter som first movers inden for data management, informations-, trænings- og planlægnings-værktøjer.

Størstedelen af værktøjerne medtaget i denne afdækning er blevet til i samspillet mellem en række partnere - hyppigst universiteter, universitetsbiblioteker og nationale lærings-, teknologi- og undervisningsfremmede organisationer. På trods af værktøjernes nationale og kontekstspecifikke ophav, tilpasning og implementering, ligger det ikke inden for rammerne af indeværende projekt at beskrive de organisatoriske dimensioner og konkrete anvendelser³⁰. Derimod fokuseres på værktøjernes egne egenskaber og potentialer for danske interessenter.

De undersøgte værktøjer/materialer og organisationer er:

²⁹ Andersen et al. (2013) "Forvaltning af forskningsdata". DEFF rapport, s. 28.

³⁰ Der findes evalueringer der har fokus på netop disse udfordringer. Se fx Molloy, L. (2012). JISC Research Data MANTRA Project at EDINA, Information Services, University of Edinburgh: Evaluation (2012). Se også: De Smaele, M., Verbakel, E., Potters, N., & Noordegraaf, M. (2013). Data Intelligence Training for Library Staff. *International Journal of Digital Curation*, 8(1), 218-228. Se også: Sallans, A., & Donnelly, M. (2012) DMPonline and DMPTool: Different strategies towards a shared goal. *International Journal of Digital Curation*, 7(2), 123-129.

- *DMPTool*³¹: Planlægningsværktøj udviklet af University of California Curation Center og California Digital Library
- *DMPonline*³²: Planlægningsværktøj udviklet af Digital Curation Centre og JISC
- *MANTRA*³³: Onlinekursus udviklet af University of Edinburgh, EDINA og JISC
- *RDMRose*³⁴: Udviklet af University of Sheffield, Sheffield University library, York University library, Leeds University library samt JISC
- *DataTrain*³⁵: Træningsmateriale og –modul udviklet af Cambridge University library, Cambridge University, Archeology Data Service og JISC
- University of Edinburghs research data management webguide³⁶
- Stanford University library's webportal³⁷ for data management services
- *Data Intelligence 4 Librarians*: Et online/offlinekursus udviklet af 3TU datacentrum og DANS.

A. Vurderingskriterier

Værktøjerne er vurderet og beskrevet på baggrund af følgende kriterier:

- Formål
- Indhold
- Læringstyper
- Anvendte læringsmaterialer
- Data management plan aspekt
- Forskningsdisciplin-specifikt
- Målgruppe(r)
- Didaktik / pædagogik
- Brugervenlighed
- Varighed og omfang
- Licensbetingelser / juridiske forhold
- Producenter og samarbejdspartnere
- Relation til andre værktøjer / services
- Relation til data management livscyklus (jf s. xxx)
- Implementeringspotentialer
- Tilgængelighed
- Sprog
- Overførbare til danske forhold

³¹ <https://dmp.cdlib.org/>

³² <https://DMPonline.dcc.ac.uk/>

³³ <http://datalib.edina.ac.uk/mantra/>

³⁴ <http://rdmrose.group.shef.ac.uk/>

³⁵ <http://www.lib.cam.ac.uk/dataman/datatrain/datatrainintro.html> og <http://archaeologydataservice.ac.uk/learning/DataTrain>

³⁶ <http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/research-data-mgmt>

³⁷ <http://library.stanford.edu/research/data-management-services>

- Tekniske forhold

Hvor nogle af ovenstående kriterier har rent beskrivende karakter, er andre mere et vurderingsspørgsmål. I de henseende hvor der vurderes, er der i skemaet (jf. bilag 2) også redegjort for præmisserne. Hvert værktøj har fået tildelt en nominal bedømmelse i form af en karakter på alle disse parametre, som giver en samlet vurdering af værktøjet til sidst³⁸. Karakteren er et vejledende og tværgående blik på værktøjerne.

B. Værktøjer rettet mod planlægningsfasen

I planlægningsfasen foregår de aktiviteter, der ikke karakteriseres direkte som forskning. Noget af det første, der kan være understøttet af et af de evaluerede værktøjer, er den data management plan (DMP)³⁹ som et stigende antal udenlandske forskningsfonde kræver i forbindelse med ansøgning om midler til projektet. De evaluerede værktøjer henvender sig i høj grad til denne fase - ligesom de fleste også er anvendelige i de senere faser. I forbindelse med DMP er især DMPonline relevant (jf. bilag 2).

En væsentlig pædagogisk kvalitet ved et værktøj af denne type er, at man som forsker tvinges til at tage stilling til en række forhold, der kan have stor betydning for udformningen og udførelsen af forskningsprojektet. Begge de to værktøjer tager udgangspunkt i de krav som de store forskningsfonde i henholdsvis UK og USA stiller. Dermed undgår forskerne at glemme eller overse for fondene vigtige punkter. Derudover kan planerne også komme forskerne selv til gavn, idet man kan bruge DMP'er til at forberede sig på hvordan man vil håndtere projektets data, og dermed forebygge senere problemer eller misforståelser, der kan have udsættende virkning på projektet. Også retslige, adgangs- og formidlingsspørgsmål tvinges man til at tage stilling til og dermed er det ideelt set en enig gruppe, der kan påbegynde projektet, og man undgår usikkerhed og diskussioner senere hen. Værktøjerne udgør også en konkret indgang til at udbyde supporterende services og formidlingsprodukter målrettet forskere.

I DMPonline og DMPTool lægges der op til at DMP er et dynamisk produkt, som dels først fuldendes senere i projektet, dels kan ændres efterhånden som projektet skrider frem, og uforudsete hændelser indtræffer. Alene det faktum at forskerne ikke i pre-award fasen vil udfylde hele DMP'en, hvilket ville være spildt arbejde, hvis projektet ikke bevilges midler, viser, at DMP'er kan have et liv under hele projektet. Hertil kommer at de betingelser og forudsætninger som projektet lever under, kan ændres, hvorfor man må ændre i planen. Begge værktøjer indeholder gode vejledninger til hvordan man anvender dem, samt eksempler på DMP'er.

Til brug i planlægningsfasen eller endda før, er der evalueret en række andre værktøjer. Disse henvender sig dels til forskere, men også biblioteksansatte og andre, der vejleder eller supporterer forskere i data-sammenhæng, kan drage nytte af dem. Hele ideen og konceptet bag DMP'er forklares på udmærket vis i de undervisningsmaterialer, der er evalueret. MANTRA er et selv-

³⁸ To personer har, uafhængig af hinanden, vurderet værktøjerne. De 2 personers bedømmelse af hvert enkelt værktøj er lagt sammen. Det er således muligt for et værktøj at score mellem 2 og 6 point, hvor 6 er bedst.

³⁹ http://en.wikipedia.org/wiki/Data_management_plan

læringsværktøj, der gennemgår hele forskningslivscyklussen, herunder altså også planlægningen. I denne forbindelse forklarer MANTRA om god praksis i almindelig datahåndtering, om de generelle regler i forhold til *responsible conduct of research* og sætter kursisterne i stand til at udfærdige DMP'er. Forud for det er der et modul, der forklarer hvad forskningsdata i det hele taget er. Dermed er kursusmodulerne sat sammen så både kandidat- og Ph.d. studerende samt etablerede forskere kan drage nytte af dem. MANTRA er bygget op, så man kan gennemgå alle modulerne i rækkefølge eller kun tage dem, som vurderes relevante for den pågældende bruger, hvad enten man er seniorforsker, ny forsker, Ph.d. studerende eller såkaldt *information professional* - biblioteksansat eller lignende. Dermed sætter MANTRA denne sidste gruppe i stand til at hjælpe de første med at udfærdige DMP'er. Værktøjerne kan således generelt supplere hinanden både den ene og den anden vej.

Et andet værktøj, RDMRose gennemgår også hele livscyklussen, men henvender sig ikke til forskere, men til førnævnte *information professionals*, især biblioteksansatte. I forhold til DMP'er, linkes der til eksterne eksempler, og der vises eksempler på checklister, som man ofte anvender ved udarbejdelsen af DMP'er. Derudover forklares og begrundes vigtigheden af arbejdet med DMP'er. I det tilfælde at man er en *information professional*, og man er i færd med at skulle bidrage til en udvikling på området, kan der findes megen nyttig information i materialet. Det har form af præsentationer i Powerpoint og aktivitetsark, som kursisterne kan bruge til at besvare de spørgsmål, der stilles i præsentationerne. Der er dermed lagt op til en høj grad af inddragelse af kursisterne og henvisning til deres egen konkrete situation.

C. Værktøjer rettet mod den udførende fase

I den udførende fase indsamles data, og der skal derfor være en infrastruktur, der muliggør ordentlig datahåndtering. Det er også i denne fase, at data analyseres, deles og gemmes og der skal derfor være taget stilling til en masse forhold omkring dette i planlægningsfasen. Er der sket ændringer i udførelse af projektet, eller er der sket ændringer i fx tidsmæssige eller tekniske forudsætninger, er det vigtigt at tage højde for det så tidligt som muligt og i den forbindelse konsultere sin DMP. Der kan være grund til at redigere i sin plan, og derfor kan værktøjer til håndtering af DMP'er være relevante. DMPonline og DMPTool tilbyder begge at gemme og redigere i de udarbejdede DMP'er. Den største forskel er at DMPonline tilbyder faste skabeloner, hvor DMPTool stiller en række spørgsmål og genererer på den baggrund en plan som efterfølgende kan modificeres. Det er ikke umiddelbart til at sige om den ene eller anden løsning er at foretrække, da det hænger nøje sammen af data management organiseringen og politikker.

Både MANTRA og RDMRose gennemgår de elementer, der udgør hovedparten af arbejdet i den udførende fase, men hvor MANTRA har forskere, Ph.d. studerende og biblioteksansatte som målgruppe, er RDMRose henvendt til den sidste gruppe alene. Man kan dog også som *information professional* anvende MANTRA som leverandør af materiale til vejledning af forskere, hvis man påtænker at afholde kurser i det. RDMRose fokuserer alene på dette aspekt, og er ikke tænkt som et vejledningsværktøj for forskerne selv. DataTrain er som RDMRose også målrettet supporterende personale ved at være rammesat som et modulbaseret kursus. Det er således et forslag til en opskrift, der skal kunne inspirere formidlere af data management indenfor specifikke domæner (social antropologi og arkæologi).

Fælles for MANTRA og RDMRose er (modsat DataTrain), at de ikke går meget dybt ned i datahåndteringen, eftersom de vilkår og behov, der hersker på de forskellige fagoråder er så forskellige, at det er uoverkommeligt at gennemgå i et fælles værktøj. Skal man dykke dybere ned, er der behov for de specialiserede IT- og fagfolk, der kender de konkrete betingelser, der gør sig gældende på det konkrete fag.

D. Værktøjer rettet mod publicering og langtidsbevaring

Når det drejer sig om publicering og bevaring, og altså i den afsluttende fase af et forskningsprojekt, er der meget at hente for forskere i MANTRA, både i forhold til begrundelse for deling, praktisk vejledning og i forhold til rettigheder og licenser. RDMRose har også med udgangspunkt i en livscyklus for *data curation* taget publicerings- og bevaringselementerne med. Publiceringsdelen beskrives i større detaljerigdom end bevaringsdelen. Materialerne behandler både data management i praksis, *data repository policies*, faglige repositories, metadata og data citation.

E. Værktøjer rettet mod vejledning, træning og support

Til formålet vejledning og support kan de evaluerede værktøjer opdeles i formidling via ren tekst, ved Power Point præsentationer samt Word dokumenter, som multimedia e-læringsmateriale der benytter sig af mange forskelligartede materialetyper, eller ved et online/offline kursus.

Både Stanford University Libraries og University of Edinburgh har valgt at lægge ren tekst vejledningsmateriale på deres hjemmeside, suppleret med links. Materialet er rettet bredt mod forskere, dækker hele data management cyklus og er ikke disciplinspecifikt. En stor del af materialet er nemt at overføre til danske forhold, mens andre omhandler specifikke amerikanske eller engelske politikker vedtaget af forskningsfonde og universitetet. Stanford University har placeret supportfunktionen på biblioteket, omend indgang til supporten er via "Research & Scholarship" på universitetets webside, mens University of Edinburgh har lagt "Research Support services" ind under "Information Services", som biblioteket hører under.

MANTRA kan bruges som et værktøj man henviser forskere til, som et selv-lærings værktøj, men man kan også bruge de enkelte læringsmaterialer, hvis man fx som biblioteksansat afholder kurser for fx Ph.d.-studerende. RDMRose's materialer kan også nemt anvendes i fx kursus-sammenhæng, og er rent teknisk nemmere at gå til, da alle materialerne ligger som PDF- eller Wordfiler. Det samme gælder Datatraining, hvor enkelte delelementer kan udtages uden større indsats som PPT eller PDF.

I Holland har man udviklet et introduktionskursus rettet mod biblioteksansatte. Formålet er at kursisterne efter kurset skal kunne rådgive forskerne i håndtering af forskningsdata. I kurset gennemgår man de vigtigste aspekter i forhold til data management, tekniske færdigheder, og rådgivning. Materialet er rettet mod en hollandsk kontekst, men er i stort omfang inspireret af

amerikanske og engelske forhold. Da kurset pt. (april 2014) har ændret navn og karakter⁴⁰ efter vores analyseperiode, samt at materialerne ikke længere er online tilgængelige, er det svært at vurdere om kurset vil kunne bruges i en dansk kontekst.

F. Opsummering

Især i planlægningsfasen af et forskningsprojekt vurderes både DMPTool og DMPonline brugbare. Et godt udgangspunkt for data management support synes netop at være selve data management-planen. Arbejdet med en data management plan kan potentielt konkretisere forhold omkring forskningsdatas livsforløb som ikke ville have været artikulert ellers. Derfor kan den udgøre et centralt redskab for fx bibliotekspersonale som sidder i en front office rolle eller som et praktisk element i et data management workshopforløb.

DMP-værktøjerne er designet til henholdsvis amerikanske og engelske forhold, hvor forskningsfondene stiller (forskellige) krav til, at ansøgerne vedlægger en DMP som en del af forskningsansøgning. Dette er endnu ikke tilfældet i Danmark, hvorfor et DMP-værktøj især er brugbart i forhold til forskerens egen planlægning af projektet. Dog præsenterer Horizon2020-piloten ansøgerne for en række krav der allerede på nuværende tidspunkt ville kunne udgøre rammen om de data management plan-skabeloner der blev lagt i en dansk version af DMPonline⁴¹.

Som læringsforløb er værktøjerne MANTRA og RDMRose begge oplagte at bringe i spil i konkrete supportforløb. Hvad enten målgruppen er forskere eller supporterende medarbejdere er de velegnede, brugervenlige og mulige at bruge i en dansk sammenhæng. Dog bør det overvejes at dele af indholdet oversættes til danske forhold da der hyppigt henvises til UK specifikke politikker, websider, forskningsfonde og cases (de konkrete software øvelser er formentlig en undtagelse). Selve teksten kan godt overvejes beholdt på engelsk, da mange forskere og Ph.d. studerende på danske universiteter ikke taler dansk. Da MANTRA er et så stort kompleks skal der også tages højde for at en dansk version vil kræve ressourcer til vedligeholdelse af fx indholdet. RDMRose har den fordel at materialerne ikke er systemafhængige, men blot distribueret på en en website. Dette forudsætter dog også at de skal bringes i spil inden for en anden ramme - eksempelvis et kursus eller en e-læringsløsning. Det gør det dog samtidig også væsentlig nemmere at håndplukke delaspekter til egen tilpasning. En anden fordel ved RDMRose er at det indeholder disciplinspecifikke cases inden for både tekniske, medicinske og samfundsvidenskabelige områder, der vil kunne bruges til at gøre support målrettet og aktuel. Datatraining giver gode eksempler på netop det disciplinspecifikke, men har ikke den samme brugervenlighed som MANTRA og RDMRose og fungerer bedst som inspiration, frem for som et direkte applicerbart værktøj. DataTrain er af den grund oplagt at kigge nærmere på hvis man har et mere snævert supportsigte.

På trods af at materialernes pædagogiske og formidlingsmæssige kvaliteter generelt er høje, har de muligvis en udfordring ved at stå alene som selv-lærings-services. Komplexiteten er nogle

⁴⁰ Det evaluerede kursus "Data intelligence 4 librarians" har skiftet navn til "Essentials 4 Data support" og fra at være et 4 dages kursus blev det nu til 2 dage.

⁴¹ "Kapitlet "Datamanagement planning" indeholder en teknisk evaluering af DMPonline softwaren. Evalueringen kommer ind på usability, funktionalitet, krav til installation og indsats ved en dansk DMPonline løsning."

steder lav og andre steder stor, hvorfor disse værktøjer virker relevante at bringe i spil sammen med mere lettilgængelige - så som webformidling eller som del af et undervisningsforløb.

Både Stanford University og University of Edinburgh har eksemplarisk og simpel formidling af data management, målrettet forskere. Introduktionerne har den åbenlyse fordel at de er nemme at komme i gang med og udgør således et oplagt først skridt i at formidle data management services. Supportpotentialer ligger først og fremmest i at det understøtter de forskere der selv er opsøgende og er åbne for selvstudier. Begge har overskuelig opbygning og dækker hele data management livscyklus, som gør at man hurtigt får et overblik over de forskellige delaspekter der er forbundet med data management. En stor del af materialet er nemt at overføre til danske forhold, mens andre omhandler specifikke amerikanske eller engelske politikker vedtaget af forskningsfonde og universitetet. Det vil sige at de kommunikative og pædagogiske rammer kan anvendes uden større tilretning, hvor det indholdsmæssige bør approprieres. Webinformation kunne tilpasses lokale institutioners egen interne organisering af data management services, såfremt en sådan supportløsning overvejes.

Bilag 1: Studieture til Holland og UK, november-december 2013

Visit Danish Research Library delegation to SURF, 21 November 2013

SURF Office, Meeting room VK1
Hojel City Center, Building D.
Graadt van Roggenweg 340,
3531 AH Utrecht

Theme:

Information and dialogue on national Research Data Services, Management and Infrastructure

Visiting party

Sacha Zurcher	Roskilde University Library
Mikkel Hvidtfeldt Andersen	Roskilde University Library
Søren Æhrendahl Mikkelsen	State and University Library, Aarhus
Diba Markus	Danish e-Infrastructure Cooperation
Christian Ertmann-Christiansen	The Royal Library, Copenhagen
Anders Conrad	The Royal Library, Copenhagen

Invited speakers (morning):

Paul van Dijk	SURFnet
Madeleine Smaele	3TU
Mark van Sanden	SURFsara

Invited speakers (afternoon):

Elly Dijk	NARCIS
Ingrid Dillo	DANS
Patrick Aerts	NLeSC

Hosting party (all day):

Marc Dupuis	SURF
Bas Cordewener	SURF
Magchiel Bijsterboch	SURF

Programme

09.30 - 09.40	Welcome and opening	Bas Cordewener
09.40 - 09.50	Introduction and background study visit	Chair Danish delegation
09.50 - 10.15	SURF, Platform IT & Research	Marc Dupuis
10.15 - 10.55	SURFnet	Paul van Dijk
10.55 - 11.35	3TU	Madeleine de Smaele
11.35 - 12.15	SURFsara	Mark van Sanden

12.15 - 13.00 Lunch

13.00 – 13.50 DANS

Ingrid Dillo

13.50 – 14.40 NARCIS

Elly Dijk

14.40 – 15.30 NLeSc

Patrick Aerts

15.30 – 16.00 Coffee & Tea & Close

	Studietur Edinburgh og Oxford		
Rejsende			
Navn	4th December 2013	5th to 6th December 2013	
Anders Sparre Conrad	Edinburgh	Oxford	
Søren Ærendahl Mikkelsen		Oxford	
Jesper Boserup Thestrup		Oxford	
Sacha Zurcher		Oxford	
Asger Væring Larsen	Edinburgh	Oxford	
Diba Markus	Edinburgh		
Christian Ertmann	Edinburgh		
	Møde med		
Edinburgh	Kevin Ashley	DCC - Digital Curation Centre	As Director, Kevin maps new territory as the DCC embarks on its third phase of evolution (2010 - 2013), where the accent is on enabling capacity and capability amongst the research community in matters of digital curation. Previously, as Head of Digital Archives at the University of London Computer Centre (1997 - 2010), he was responsible for a multi-disciplinary group that provided services related to the preservation and reusability of digital resources on behalf of other organisations, as well as programmes of research, development and training. His group operated NDAD (the National Digital Archive of Datasets) for The National Archives of the UK for over twelve years, capturing, preserving, describing and releasing government data as part of TNA's collections. As a past or present member of numerous advisory and steering groups, including JISC's Infrastructure and Resources Committee, JIIE, the Repositories and Preservation Advisory Group, the Advisory Council for ERPANET and the Archives Hub Steering Committee, Kevin has contributed widely to the research information community. - See more at: http://www.dcc.ac.uk/about-us/dcc-staff-directory/kevin-

			ashley#sthash.N8AQSRF7.dpuf
	Martin Donnelly	DCC	Martin's academic background is in English Literature and Information Technology, and before joining the University of Edinburgh he was Technology Assessor and Tutor/Lecturer at the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow. Martin's curation focus is on data management planning and policy. He wrote a book chapter on data management planning in 2012 (Pryor ed., Managing Research Data, London: Facet), and was co-author of the DCC's original "Checklist for a Data Management Plan". He also conceived and project managed the first three iterations of the DCC's DMPonline tool, and has presented about this in the UK, the US, Canada, Australia, New Zealand, Denmark, Italy and Belgium. Externally, Martin was a founder member of the steering group of the US DMPTool, sits on the Advisory Board for Databib, and acts as Secretary to the UK CODATA national committee. An erstwhile HE administrator, Martin previously served as Policy Officer at the University of Glasgow and as Quality Officer (Strategic Development) at Edinburgh College of Art, and as such is interested in the interlinking of research and administrative information systems. Also interested in arts and humanities data management, government data/policy, and literary and cultural theory. - See more at: http://www.dcc.ac.uk/about-us/dcc-staff-directory/martin-donnelly#sthash.BUhMlvCx.dpuf
	Robin Rice	Data Library - EDINA	Robin Rice is Data Librarian for the University of Edinburgh, based at EDINA and Data Library, a division of Information Services (IS). She is the service manager of both the University's Data Library and the institutional data repository, Edinburgh DataShare, and serves on the EDINA management team and the IS Research Data Management Implementation Committee. She has led a number of JISC-funded projects aimed at building capacity

			for data curation and research data management services at the University of Edinburgh. She was the Project Coordinator of the Digital Curation Centre during its set-up phase in 2004-05. Robin has been a data librarian for twenty years, in the UK and at the University of Wisconsin. She is active in IASSIST, an international membership organisation for academic data professionals, and currently chairs the Communications Committee.
	John Scally	The University of Edinburgh	Dr. John Scally is Director of Library and University Collections, including the university-wide Information Services. John Scally has been appointed chair of the university's RDM Policy Implementation Committee, that is charged with delivering services that will meet the University's Research Data Management Policy, as decided May 2011. As of June 2014, John Scally has been appointed Chief Executive and National Librarian at the Edinburgh based National Library of Scotland.
	Stuart Lewis	The University of Edinburgh	Stuart Lewis is head of the Digital Library and part of the senior management team of the university library. As such, he is responsible for the program to deliver storage space for research data to all active researchers at the Edinburgh University.
Oxford	Wolfram Horstmann	Bodleian Libraries	Dr Wolfram Horstmann joined the Bodleian Libraries as Associate Director for Digital Library Programmes and Information Technologies in 2012. In consultation with all relevant parts of the Libraries, he is developing a Digital Strategy for the Bodleian Libraries, put into practice by his team, the Bodleian Digital Library Systems and Services. As Chief Information Officer for Scholarly Information at Bielefeld University – which has one of the highest ranked libraries in Germany – he was responsible for strategic development of the institutional services for content, data and tools in research and education between 2007 and 2011. He has also served as European project manager for the State and University Library of Goettingen and as head of Publishing Services at the Academic Library Centre in Cologne. He is a biologist by training, having worked in the field of computational neuroscience before he actively turned his attention towards scholarly communication and libraries.

Annual Workshop of the International Association of Scientific and Technological Libraries ([IATUL](#))

Research Data Management: Finding our Role

University of Oxford
5-6 December 2013

Workshop Programme

Day One: Thursday 5 December 2013

Divinity School, Bodleian Library

17:30-19:00 Welcome Reception - Divinity School, Bodleian Library (network with colleagues from Oxford)

Day Two: Friday 6 December 2013

Arco Building, Keble College

08:30-09:00 Coffee and Registration

09:00-09:20 Welcome and Introduction

09:20-09:40 [An Eight Year Odyssey in Data Management: Purdue University](#) - Dr James L. Mullins, Dean of Libraries, Purdue University, USA (video at <http://www.youtube.com/watch?v=jYBcsfiRhio>)

09:40-10:00 *Dipping our Toes in Data... Exploring our Role* - Elisabeth Newbold, Content & Collections Leader, The British Library, UK

10:00-10:20 [PIDs and DOI Registration with DataCite](#) - Frauke Ziedorn, DOI service/DataCite, German National Library of Science & Technology, Hannover, Germany

10:20-10:40 [Managing Research Data: the Organisational Challenge at Oxford](#) - James Wilson, Senior Project Manager, University of Oxford, UK

10:40-11:00 Break

11:00-11:20 [Harnessing Bodleian Expertise for Oxford's Research Data Management](#) - Sally Rumsey, Digital Research Librarian & Senior Project Manager, University of Oxford, UK

11:20-11:40 [IT Utilities to Exploit Data as a Commodity in Digital Economy](#) - Steve Brewer, Network Coordinator, IT as a Utility Network, University of Southampton, UK

11:40-12:00 [Zenodo](#) - Lars Holm Nielsen, Collaboration & Information Services,
IT Department, CERN, Geneva, Switzerland

12:00-12:20 [How to Deal with Petabytes of Data: the LHC Computing Grid](#) - Peter Gronbech,
GridPP Project Manager, University of Oxford, UK

12:20-13:30 Lunch in the Divinity School, Bodleian Library

13:30-13:50 [Building Confidence: Training Librarians in Research Data Management](#) - Robin
Rice, Data Librarian, University of Edinburgh, UK

13:50-14:10 [The Principles of Guerrilla Warfare and the Fulfilment of the Library's RDM Role](#) -
Arjen Voogt, Consultant, University of Groningen, Netherlands

14:10-14:30 [Strengthening the University of Sheffield's Research Environment by Building
RDM Capability and Capacity](#) - Laurian Williamson, Research Data Management
Coordinator, University of Sheffield, UK

14:30-14:50 [From JISC Project to Embedded Data Service](#) - Dr Catherine Pink, Research
Data Scientist, University of Bath, UK

14:50-15:15 Questions and Answers

15:15-16:00 Closing Discussion and Drinks

**16:30-17:00 Christmas music at the Radcliffe Science Library from the Bodleian Libraries
Choir, followed by a tour of the Library.**

Bilag 2: Vurdering af informations-, trænings- og planlægningsværktøjer for data management

Hent på <http://goo.gl/gtr7po>

Bilag 2 : Vurdering af informations-, trænings- og planlægningsværktøjer for data management

Kriterier/Navn	MANTRA		Data Intelligence 4 librarians		DataTrain		Stanford University Libraries		University of Edinburgh		DMPOnline		DMPTool		RDMRose		Beskrivelse af kriterier
	Vurdering	Kommentar	Vurdering	Kommentar	Vurdering	Kommentar	Vurdering	Kommentar	Vurdering	Kommentar	Vurdering	Kommentar	Vurdering	Kommentar	Vurdering	Kommentar	
Type		Elæringsmateriale, interaktiv multimedie-kursusplatform suppleret af guidelines via webside		Kursus: A didactical mix of group meetings, online study and assignments targeted data curation. The contents of the website are intended as reference material.		Struktureret e-undervisningsmateriale. 1. Arkæologi: Eight presentations, followed by group discussion and written exercises. 2. Socialantropologi: 4 temablokke designet som interaktiv workshop		Vejledningsmateriale på web med links		Vejledningsmateriale på web med links		8-trin webservice-guide til udarbejdelse af generisk Data Management Plan. Inkluderer også en fortløbende checkliste.		11 trins web-guide til udfærdigelse af generisk DMP. Inkluderer generel guide		Materialet er designet til selv-undervisning via web. Materialet kan downloades og anvendes i sine enkelte dele hver for sig. Der findes en "light-version".	Er det online eller face-to-face, selvlæringsmateriale eller kursus?
Målgruppe		Alle målgrupper rammes. Henvender sig fortrinsvist til PhD og yngre forskere		Biblioteks-ansatte		1. The teaching material is targeted at co-ordinators of Core Research Skills courses for first year post-graduate research students in archaeology. 2. Socialantropologi: målrettet pre-fieldwork PhD'ere		Primært forskere og early career researchers, men Stanford er også åben for at inkludere research data management allerede i studerendes curriculum hvis en lærer ønsker det.		Forskere		Forskere		Forskere		Målgruppen er informationel professionals, herunder bibliotekarer og informationsspecialister. Materialet er på postgrad niveau.	Er målgruppen forskere, PhD-studerende, eller biblioteksansatte?
Materialetyper		Tekst, billeder, illustrationer, videointerviews, quizzer, guidelines, DMP-templates, software practicals, checklists, referencer til selvstudiemateriale. Materialet er integreret i et modulært system under forskellige tematikker.		Tekst, guidelines, checklists, referencer til selvstudiemateriale, video (på hollandsk), personlig support og træning. Kurset er opbygget via forskellige moduler, hver med sin egen tematik.		Primært bestående af tekst, testimonials, guidelines, checklists og referencer/links til selvstudier. Der er en survey som de studerende skal udfylde inden kurset således at det kan målrettes.		Ren tekst		Ren tekst		Tekst		Tekst samt en enkelt instruktionsvideo		Materialet består af Power Point præsentationer samt word dokumenter man som bruger kan besvare spørgsmål i.	Hvordan er materialet sat sammen?
Varighed og omfang		Materialet er meget omfattende, men inddelt i mindre moduler der kan tilgås særskilt. Det estimeres at materialet kan gennemgås på egen hånd på 10 timer.		4 dages kursus plus egen tid til opgaver og læsning. I alt omkring 7 dage.		1. Arkæologi: 4 x 2 timer. 2. Socialantropologi: 3 x 1.5 timer (minimum)		Web-introduktion med links til yderlige information.		Materialet er mere omfattende end Stanfords, men er stadig "kun" en web-introduktion		Materialet er ikke omfattende. Afhænger af forskningsprojektets karakter og grundigheden af den information brugeren selv fylder ind.		Materialet er ikke omfattende. Afhænger af forskningsprojektets karakter og grundigheden af den information brugeren selv fylder ind.		Materialet er inddelt i 8 sektioner, der hver ansås at tage en halv dag (information fra RDMRose) så ialt vil det tage 4 dage at gennemgå	Hvor lang tid tager det at læse materialet/tage kurset og hvor omfattende er materialet?
Licens/Jura		CC-BY		CC-BY		CC-BY-NC-SA		Materialet er omfattet af almindelig ophavsret		Materialet er omfattet af almindelig ophavsret		n/a		Materialet er omfattet af almindelig ophavsret		Materialet er frigivet under en CC-BY-SA licens	De juridiske forhold og evt. licenstyper beskrives - hvilke regler gælder der for (gen)anvendelse af materialet?
Producent		EDINA and Data Library, University of Edinburgh		3TU Datacentrum		The University of York and Archaeological Data Service		Stanford University Libraries		University of Edinburgh		DCC (Digital Curation Centre) og hosted af University of Edinburgh		University of California Curation Center of the California Digital Library		The JISC funded RDMRose project (Jul 2012-Jun 2013) was a collaboration between the libraries of the University of Leeds, Sheffield and York, with the Information School at Sheffield to provide an Open Educational	Hvem har produceret materialet?
Samarbejdspartnere		Graduate Schools: University of Edinburgh's School of Social and Political Science, School of GeoSciences and the Doctorate in Clinical Psychology, samt andre forskere og PhD'ere fra University of Edinburgh.		DANS - Data Archiving and Networking Services		Lavet i samarbejde med undervisere på Department of Social Anthropology og Department of Archaeology, University of Cambridge. Projektet blev pilottestet som del af en skill programme i ovenstående Departments.		nej		nej		Persondata og projektdata lagres hos University of Edinburgh		The University of California Curation Center (UC3) at the California Digital Library, DataONE. Digital Curation Centre (UK), Smithsonian Institution, University of California, Los Angeles Library, University of California, San Diego Libraries, University of Illinois, Urbana-Champaign, University of Virginia Library.		JISC samt US advisory board	Hvis materialet er udviklet i samarbejde med repræsentanter fra målgruppen anføres det her.
Nominel vurdering af 2 personer (fra 1 - 3 hvor 3 er højest)																	
Formål - oplyder det formålet, som det er defineret under kriterier i kolonne 1?	6	Multiple	5	Deltagere skal kunne rådgive forskerne i håndtering af forskningsdata. Kurset er på introduktionsniveau.	5	To equip first year post-graduate students with essential skills in looking after their research data for their PhD	5	Multiple	6	Multiple	4	To help create and maintain different versions of Data Management Plans. To provide useful guidance on data management issues and how	3	Create ready-to-use data management plans for specific funding agencies. Meet requirements for data management plans. Get step-by-step	6	RDMRose developed and adapted learning materials about RDM to meet the specific needs of liaison librarians in university libraries, both for practitioners' CPO and for	Målgruppen skal forstå vigtigheden af data management generelt, og forstå hvad dette har af implikationer for forskerne. 3. gives når målgruppen skal være i stand til at dokumentere forskning via data; 2. gives når målgruppen skal være i stand til at håndtere og organisere forskningsdata; 1. gives når målgruppen skal forstå formlen med en
Disciplinspecifichet - henvender materialet sig til et fag eller er det generelt?	6	Ikke specifikt, målsætning er at være bred / at kunne udbyrdes på tværs af discipliner	5	Ikke specifikt på bestemte fagområder	3	Materialet er målrettet arkæologi og socialantropologi, men har generiske aspekter	6	Ikke specifikt, målsætning er at være bred / at kunne udbyrdes på tværs af discipliner	6	Ikke specifikt, målsætning er at være bred / at kunne udbyrdes på tværs af discipliner	6	Ikke specifikt.	6	Ikke specifikt	6	Materialet inddrager 6 meget forskelligartede fagområder / cases	Variereet fokus foretrækkes, men der skal være taget hensyn til områdespecifikke (nat/tek/sund vs. hum/samf) elementer. 3. gives til materialer med en kombination af generelt fokus med hensyntagen til områdespecifikke elementer; 2. til materialer med et generelt fokus;
Indhold - er materialet overførbart til danske forhold?	4	Dele af materialet er nemt at overføre til danske forhold, mens andre omhandler engelske forhold, fx politikker vedtaget af forskningsfonde og universitetet	4	Afhængig hvordan Danmark kommer til at udforme research data management, kan materialet delvis overføres til danske forhold. Det hollandske materiale er i stor omfang inspireret af England.	4	Dele af materialet er nemt at overføre til danske forhold, mens andre omhandler amerikanske forhold, fx politikker vedtaget af forskningsfonde og universitetet	4	Dele af materialet er nemt at overføre til danske forhold, mens andre omhandler engelske forhold, fx politikker vedtaget af forskningsfonde og universitetet	4	Dele af materialet er nemt at overføre til danske forhold, mens andre omhandler engelske forhold, fx politikker vedtaget af forskningsfonde og universitetet	4	Materialet er skræddersyet til forhold med udgangspunkt i analyse af de enkelte engelske funding bodies' data-krav og specifikationer. Der findes dog en generisk skabelon. Softwaren findes i Github, og DCC vil gerne være behjælpelige med overførsel	2	Materialet er skræddersyet til amerikanske forhold med udgangspunkt i analyse af en række bevillingsgivers datakrav og politikker.	4	Materialet er målrettet britiske forhold, herunder hensyntagen til britiske forskningsfonde og universitetskrav. Meget af materialet er på et så generelt niveau at det sagtens kan oversættes til danske forhold - fx temaer om forskningsmiljøer, data og livscyklus	3. gives til materialer der nemt kan overføres til danske forhold; 2. gives til materialer der med nogen bearbejdning kan overføres til danske forhold; 1. gives til materialer der ikke kan overføres til danske forhold.
Pædagogisk tilgang - hvad er læringsmetoden (foredrag, hands-on, problemløserorienteret)?	6	Introduktioner i tekst, konkrete cases i video, opgaver udgår tilsammen en frugtbar konkret problembaseret tilgang til læring.	6	Kursus (både online og face-to-face), gruppearbejde, assignments, samt videotutorial	4	Undervisningsmateriale i form af slides plus pdf'er med noter til hver slide. Materialet er har status af informationsmateriale og vejledende rådgivning. Surveyen giver dog god mulighed for at målrette kurset konkrete behov.	2	Materialet vurderes som værende ret en-dimensionelt	2	Materialet vurderes som værende ret en-dimensionelt	4	Materialet er instruktivt og indeholder informativt vejledningsmateriale. Materialet er ikke ment som undervisningsmateriale, men som skabeloner til udfyldelse af DMP'er. Men man bliver kun præsenteret for de felter, der er nødvendige i forhold til den fond	4	Materialet er instruktivt og indeholder informativt vejledningsmateriale. Materialet er ikke ment som undervisningsmateriale, men som skabeloner til udfyldelse af DMP'er. (Ager: Der er både tekstuel og visuel guide til anvendelse af DMPTool)	4	Materialet er alene baseret på Power points og tekstdokumenter. Der er dog enkelte video'er. Der er mulighed for at give feed-back, men om det er muligt, at få hjælp fremgår ikke af materialet. Læringsstilen i materialet er problembaseret med en vekslen mellem information og selvstændig aktivitet.	3. gives til materiale der har varierede tilgange. Fx foredrag, "hands-on", selvlæring, problemløserorienteret; 2. gives til materialer, der er begrænset i variation; 1. gives til ren tekst eller ren lecture. For materialer der har fået 1 eller 2, kan karakteren hæves med 1, hvis der er mulighed for efterfølgende personlig support.

Bilag 2 : Vurdering af informations-, trænings- og planlægningsværktøjer for data management

Hent på <http://goo.gl/gtr7po>

Brugervenlighed - hvor brugervenligt er det undervisningsmateriale der er online tilgængeligt? Hvordan er det bygget op og er det nemt at gå til?	5	Nem navigering og opdeling efter emne. Design og useende virker umoderne. Nogenlunde intuitiv begrebslogik.	6	Nem navigering og opdeling efter emne	5	Opdeling efter emne. Design fordrer ikke en jævn læseoplevelse. Materialet et systematisk præsenteret. Navigation er nogenlunde intuitiv.	6	Materialet er nemt at anvende	6	Materialet er nemt at anvende	6	Høj brugervenlighed	4	Materialet er nemt at anvende. Dog er felterne ikke opdelt, men samlet i større enheder, hvilket kan gøre det mere uoverskueligt. Gode hjælpetekster og links til andre sider med hjælp.	6	Materialet er nemt at gå til. Dog er det meget omfattende, så alene omfanget kan gøre det svært at overskue, men det er inddelt i sektioner med overskrifter, der tydeligt beskriver indholdet. Light-	3. gives til emneopdeling, sprogligt veformuleret materiale og intuitiv navigation i en høj omfang; 2. gives til emneopdeling, sprogligt veformuleret materiale og intuitiv navigation i begrænset omfang; 1. gives til emneopdeling, sprogligt veformuleret materiale og intuitiv navigation i yderst begrænset omfang.
Sammenhæng med andre services - hvor tæt knyttet er materialet til andre services eller systemer?	6	Ikke låst i forhold til særlig service eller system. Materialet kan fint stå alene som selvstudie, men kan også anvendes som del af et kursusforløb. Der refereres i software practicals til proprietært software.	4	Materialet er tænkt ind i sammenhæng med face-to-face kurset	3	Materialet er målrettet arkæologi-repositoriet ADS og Cambridge's Dspace installation. Der redegøres for hvor undervisere fra andre institutioner bør ændre deres brug af materialet.	4	Nogle punkter er ikke beskrevet men består af links til interne eller eksterne systemer. Der er tegnet licens til et antal services	6	Man kan godt anvende materialet uden tilknytning til andre systemer eller services, men der links til dem som forskerne kan benytte sig af på Edinburgh.	6	DMPOnline referer til DCC's online ressourcecenter hvor forhold omkring DMP udbydes og hvor videre vejledning er tilgængeligt (fx DCC CARDIO)	6	Der er særskilte links til og templates fra bevillingsgivere. Dette er tilsyneladende også inkluderet i de enkelte DMP-skabeloner. Den enkelte DMP er afhængig af hvilken fond, der giver penge. Der er ikke en generisk plan.	5	Materialet læner sig meget op af DCC's webiders indhold, så en oversættelse af siderne vil skulle tage hensyn til dette, idet der ikke findes tilsvarende sider på dansk	3. gives til materialet der er uafhængigt af andre systemer eller services; 2. gives til materialet der er delvist afhængigt af andre systemer eller services; 1. gives til materialet der er ubrugeligt uden tilknyttede systemer eller services.
Data Livscyklus - hvilke dele af datas livscyklus er materialet centreret omkring?	6	Hele cyklus dækkes.	6	Hele cyklus dækkes (baseret på UK Data Archive's Life Cycle fra 2011)	6	Hele cyklus dækkes	4	Materialet er struktureret trinvis og dækker hele cyklussen	5	Materialet er struktureret trinvis og dækker hele cyklussen	6	Materialet dækker planlægningsfasen. Typisk vil DMP knytte sig til planlægningsfasen og kun opdateres undervejs, hvis der	6	Materialet dækker planlægningsfasen. Typisk vil DMP knytte sig til planlægningsfasen og kun opdateres undervejs, hvis der er	6	Materialet behandler dette tema indgående og beskriver selv livscyklussen	3. gives til materialer der dækker hele forskningscyklussen; 2. gives til materialer der delvist dækker forskningscyklussen; 1. gives til materialer, der kun dækker 1 eller 2 trin.
DMP templates - er der adgang til eksempler på DMP'er?	6	Templates er indeholdt. UKDA (UK Data Archive) research data lifecycle diagram og The JISC (Joint Information Systems Committee) research lifecycle diagram præsenteres i materialet som inddækket i kurset.	6	Templates er indeholdt (forskellige eksempler fra DMP's er med i kurset og supplerede materiale online).	6	Templates er indeholdt.	6	Templates er indeholdt	6	DMP-templates er ikke inkluderet, men der er eksempler, og der links til DMPOnline og MANTRA.	6	Der links i stiet til eksempler af DMP'er fra Rural Economy and Land Use Programme (RELU). Selve site indeholder ikke DMP'er.	6	Stiet indeholder eksempler på DMP'er. Der er skabeloner til alle fondenes DMP'er og det er muligt at dele sin plan med andre (man får et link til planen i pdf, som man kan sende til andre. Enkelte eksempler til DMP'er findes.	6	Materialet indeholder ikke selv eksempler, men der links til en række forskellige eksempler på DMP'er. desuden findes eksempel på checkliste.	3. gives til materialer der indeholder forskelligartede templates på DMP'er; 2. gives til materialer indeholdende en generisk template; 1. gives til materialer uden templates.
Kan materialet kun bruges af målgruppen eller indeholder det også information om anvendelse af materialet i en undervisningssituation?	6	Materialet er omfattende, og er designet til selvstudie, men kan fint anvendes i face-to-face undervisning. Der er desuden selvstudie materiale til Data curation profiles og Data interviews.	5	Sigtet er træning af biblioteksansatte	6	Materialet er skabt til underviser- selv om det er universitets- undervisere. Der medfølger forslag til 1-dagskursusplan	2	Materialet er udelukkende bestemt for forskere	2	Materialet er udelukkende bestemt for forskere	4	Materialet er for forskere og forskningsgrupper og er ikke decideret undervisningsmateriale. Det kan givetvis anvendes som øvelsesredskab i undervisningssituationer	4	Materialet er for forskere og forskningsgrupper og er ikke decideret undervisningsmateriale. Det kan givetvis anvendes som øvelsesredskab i undervisningssituationer	5	Materialet er lavet, så det mere eller mindre direkte kan anvendes i en undervisningssituation. Da materialet er målrettet information professionals, opfylder det kriteriet om, at der skal være train-the-trainer elementer indeholdt.	3. gives til materialer der henvender sig både til forskere og trainere; 2. gives til materialer der kun henvender sig til forskere eller trainere; Der gives ikke karakteren 1.
Tilgængelighed - hvor let tilgængeligt er materialet?	6	Materialet er frit tilgængeligt på nettet.	3	Der er kun plads til 16 personer på kurset, der afholdes i Holland. Afholdelsesfrekvens usikker. Materiale er frit tilgængeligt på nettet, men er supplerede til face-	6	Materialet er frit tilgængeligt på nettet	6	Materialet er frit tilgængeligt på nettet	6	Materialet er frit tilgængeligt på nettet	6	Frit tilgængeligt. Profil skal oprettes.	6	Frit tilgængeligt. Profil skal oprettes.	6	Materialet ligger frit tilgængeligt på Internet	3. gives til materialer der er frit tilgængeligt på nettet; 2. gives hvis det er frit tilgængeligt efter endt kursus; 1. gives hvis man kan få adgang ved betaling.
Sprog - hvilke(t) sprog findes materialet på?	6	engelsk	5	Både på hollandsk og engelsk. Kurset gives på hollandsk. Online	6	engelsk	6	engelsk	6	engelsk	6	engelsk	6	engelsk	6	engelsk	Dansk og engelsk foretrækkes. 3. gives til engelsk; 2. gives til dansk; 1. gives til andre sprog.
Teknik - hvor nemt er det at eksportere materialet til andre formater?	3	Ingen eksport-facilitet. Slides ligger indbygget i fremviser-softwaren. Dog ligger multimedie materiale og opgaver tilgængeligt særskilt.	2	Ingen eksport-facilitet. Kursusmateriale udleveres på kurset.	6	Materialerne ligger i ppt, pdf og rtf-formaterne. Materialet er generelt blot præsenteret på websider.	4	Materialet findes i html	4	Materialet findes i html	6	Webservice der kan eksporteres som html, tekst eller pdf. Som nævnt ligger softwaren i GitHub, så hvis man vil benytte denne infrastruktur til at præsentere skabelonerne, skal denne software installeres. DCC vil gerne være	6	Webservice der kan eksporteres som rtf, tekst eller pdf. Nem download.	4	Materialet indeholder ikke en egentlig eksportfunktion, hvorfor det ikke kan få 3. Det ligger dog i ppt og doc formaterne og kan sektionvist downloades	3. gives til materialer hvor eksport-faciliteter gør det nemt at trække materialer ud eller downloade dem; 2. gives hvis der i materialet er indbygget noget software, som besværliggør udtræk; 1. gives hvis det kræver fx installation af bestemt software for at tilgå materialet.
Samlet karakter	66		57		60		55		59		64		59		64		
Generel samlet bedømmelse		Elæringsmateriale, interaktiv multimedie-kursusplatform suppleret af guidelines via website. MANTRA er state of the art hvad angår multimedial undervisningsmateriale i Data		Omfattende materiale, mættet primært på nationale forhold, men materiale er meget inspireret af amerikanske og engelske forhold.		Meget anvendeligt materiale, der dog har et snævert faldt sigte. Godt udgangspunkt for disciplinsspecifikt undervisningsmateriale. Brugervenligheden er ikke lige så		Materialet er ikke særligt grundigt omend de bagved liggende services er omfattende. Der er meget tekst, og mange links. Opbygning er dog overskuelig.		Indholdsmæssigt mere omfangsrigt end Standforfs og dermed mere anvendeligt. På den anden side stiller det større krav til formidling som ikke nødvendigvis opfyldes lige godt i alle henseende		DMPOnline gør det nemt og overskueligt at producere en plan for sin data om end den ikke leverer rum for mange individuelle forhold. Forskere skal have teknisk viden om de data de indsamler		DMPTool fungerer meget intuitivt og er gennemtænkt i forhold til anvendelighed. På samme vis som DMPOnline er materialet skræddersyet til nationale forhold hvilket foranleder en		RDMRose er skabt til at undervise ud fra. Materialerne ligger frit tilgængeligt og kan genbruges og/eller ændres som man ønsker det. Materialerne er af høj kvalitet og vil kunne medvirke til at forberede fx	

Bilag 3: DEFF-ansøgningen *Data management i praksis* **Data management i praksis**

Ansøger og økonomisk/juridisk ansvarlig

2. Det Kongelige Bibliotek 3. Postadresse: Postbox 2149 5. DK-1016 6. København K 7. +45 33 47 47 47 8. E-postadresse: kb@kb.dk 9. CVR-nr: 28 98 88 42 10. P-nummer – AEN 5798 000795297 11. Bankregistreringsnummer: Regnr. 0216 12. Kontonummer: 4069032583 IBAN: DK2002164069032583 SWIFT: DABADKKK	13. Birte Christensen-Dalsgaard 14. Vicedirektør 15. bcd@kb.dk 16. Postboks 2149 17. Postadresse (evt. fortsat) 18. 1016 19. København K 20. Medansøgere: SB, DTIC, DEIC (søger ikke penge), SDU, RUB, Statens Arkiver (DDA), AUB, CBS Bibliotek
---	--

Projektet

21. Hvilke(t) indsatsområde(r) hører projektet under?

Data management af forskningsdata

22. Overordnet beskrivelse af projektet.

Projektets mål er at etablere et antal kørende tjenester til understøttelse af alle dele af data management livscyklus for forskningsdata, fra data management planer til digital bevaring.

Tjenesterne vil blive etableret i samarbejde med udvalgte igangværende forskningsprojekter og –infrastrukturer, fordelt over flere forskningsmæssige hovedområder. Konkret vil udviklingen ske i et antal cases, som hver skal levere en løsning i forhold til en bestemt forskergruppe. Hver case vil blive uddelegeret til en eller flere projektpartnere, og hver case vil have en hovedansvarlig part. De enkelte cases vil alle omfatte den fulde livscyklus, men der vil samtidig blive organiseret en koordinering af de forskellige cases i forhold til en række temaer, inspireret ud fra livscyklustrin.

De implementerede løsninger vil tage udgangspunkt i de evalueringsresultater og anbefalinger som kommer ud af det igangværende *Fælles Infrastruktur for Forskningsdata (FIF)*- projekt. De vil være af en kvalitet så de efterfølgende vil kunne benyttes til forvaltning af forskningsdata af de pågældende forskerteams og eventuelt af beslægtede forskningsprojekter. Løsningerne vil samtidig tjene som grundlag for fortsat evaluering af både systemer, teknologier og af værdien af de tjenester som forskningsbibliotekerne vil kunne tilbyde på området. Projektet ønsker dermed ad praktisk vej også at yde et bidrag til den igangværende strategiproces.

Figuren nedenfor illustrerer projektets organisation i cases og tværgående temaer:

HUM cases (SB og KB)	SAMF/SUND cases (DDA, SDU, RUC)	NAT cases (KB)	TEK cases (DTU)	DeIC Sekretariat
Data Management Planning				Koordinering af tværgående båndssamspil med strategi- og udrulningsprocesser i DeIC, DEFF, DU regi
Data capture, storage and documentation (basic metadata)				
Data identification, citation and discovery				
Select and deposit for long term preservation				
Training and marketing toolkits				

I de enkelte cases involveres forskerne og konkrete forskningsprojekter, hvorved deres behov og praktiske udfordringer tages med i den færdige løsning.

HUM cases

- Statsbibliotekets cases koncentrerer sig om eksisterende forskningsprojekter inden for humaniora, bla. LARM og DigHumLab tema-2a (Netlab). Der er i disse projekter allerede udtrykt behov for håndtering og bevaring af forskningsdata. Der vil være et særligt fokus på forholdet mellem kildedata (fx. kulturarv) og nye forskningsdata produceret i projekterne
- Det Kongelige Bibliotek vil arbejde med tekstkorpora som forskningsdata specielt hvordan TEI header kan bruges til at strukturere forskningsrelaterede information, et arbejde som startede i Clarin og fortsætter i DigHumLab tema 1 samt vil være anvendeligt i forhold forskningsbaserede udgivelser som Søren Kierkegaards Skrifter og Holberg udgivelsen.

SAMF/SUND cases

- Syddansk Universitetsbibliotek og DDA's ene case koncentrerer sig om forskningsdata, der kombinerer surveys, fysiske målinger, registerudtræk og evt. andet. Hele datas livscyklus vil være i fokus, og der vil desuden være fokus på kobling og citation mellem forfattere, publikationer og datasæt. Et forskningsprojekt hos Statens Institut for Folkesundhed, SDU vil danne fundament for casen.
- Syddansk Universitetsbibliotek og DDA's anden case vil fokusere på projekter inden for kommunalforskning. Her vil der med casen arbejdes med, hvordan opbygning af forskningsdokumentation i henhold til 'best practices' gennem hele forskningsprocessen både er med til styrke forskerens arbejde, datadeling mellem forskere, bevaring af forskningsdata, citeringsmuligheder mv.
- RUBs case forholder sig til kvalitative data. Det drejer sig om typisk samfundsvidenskabelig/humanistisk empiri i form af videomateriale og transskriberet

interviewtekst. RUB har med interessetilkendegivelse fra forskningsprojektet CALPIU⁴² specifik fokus på langtidsopbevaring og adgangsstyring.

NAT cases

- Det Kongelige Bibliotek vil adressere etableringen af proveniens data som en integreret del af forskningsprocess og vil have fokus på et stort internationalt netværk omkring Kepler Data. Alle dele, primærdata, forarbejdede data, ansøgninger, anvendte programmer samt publikationer vil blive struktureret og annoteret i forhold til relevante internationale standarder.

TEK cases

- DTIC's cases ligger på det teknisk-naturvidenskabelige område. Der vil bl.a. være fokus på cases inden for astrofysik og geomagnetisme med problematikker omkring såvel data, der er indhentet af forskergruppen selv, som data, der opstår ved efterbehandling af data fra ESA ol. internationale eksperimenter og dataleverandører. Der vil også være fokus på problematikker omkring visualisering af datasæt og de bevaringsmæssige udfordringer der rejses herved. Her tænkes især på eksponering og meritering, f.eks. ifm. citation af datasæt.

Beskrivelse af de fem tværgående temaer (inspireret af DCCs livscyklus, <http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services>):

1. **Data management planning.** Der bygges på resultaterne af FIF-projektet og etableres adgang til et værktøj til udarbejdelse af data management planer. Der udarbejdes data management planer for hver case, og værktøjet evalueres på baggrund af de indhøstede erfaringer.
2. **Data capture, storage and documentation.** Der skabes repositories for de aktive forskningsdata, der understøtter metadatastandarder relevante for projektets cases, mulighed for citation af datasæt, og digitale objekter (f.eks. til formidling af data). Optimale data capture teknikker (f.eks. integration med DeIC Box) efterforskes og implementeres.
3. **Data identification, citation and discovery.** Forskerne forventes identificeret vha. ORCID-identifikatorer. Datasæt forventes identificeret ved tildeling af DataCite DOI'er og/eller DanPID'er. Metadata om datasæt høstes fra såvel lokale repositories som de nationale langtidsarkiver og gøres søgbare i nationale og internationale forskningsdatabaser. Krydslinking af forskere, forskningspublikationer, datasæt og relaterede digitale ressourcer (f.eks. visualiseringer) udforskes i paradigmatiskke eksempler for hver case.
4. **Select and deposit for long term preservation.** Der udarbejdes strategier og kriterier for udvælgelse og hensigtsmæssig pakning (eksempelvis WARC) af materiale til langtidsbevaring. Workflows og integration til bevaringsløsninger baseret på Dansk Dataarkiv og softwareudviklingen fra Det Nationale Bitmagasin implementeres.

⁴² Research Centre for the Study for Cultural and Linguistic Practices in the International University

5. **Training and marketing toolkits.** Der udarbejdes informationsmateriale, vejledninger, træningsmateriale, kursusmateriale. På de deltagende bibliotekers universiteter gennemføres informationskampagner for at udvide anvendelsen af projektets løsninger. Der vil være særligt fokus på at udarbejde skabeloner for kurser i data management for PhD'er.

Der nedsættes en temaleder til hvert tværgående tema. Opgaven er at uddrage det generelle vedr. temaet og påvirke de enkelte cases til at anvende fælles begreber og værktøjer inden for temaet, hvor det er muligt.

Den blå søjle til højre på figuren illustrerer projektets koordination med DeIC og DEFF. Her koordineres også de tværgående temaer, og her arrangeres de workshops og andre temamøder, der afvikles som en del af projektet hen over hele projektperioden.

23. Hvordan løser projektet en problemstilling, der berører hele FFU-bibliotekssektoren?

Projektet vil udgøre en pilotindsats, der støtter op om den nationale strategi for data management. Der vil blive arbejdet på to fronter, som hver især er med til at undersøge og understøtte bibliotekers rolle i forhold til forvaltning af forskningsdata, dels ved

- at forene en række nationale og internationale standardiseringsinitiativer omkring metadata med en række konkrete forskningscases til at sikre best practice i en række områder
- at arbejde med løbende dokumentation samt metoder til at sikre dette på en for forskerne så lidt omstændelig måde som mulig.

Projektet vil også gennem en aktiv formidlingsaktivitet sikre, at resultatet af studierne deles bredt i den danske forskningsverden, og dermed er med til at sikre en forståelse blandt forskerne for vigtigheden af såvel autoritative informationer som tidligt at sikre data for eventuel videre brug og publicering.

24. Hvordan understøtter projektet implementeringen af DEFFs strategi 2012-2016?

Projektet vil medvirke til at forskningsdata bliver til "managed data" og tilgængeliggøres, og i stigende grad benyttes som e-ressourcer. Projektet styrker samarbejdet mellem forskningsbibliotekerne ved at inddrage en række biblioteker i udviklingen af forskningsdata som et biblioteksområde. Projektet omfatter desuden DeIC og Statens Arkiver som nye samarbejdspartnere. Endelig vil projektet styrke tilgængeliggørelsen og formidlingen af forskningsresultater til offentligheden.

26. Hvordan skaber projektet værdi for Danmark som helhed?

Projektet vil bidrage til at give et større afkast af de danske forskningsresultater. Ved at skabe services og metoder, der giver mulighed for at opsamle, dele, formidle, videreudnytte, videreudbygge, citere, kontrollere og bevare forskningsdata kan forskere forbedre synligheden, troværdigheden og værdien af deres forskning. De får mulighed for at give offentligheden inklusiv erhvervslivet bedre adgang til forskningsresultater og dermed samtidig adgang til at aktivere deres aktive deltagelse og bidrag.

27. Hvad vil der ske, hvis projektet ikke bliver gennemført?

Bibliotekes mulige rolle i forhold til forskningsdata har været diskuteret i en årrække på strategisk plan og gennem undersøgende projekter. Manglende praktiske løsninger vil betyde at væsentlige beslutninger i forhold til biblioteks rolle bliver truffet på et rent

teoretisk grundlag. Det vil gøre det vanskeligt at fundere en videre udvikling i forskningspraksis og sætte bibliotekerne bagud i forhold til andre aktører på området.

28. Har andre nationalt eller internationalt arbejdet med samme emne?

Projektet er inspireret af etablerede data management infrastrukturer i lande som USA (Stanford), England (Oxford, Edinburgh), Holland (DANS, 3TU.Datacentrum, SURF), Tyskland og Australien (ANDS). Der har også været gennemført behovsafdækkende projekter (bl.a. via SURF). Projektet vil erfaringsudveksle med kolleger fra flere af disse lande, og hvor det er muligt anvende open source værktøjer med oprindelse i disse landes initiativer. På nationalt plan bygges der videre på erfaringerne fra tidligere DEFF-projekter på området.

29. Målgruppe

Den primære målgruppe er de forskere, der er inddraget i de forskellige cases. Både dem som producerer og dem som aftager forskningsdata. Af sekundære målgrupper kan nævnes

- 1) Forskningsbibliotekerne, som skal finde deres rolle i data management livscyklussen og
- 2) Universiteterne og de enkelte forskningsområder, som tilbydes services indenfor data management.

30. Hvilke målbare succeskriterier opstilles for projektet, og hvordan skal disse vurderes?

1. For hver case konstateres det at der er etableret tjenester til alle livscyklustrinene.
2. Det konstateres at hver case har gennemgået en accepttest blandt forskerne, og at resultatet af testen er dokumenteret. Yderligere forslag, behov og ønsker fra forskerne indgår i dokumentationen.
3. For hver af de tværgående temaer evalueres der på hvilke dele af de udviklede tjenester der egner sig til samlede nationale løsninger inden for temaet, på tværs af forskningsområder. Det konstateres at denne evaluering er dokumenteret i en rapport.
4. Det konstateres at der er blevet udviklet træningsmateriale på nettet til fælles brug på tværs af alle projektets cases.
5. Det konstateres at der er lagt en plan for markedsføring og outreach i udvalgte forskergrupper.

31. Har projektet en blivende effekt? Hvordan sikres projektets leverancer i en driftsfase? Og hvordan har projektet haft en effekt på/for sektoren som helhed efter afslutning?

Det forventes af de løsninger som benyttes, indkøbes eller udvikles vil fortsætte efter projektets afslutning. Institutionerne bag projektet garanterer, at publicerede og langtidsbevarede forskningsdata kan tilgås i minimum 10 år. Herudover vil den indhøstede viden indgå i det videre arbejde med at etablere nationale løsninger for forskningsdata, med strategiudvikling på området, med kompetenceopbygning og i det hele taget med at positionere bibliotekerne.

32. Tidsplan

Projektet løber fra august 2014 til juli 2016.

33. Formidling af resultater

Tema 5, *Training and marketing toolkits*, har som mål at etablere træning og information i forhold til projektets primære målgruppe, forskerne. Herudover vil der blive udviklet en kommunikationsplan som fastlægger niveauet af formidling til øvrige interessenter. Formatet og omfanget på den afsluttende resultatformidling vil blive aftalt nærmere i foråret 2016.

34. Samarbejdspartnere. Hvilke samarbejdspartnere er der indgået aftale med, og hvilke samarbejdspartnere ønskes der på sigt at inddrage i projektet?

Der er indgået aftale med alle medansøgerne. Herudover forventes DigHumLab inddraget i de to humanistiske cases.

Økonomi**35. Budget**

	mandtid	ansøgning	egenfinan- siering	total
SB - HUM cases (inkl. lokal server og storage)	24 MM	1,000,000	300,000	1,300,000
KB - HUM + NAT cases (inkl. lokal server og storage)	24 MM	1,000,000	300,000	1,300,000
DTIC - TEK cases (inkl. lokal server og storage)	24 MM	1,000,000	300,000	1,300,000
SDU - SAMF/SUND cases	4 MM	150,000	50,000	200,000
RUC - SAMF/SUND cases	4 MM	150,000	50,000	200,000
DDA - SAMF/SUND cases	4 MM	50,000	150,000	200,000
AUB	4 MM	150,000	50,000	200,000
CBS	4 MM	150,000	50,000	200,000
Projektledelse	4 MM	200,000		200,000
DeIC - koordinering	6 MM		300,000	300,000
Møder + Rejser		100,000		100,000
Website + kommunikation		100,000		100,000
Total		4,050,000	1,550,000	5,600,000

Budgettet er for enkelte mindre poster foreløbigt og vil blive endeligt fastlagt i en eventuel ansøgning. Forskerinddragelse i de forskellige cases betales af case-budgettet således at SB, KB og DTIC selv laver aftaler om evt. økonomisk kompensation ift. projektdeltagelsen.

Generelt**41. Bemærkninger**

De opsummerende punkter 36-40 vil blive udfyldt i en eventuel egentlig ansøgning.

Dato: 1. november 2013

Underskrift af den økonomisk ansvarlige

Birte Christensen-Dalsgaard