Notat

Datamanagement i Danmark

15. september 2019

Udarbejdet af et udvalg bestående af:

AAU Prodekan Torben Larsen
AU Projektleder Birte Christensen-Dalsgaard (formand)
CBS Chefkonsulent Lars Nondal
DTU Sektionsleder Michael Rasmussen
KU Professor Kasper Møller Hansen
RUC Vicedirektør for Digitalisering Galina Ianchina
SDU Institutleder Peder Thusgaard Ruhoff

samt

DeiC COO for datamanagement Anders Sparre Conrad (sekretariatsbetjening)

Gruppen har undervejs fået assistance fra:
Ulrik Sørensen Rohde, KU

Opgavebeskrivelse og baggrund

DeiCs bestyrelsesmøde den 26. april 2019 havde de enkelte universiteters visioner på datamanagement på dagsordenen. I det udsendte referat står:

I forbindelse med punkt 3 blev bestyrelsen enige om at arbejde med en definition, hvor "storage" er det system, der anvendes til opbevare data, og "datamanagement" er de lag, der ligger ovenpå som services, politikker o.lign. De enkelte bestyrelsesmedlemmer fremlagde visionen for det nationale niveau af storage og datamanagement set fra deres universitet.

Fremlæggelserne viste et differentieret behov og forventninger til det nationale samarbejde. Der skal nedsættes en arbejdsgruppe, der kan se nærmere på hvilke aktiviteter, det vil give mening at gøre i fællesskab, jf. yderligere behandling under punkt 6 og 8."

Efter bestyrelsesmødet blev der udarbejdet et kommissorium for bl.a. en arbejdsgruppe, som skulle have fokus på datamanagement og storage. I henhold til kommissoriet (bilag 1) skal arbejdsgruppen senest den 15. september 2019 levere en rapport, der belyser følgende områder:

- Hvilken storage-infrastruktur skal være tilgængelig på nationalt niveau for at understøtte FAIR-data og open science og dermed øget tværgående samarbejde?
- Hvordan sikres sammenhæng mellem lokale, nationale og internationale løsninger, herunder aktiv dansk deltagelse i EOSC-initiativer?
- Hvilke trin i datas livscyklus skal dækkes af nationale services, herunder tanker om understøttende support og kompetenceudvikling?
- Hvilke udviklingsressourcer og hvilken organisering er der behov for for at sikre en bæredygtig national udvikling i forhold til den internationale?
- Anbefalinger til det nationale samarbejde i DM Forum, herunder om det skal fortsætte og i givet fald med hvilke opgaver og budget.
- Hvordan håndteres deltagelse i internationale medlemsskaber og samarbejde og med hvilket engagement og investering?
- Hvordan sikres national videndeling af internationale resultater og aktiviteter med henblik på anvendelse såvel teknisk som forskningsmæssigt?

Indholdsfortegnelse

1. Anbefalinger

Anbefalingerne falder i fire grupper:

- En generel anbefaling om at udvikle en generel tjeneste til målgruppen af de mange enkeltforskere eller forskere, som arbejder i mindre forskningsgrupper og i mindre digitale forskningsfelter
- En anbefaling om udvikling af nationale tjenester og etablering af nationale storage-faciliteter overfor alle
- Anbefalinger omkring nationale licensindkøb og aftaler med potentielle leverandører
- Og endelig, anbefalinger om organisering og relationer til internationale projekter, herunder EOSC.

Det anbefales at udvikle en række grundlæggende tjenester til forskere med begrænset behov.
Gruppens arbejde og dialog med forskere og institutioner i de sidste par måneder viser, at der er en meget stor gruppe forskere specielt fra humaniora og samfundsvidenskab, som har nogle ret små krav, som i dag ikke ser nogen løsning på deres datamanagement-behov. De vil typisk have behov for nemme løsninger til at gemme deres resultater, udveksle disse med forskere i ind- og udlandet og gemme data med en DOI tidligt i forskningsprocessen.

Det anbefales at undersøge om der er interesse for at indkøbe nationale licenser til en række værktøjer til eksempelvis elektroniske labboks eller business intelligence miljøer/værktøjer.

Det anbefales at tilbyde et integreret system, hvor forskere kan styre projekter gennem de forskellige faser og sikre, at data og tilhørende beskrivelser lagres forsvarligt, og at dette nemt kan udveksles med andre forskere inden for og uden for egen institution.
Et sådant nationalt system kan også være med til at fremme mobiliteten. Det anbefales at identificere minimumskrav og undersøge muligheden for at implementere en eksisterende løsning. Det er centralt at forskernes behov og incitamenter for at benyttet et fremtidigt system tænkes ind i systemet fra start. Et sådant processystem vil trække på en række af de efterfølgende tjenester.

Det anbefales at etablere en fælles søgetjeneste for data.
Vores undersøgelse afdækker et behov for en brugervenlig, national søgetjeneste, som går på tværs af alle danske registre med beskrivelse af forskningsdata. At gennemføre dette kræver, at alle registre overholder standarder, som muliggør en sådan udveksling af data. Efterfølgende skal en sådan tjeneste implementeres – i første omgang med data fra de registre, som er klar til at synkronisere med den nationale tjeneste. Der bør påbegyndes et arbejde, som dels skal afdække relevante standarder for udveksling af metadata til en national søgetjeneste, dels se på, hvilke registre som ville kunne levere strukturerede data til en samlet søgetjeneste.

Det anbefales at etablere trusted repository-funktioner på nationalt niveau.
Der efterspørges muligheder for at gemme datasæt nationalt i trusted repositories med

permanente DOI'er, som alternativ til eller parallelt med, at disse afleveres til tidsskrifterne. Disse skal også understøtte andre standarder for PID'er, herunder brugen af ORCID.

Det anbefales at tilbyde "sync & share" tjenester nationalt.
Forskeres behov for at dele både små og store mængder data skal understøttes – både i form af lager, hvor forskeren har fulde administrative rettigheder, og i form af sikker fildeling via nettet. Der er i dag løsninger, men det anbefales at undersøge muligheden for også at kunne benytte kommercielle tjenester som Dropbox til mindre datamængder.

Det anbefales at etablere tilstrækkelig volumen af nationalt tilgængelig storage.
Forskernes behov for lager stiger eksponentielt som resultat af dataopsamling fra nye og forbedrede digitale metoder. Det er derfor vigtigt at forhandle gode vilkår for adgang til storage, som tilfredsstiller forskellige parametre som hurtig vs. langsom adgang, stor vs. lille kapacitet og sikkerhed. Det skal gøres intuitivt og nemt for forskerne at migrere fra én platform til en anden.

For at dette kan gennemføres forsvarligt, anbefales det at gennemføre en grundig undersøgelse af de nuværende behov for og krav til lagerplads, som dækker alle fakulteter og forskningstraditioner, fordelt på behov for:

- lager i forbindelse med beregninger
- lager, som bruges til at gemme data – enten fordi de aktuelt ikke er basis for en analyse, eller fordi de gemmes af hensyn til eventuel genbrug.

Specielt behovet i forbindelse med opbevaring af data, der gemmes af hensyn til genbrug er i dag dårligt belyst. Kortlægningen af behov skal afdække krav fra alle fakulteter og forskningstraditioner og fremtidige strukturer bør tage udgangspunkt i forskernes behov og gennemtænke forskernes incitamenter til at benytte fremtidige tjenester.

Det anbefales at implementere et politik-implementeringslag mellem lager og applikationer.
I lighed med Holland kan det gøres i iRODS. Det anbefales at implementere nogle af de velfungerende systemer, som anvendes i Holland såsom B2SAFE og YODA, for dermed at opbygge relevante erfaringer og få den nødvendige erfaring til at træffe et valg.

Det anbefales at Back Office etableres som en professionel organisation med en faglig tyngde.
De nødvendige kompetencer bør identificeres, og det bør besluttes, hvor disse er forankret – i et fagligt miljø eller i en central institution. Når kompetenceprofilerne er identificeret, bør stillingerne slås op.

Det anbefales at DM Forum fortsætter som kompetenceudviklings- og samarbejdsorgan for Front Office.
DM Forum binder deltagerne i de forskellige instanser af Front Office sammen. Det er vigtigt, at disse får veldefinerede rammer og budget til at gennemføre projekter og gennemføre den nødvendige kompetenceopbygning.

Det anbefales at aktiviteter omkring langtidsbevaring udsættes til senere i strategiperioden.

2. Indledning til rapporten

I EU-Kommissionens rapport *Turning FAIR Into Reality*[1] benyttes begrebet **FAIR Digital Objects** som grundlag for implementeringen af et teknisk økosystem. Det digitale objekt består af data, kode og andre forskningsoutputs og defineres således:

"At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse."

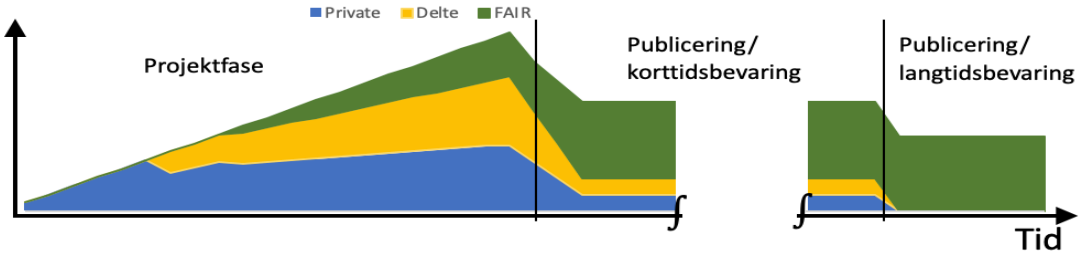
Formålet med datamanagement er at sikre integritet, sikkerhed og tilgængelighed af data, og et mål er at ikke kun data, men hele forskningsobjektet gøres tilgængeligt. Det sker ved at udstikke politikker for håndteringen som implementeres gennem brug af en række systemer, som understøtter disse. Udover at sætte fokus på kvalitet og sikkerhed i håndteringen af forskningsdata handler god datamanagement om, at relevante metadata etableres, således at forskningsresultatet er reproducerbart, og således at data og kode kan genbruges.

De metadata, som ikke autogenereres, vil typisk blive tilføjet i løbet af forskningsprojektet. I et typisk forløb går forskningsdata fra at være private over at blive delt med fagfæller til at blive publiceret. Metadata kan inddeles i flere kategorier som eksempelvis deskriptive, administrative, tekniske og strukturelle. De er basis for et FAIR økosystem som fuldt udfoldet vil muliggøre maskine til maskine anvendelse af alle åbne elementer af forskningsobjektet. Vi forventer, at det fortsat vil være nødvendigt med en menneskelig vurdering og specifikke udtræk i forhold til eksempelvis personfølsomme data.

Det vigtige er, at metadata er så åbne som muligt, da de er indgangen til, at alle elementer af forskningsobjektet kan genfindes og genbruges via en eller flere søgeservices.

Nedenfor er vist et skematisk forløb af data, hvor data indsamles, deles og publiceres i løbet af projektperioden – og hvor der sker en vis kassation efter projektets afslutning.

[1] Turning FAIR into reality, Final report and action plan from the European Commission expert group on FAIR data, https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf. Citat er fra s. 35



Karakteristika	Projektet er aktivt, der indsamles data, som initielt bliver lagt på lokale servere. Efterfølgende bliver delemængder dlet med andre forskere i lukkede fora og endelig bliver de gjort FAIR ved at metadata publiceres	Projektet er afsluttet. Forskeren arbejder med andre problemstillinger. De data, som er nødvendige for at kunne reproducere resultater gemmes også selv om disse ikke er gjort FAIR	Bevaring af hensyn til Code of Conduct er overstået. Programmer, som er brugt ved produktion, findes måske ikke længere.

Figur 1 Illustration af de tre faser, som data i forbindelse med projekter gennemlever

Som illustreret arbejder vi med tre faser i forhold til datas livscyklus, 1) projektfasen, hvor data er aktive og skal være nemt tilgængelige, 2) en korttidsbevaringsfase, hvor data gemmes af hensyn til reproducerbarhed (Code of Conduct), og 3) langtidsbevaring, hvor data skal være tilgængelige, længe efter at de værktøjer, som blev brugt til at producere dem, er afløst af andre. Kravene til storage varierer også gennem de tre faser – fra krav om hurtig adgang i projektfasen til langsomme, stabile og billigere lagermedier ved langtidsbevaring. Dette vender vi tilbage til senere.

Til udformning af dette notat har det kun i begrænset omfang været muligt at foretage informationsindsamling fra universiteterne. Vi forventer at dette rettes op gennem en høringsfase bl.a. hos universiteterne. Vi har indhentet oplysninger fra relevante ressourcepersoner på universiteterne samt foretaget en mindre undersøgelse på Københavns Universitet (Bilag 2) og brugt en tidligere brugbarhedsanalyse fra Aarhus Universitet (ikke offentlig). Derudover trækkes der på erfaringer blandt udvalgets medlemmer, diskussioner i DM Forum samt tidligere rapporter og projekter m.m., primært DeiC-rapporten "Digital infrastruktur til forskning i verdensklasse i 2025"[2] fra 2017 (DI2025 i det efterfølgende) og i mindre omfang FAIR-rapporten "Foranalyse: Indførelse af FAIR data i Danmark"[3] ligeledes fra 2017 (FAIR2017).

Nedenfor gennemgås kort nogle af hovedobservationerne.

Rapporten DI2025, som var en del af strategiprocessen, der førte frem til "Strategi for nationalt samarbejde om digital forskningsinfrastruktur" fra december 2018, indeholder en "Behovskortlægning" inden for de videnskabelige hovedområder baseret på forskerinterviews og inputs fra universiteternes dekaner[4].

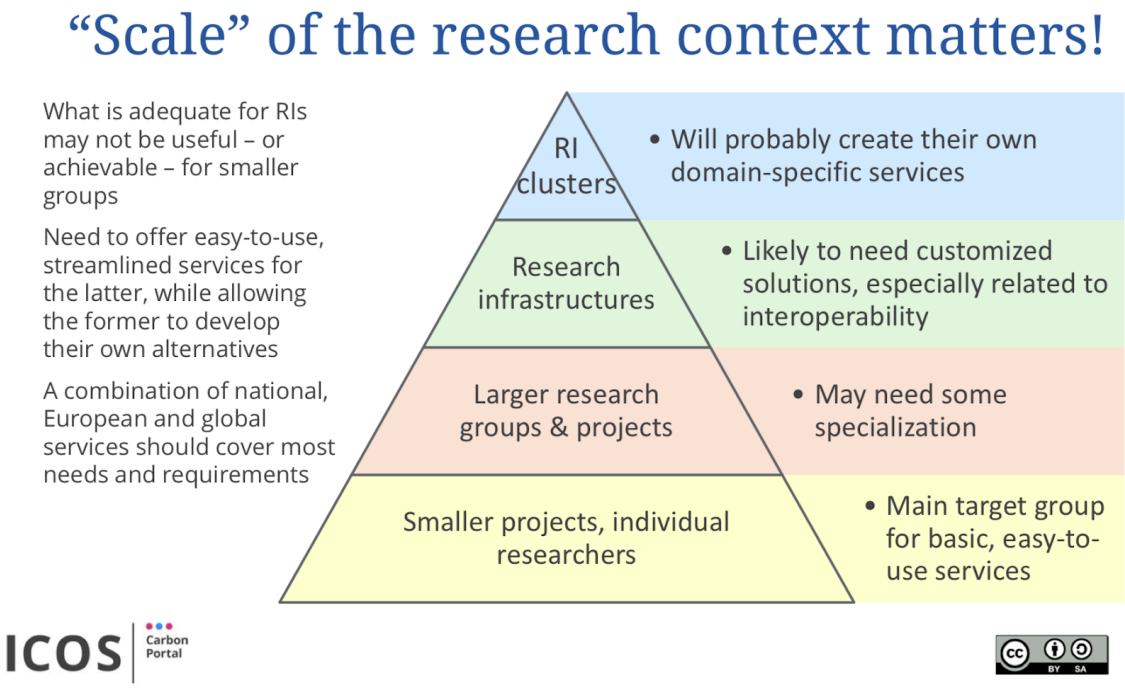
[2] https://www.deic.dk/da/analyserapport-digital-infrastruktur-til-forskning-i-verdensklasse-2025
[3] https://ufm.dk/publikationer/2018/foranalyse-indforelse-af-fair-data-i-danmark
[4] Se s. 10-21 i rapporten samt Bilag til rapporten, s. 161-171.

Det kan være svært definitivt at adskille de udtalte behov i hhv. 'storage' og 'datamanagement'-behov, da der kan være et stort overlap mellem disse.

Vores undersøgelser og review tyder på, at:

- Forskerbehov i forbindelse med storage var ofte forekommende.
- Forskerbehov ved datamanagement er lidt sjældnere forekommende og ofte mere abstrakte og principielle.
- Den samme løsning kan ikke bruges til alle og løsninger vil ofte afhænge af forskningsfeltet.
- Fagspecifikke nationale og internationale forskningsinfrastrukturer spiller en stor rolle.
- Størrelsen af forskningsgrupper og infrastruktur vil have betydning for løsningsbehov.

En uddybning af den sidste pointe er givet i nedenstående figur 2, som stammer fra en workshop om datamanagementpolitik i Stockholm. Her viste Margaret Hellström nedenstående figur:



Figur 2 - Størrelsen af forskningsfeltet har betydning for løsningsbehov, Margaret Hellström

"Scale matters" og en diskussion af understøttelse af dansk forskning er derfor også en diskussion om, hvem der skal understøttes og hvordan.

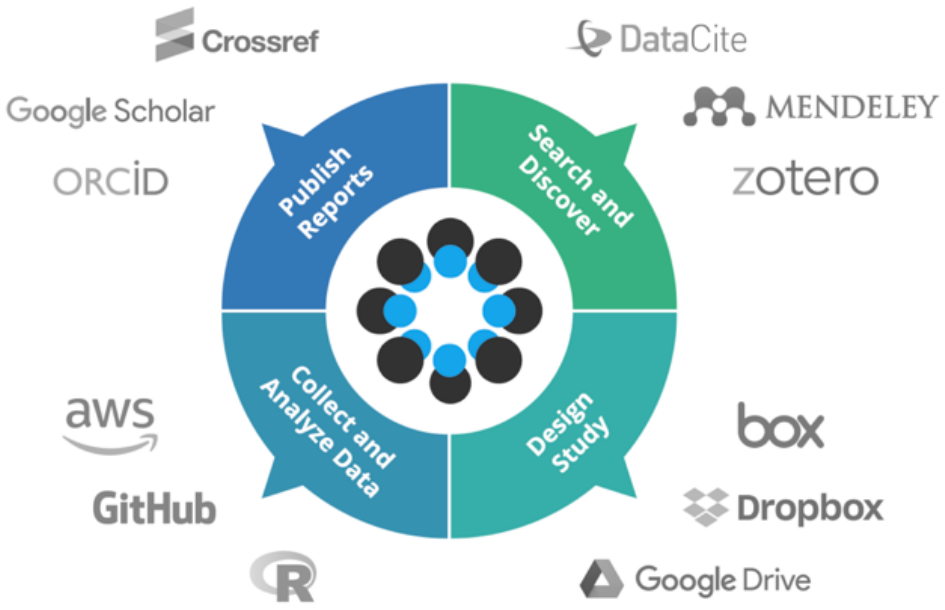
Undersøgelser af krav og forventninger, som beskrives i næste afsnit, udviser et meget komplekst landskab – fra ønsker om adgang til lager til ønsker om komplette nationale datamanagementløsninger. Skal en national løsning være attraktiv, skal den klart føre til en rationalisering af arbejdsopgaver og resultere i en samlet set bedre og billigere løsning end en række lokale løsninger.

3. Datamanagement i forhold til forskningslivscyklus

Flere har udtrykt et ønske om et samlet, integreret workflow-system, hvor forskeren aktiverer en række værktøjer som en integreret del af forskningsprocessen. Fokus er ikke alene på de enkelte værktøjer, men også på interoperabilitet mellem værktøjerne, dataflow og den dokumentation, som genereres. I den ideelle verden fanges og dokumenteres processer, efterhånden som projektet har fremdrift. Virkelighedens verden er dog en anden, hvor disse systemer ofte er selvstændige systemer, og hvor det er forskeren, som skal skabe sammenhæng.

Flere datadrevne forskningsretninger har allerede etableret dele af en sådan sammenhæng som det ses i nogle ESFRI infrastrukturer eller gennem brug af kommercielle/open source produkter som RedCap og Jupyter Notebooks. I lighed med situationen i Finland, kan man overveje at indkøbe nationale licenser til de mest brugte værktøjer af denne art.

Som et eksempel på de forskellige værktøjer, som kan indgå i et process-workflow vises nedenfor Open Science Framework, OSF, som er et gratis open source-processtyringsværktøj, som understøtter forskerne i hele forskningsprocessen.



Figur 3 - Open Science Framework[5]

[5] https://cos.io/our-products/osf/

3.1. Review af implementeringen af datamanagement i Holland

Som indledning til arbejdet blev situationen i Holland analyseret, da mange tjenester her udbydes i samarbejde mellem universiteter, som fx 4TU-samarbejdet. Resultatet af analysen er beskrevet i bilag 3. I forhold til understøttelse af forskningsprocessen kan vi konstatere:

Der bruges et mix af egenudviklede lokale løsninger, centralt udbudte løsninger samt internationale løsninger. De nationale løsninger kan være drevet af SURFsara eller DANS, men kan også udbydes og vedligeholdes af andre institutioner – enten i nationale samarbejder som 4TU.Centre for Research Data eller gennem af internationale infrastrukturer.

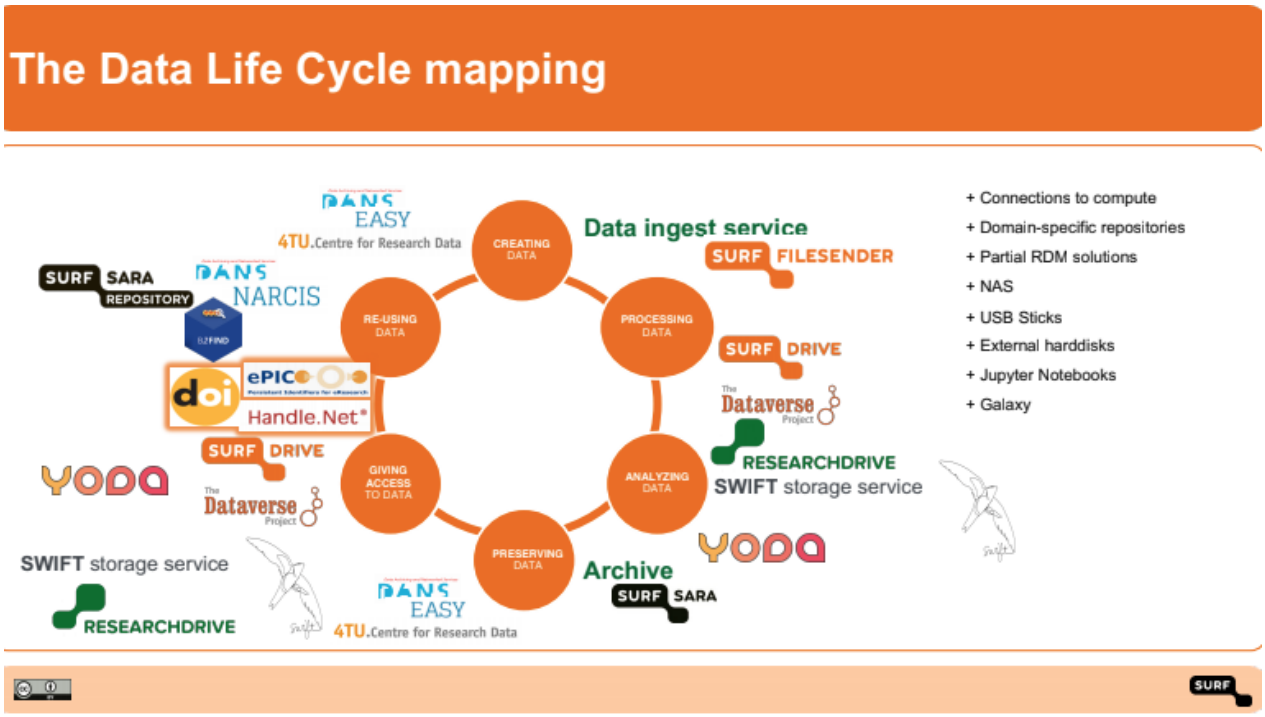
Der er en differentieret tilgang til lagring af data, som afspejler krav til åbenhed og størrelse af datasæt. Det skal her bemærkes, at man i Holland prioriterer at certificere repositorier. Der er 16 repositorier, som har CoreTrustSeal eller lignende. Til sammenligning er der i Danmark kun tre repositorier, som er certificeret.

Inspireret af situationen i Holland foreslås, at følgende typer tjenester kunne være relevante i en dansk kontekst. De tjenester, som udbydes i et fællesskab nationalt og internationalt, er markeret med fed skrift. Det skal bemærkes at enkelte af de tjenester der udbydes i Holland også allerede er til rådighed som nationale tjenester i Danmark, såsom DMPonline og Datacite DOI.

Generisk tjeneste	Eksempler på tilbudte løsninger i Holland
Data Management Planning	**DMPonline**
Datakatalog	Et nationalt datakatalog for forskningsdata med søgefacilitet (**NARCIS**). Suppleres med brug af bl.a. R3Data.org
Workflowstyring	Mulighed for at få systemet til at understøtte de forskellige faser (f.eks. **YODA**). Brug af iRODS synes at fungere godt
Udlevering og styring af identifikatorer (DOI, ORCID og andre)	Integreret
Deling af filer	Der udbydes forskellige services, der muliggør effektiv overførsel af store datasæt (**HPN-SSH eller GridFTP, Data Migration Facility DMF. SURFfilesender**)
Storage af forskningsdata	Lagring og deling af forskningsdata – løsning afhænger af type og størrelse af data (f.eks. **DataverseNL, Object Store, Data Archive, figshare, 4TU.Centre for Research Data, Zenodo**)

Langtidsarkivering	Certificeret langtidsarkivering af forskningsdata. Understøtter DOI og standard Dublin Core-metadata. Det er ikke muligt at strukturere data hierarkisk eller versionere (f.eks. **EASY**)

Den hollandske løsning er samlet i nedenstående tegning, hvor fokus er på de tjenester, som understøtter datahåndteringen:



Figur 4 - Den hollandske løsning. Kilde: SURF

3.2. Status for implementeringen af datamanagement i Danmark

Alle danske universiteter har etableret en form for og et niveau af RDM Support. Organiseringen er forskellig, som oftest i krydsfeltet mellem bibliotek, it-afdeling, forskningsstøtteenheder, Open Science-enheder, juridiske afdelinger mm.

Næsten alle danske universiteter har vedtaget lokale politikker vedr. forskernes håndtering af data. Scope for disse politikker kan variere, men typisk henviser de til Code-of-Conduct. Flere, men ikke alle, refererer til FAIR-principperne.

På nationalt niveau konstaterer vi, at der ikke er formuleret en national strategi for FAIR, selv om en udarbejdelse af en sådan er anbefalet i strategien for forskningsinfrastruktur. Så vidt vi er orienteret, stilles der ikke eksplicitte krav til at data skal være FAIR fra forskningsrådene, hvorimod deltagelse i H2020-projekter kræver deltagelse i Open Data Pilot, herunder publicering efter FAIR principperne. Dette krav forventes forstærket i Horizon Europe-programmet.

Hvad angår forskernes udtalte RDM-behov, er disse typisk mere principielle i form af tilslutning til at data skal kunne deles og i øvrigt være så FAIR som mulig. Den principielle tilslutning følges ofte op med et stærkt behov for mere support til den praktiske håndtering af den slags krav. Et andet eksternt krav, der udfordrer forskerne i den daglige datahåndtering, er GDPR.

Forskerne efterlyser hjælp i form af både services/direkte support og systemer til at håndtere det samlede workflow i hele datas livscyklus, dvs. den praktiske håndtering og de forskellige arbejdsopgaver på de forskellige trin, fra den indledende datafangst eller dataanskaffelse, over analysefasen og publicering til langtidsbevaring (eller sletning).

Vedrørende RDM-services/-support efterlyses generelt bedre og nemmere tilgængelig, lokalt forankret support. F.eks. til praktiske opgaver som tildeling af DOIs til datasæt ved publicering, metadatering, publiceringslicenser, arkivering, valg af repositorium m.m.

Det er sjældent det ord, forskerne bruger, men på service-/support-niveauet svarer dette til, at de indirekte efterlyser hjælp og support fra ”data stewards”, der er lokalt forankrede og har forståelse for fagområdets datalivscyklus og en generel indsigt i fagområdet/domænet. Der findes flere forskellige definitioner af, hvad “data stewardship”, se TU Delfts definition i bilag 3, afsnit 6.

Den enkelte forsker ønsker ikke en fragmenteret infrastruktur bestående af en række adskilte komponenter, men en så sammenhængende struktur som muligt, med smidige og sømløse overgange fra den ene fase til den anden, og hvor forskeren ikke behøver at vide eller interessere sig for, om den enkelte tjeneste stilles til rådighed lokalt eller nationalt, evt. internationalt. Specielt i forbindelse med GDPR efterspørges beslutningsstøttesystemer og positivlister eller lignende af godkendte systemer, der kan hjælpe forskeren til en forsvarlig håndtering af persondata.

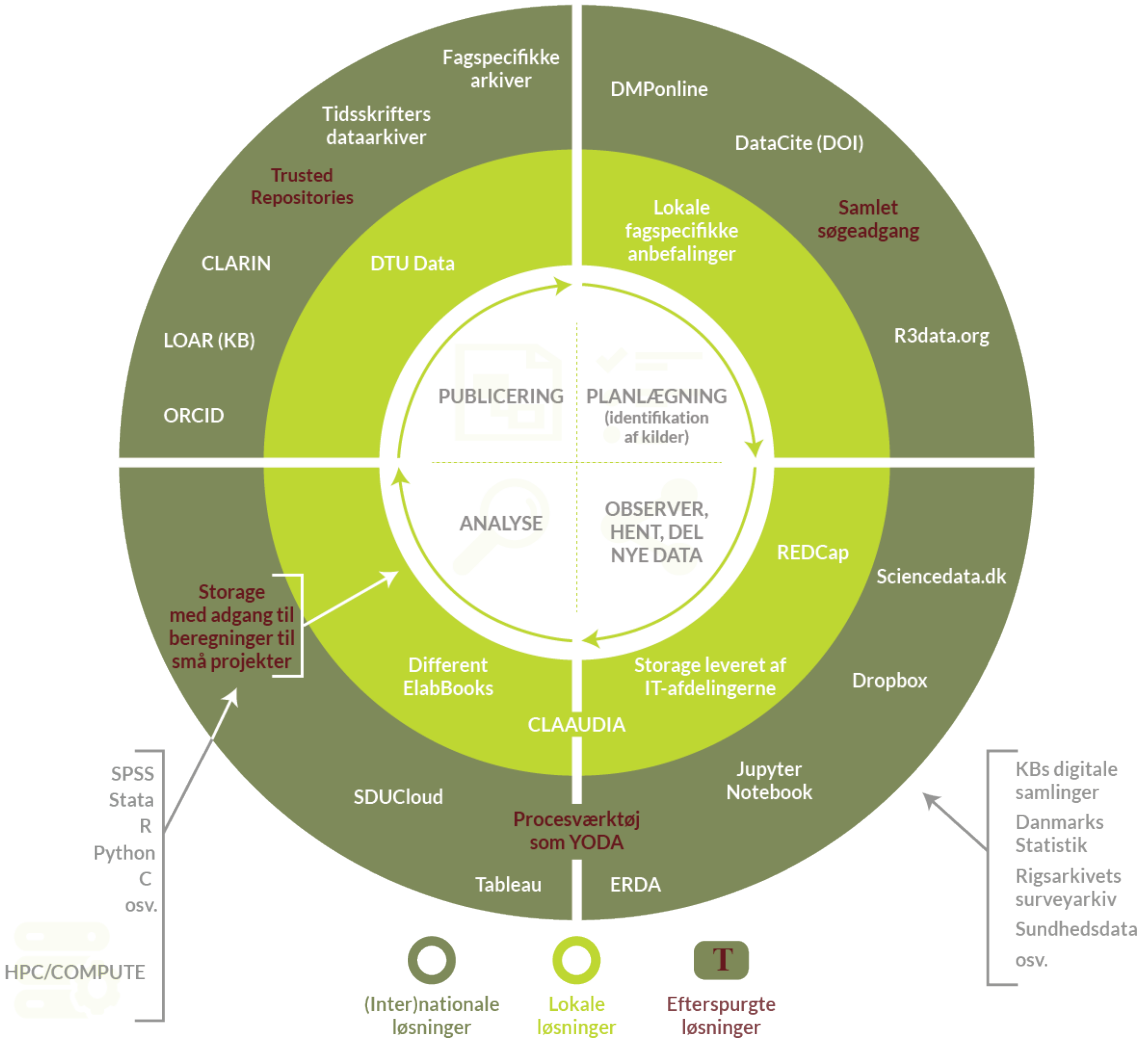
Ideelt set bygges en sådan national platform på lokale løsninger, der "fødereres" til en fælles platform, der sikrer overholdelse af FAIR-principperne, sikrer sikker deling af data med samarbejdspartnere og er sømløst forbundet med andre relevante infrastrukturer (HPC, nationale arkiver, ESFRIer, EOSC etc.).

Der gives eksplicit udtryk for, at der fra forskernes synsvinkel er brug for datalagring og datadeling *”as a service”* – dvs. fleksibelt, skalerbart og nemt for forskeren. I DeiC-rapporten s. 21 formuleres det som et ønske om *”udvikling af eScience/e-infrastruktur som en service”*.

Dette perspektiv udfoldes nærmere i arkitekturafsnittet nedenfor.

3.3. Mulige danske komponenter

Som bemærket i FAIR-rapporten (FAIR2017) er Danmark præget af et meget fragmenteret landskab når det kommer til mulige løsninger indenfor RDM. Nedenfor har vi vist eksempler på eksisterende løsninger efter samme model som for Open Science Foundation (fig. 3). Vores undersøgelse har identificeret nogle oplagte mangler, som er indikeret med rød.



Figur 5 – Et bud på komponenter til en dansk løsning baseret på OSF-modellen. Den ydre cirkel er nationale og internationale løsninger eller tjenester, den inderste er lokale løsninger eller lokalt implementerede løsninger. Ikke eksisterende løsninger, som er efterspurgt, er vist med rødt.

Aktørerne i forhold til løsninger er, ud over universiteternes IT-afdelinger og biblioteker, også kommercielle virksomheder, EOSC serviceudbydere samt nationale og internationale forskningsinfrastrukturer. Bl.a. er Danmark medlem af en række ESFRI-forskningsinfrastrukturer, som har udviklet fælles datainfrastrukturer inden for forskellige forskningsområder. Det gælder eksempelvis CLARIN ERIC til sprogforskning, ICOS ERIC og dens Carbon Portal til klimaforskning og ELIXIR til forskning inden for life science. Det bør undersøges, i hvilket omfang denne type infrastrukturer vil levere tjenester til danske forskere generelt og på hvilke vilkår.

De løsninger, som er skitseret ovenfor for storage og publicering, er typisk det, vi vil kalde korttidsbevaring. Opgaven om langtidsbevaring kræver en række yderligere funktioner og systemunderstøttelse såsom formatkonvertering og strukturbeskrivelse, som sikrer, at data kan fortolkes også om 50 år. Langtidsbevaring forudsætter givetvis at

korttidsbevaringsfunktioner er etableret og velfungerende på nationalt niveau. Gruppen anbefaler derfor at løsning af langtidsbevaring bliver udskudt til senere i strategiperioden.

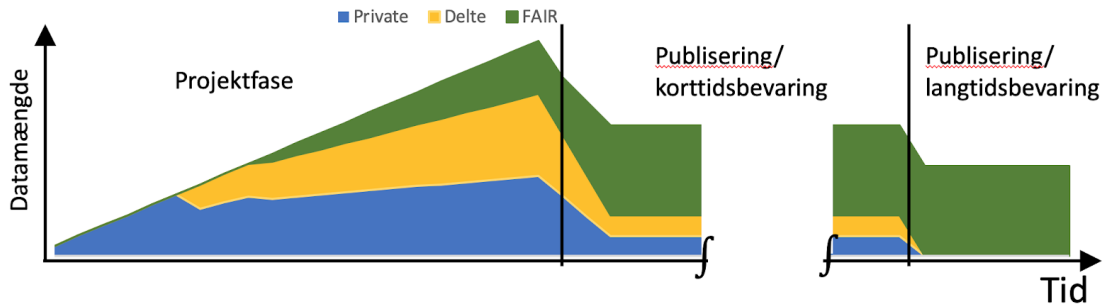
Der er indenfor kulturbevaringinstitutionerne fokus på disse opgaver, og det vil være naturligt, at de spiller en aktiv rolle i denne fase. Vi forventer at en række af disse opgaver vil blive løst internationalt.

4. Storage-infrastruktur

Inden for storage centrerer forskerbehovene sig omkring problemer med håndtering af stærkt stigende datamængder og en langt større variation af datatyper (flere nye datatyper). Institutioner med egne lagringsløsninger ser oplagt et mindre behov for etablering af nationale lagringsløsninger, mens mindre institutioner uden egen lagringsløsning meget gerne ser nationale løsninger.

Data kan karakteriseres på mange måder. I dette notat har vi valgt to uafhængige dimensioner, som har betydning for dels det administrative lag og dels typen af storage. Det ene er, om data er åbne, om de deles, eller om de er FAIR. Det andet er, om data aktivt bruges, om de korttidsbevares, eller om de skal bevares efter de 5-15 år, som foreskrives af code of conduct.

Nedenfor er vist samme skematiske forløb af data som i figur 1, hvor data indsamles, deles og publiceres i løbet af projektperioden – og hvor der sker en vis kassation. Her suppleres med karakteristika for storage i de forskellige faser.



Storage				
Hastighed	Hurtig til middel	Middel til langsom	Langsom
I/O	Stor kapacitet til middel	Middel til lav	Lav

Figur 6 – En illustration af dataproduktion i et projekt. Figuren, som har samme X-akse og princip for farvekode som i figur 1 (den rød- og grønternede er her helt grøn), illustrerer dels den blandede situation i projektfasen – og at kun FAIR-data bør langtidsbevares. Der sker ofte en migration af data mellem de forskelige faser, da krav til lagermedier ændres.

4.1. Universiteternes forventninger og behov

Arbejdsgruppens deltagerne har, så godt det har været muligt, fået input fra deres respektive universiteter. Dette er af oplagte årsager på meget forskellig form, men viser tendenser, som efterfølgende bør undersøges for at give et mere solidt beslutningsgrundlag.

Vores vurdering på basis af de informationer, som er samlet, er at Danmark meget snart når op på 100 PB. Mange har peget på en eksplosion i lagerbehov grundet de nye instrumenters meget højere opløsning, nye digitale observationsmetoder inden for alle felter og på grund af nye strategier for indsamling og behandling af observationer, hvor kvantitativ forskning nu supplerer kvalitative observationsmetoder. Forventningen er derfor, at dette tal vil være mangedoblet om fem år (eksempelvis har SUND på AU i dag 2 PB data og forventer i størrelsesordenen 50 PB om 5 år). Meget data behøver ikke ligge på ultrahurtige diske; de kan loades ind i hurtigt lager, når de er basis for analyser.

Til projektfasen efterlyses lagringsmuligheder, der muliggør aktiv deling, primært med aktive samarbejdspartnere i analysefasen (sync and share/Dropbox-lignende), men også senere deling af færdige forskningsdata (publicering med DOIs, trusted repositories).

Forskerne vil meget gerne kunne blive ved med at bruge Dropbox – af hensyn til brugervenligheden, ikke mindst i forbindelse med samarbejdet med andre forskere. Alternativt skal der kunne stilles et tilsvarende system, med fuldstændig samme funktionalitet og brugervenlighed, til rådighed. Gerne et kommercielt produkt. Der efterlyses en undersøgelse af, om det er muligt at lave en national licensaftale med Dropbox for alle danske forskere med mindre datasæt. Dette er allerede muligt gennem Dropbox Education.

Der efterlyses sikre lagringsløsninger for personfølsomme data/GDPR, og der efterlyses løsninger, der sikrer overholdelse af informationssikkerhedskrav og compliance med f.eks. ISO 27001.

Specielt for SUND og NAT/TEK er der et ønske om at sikre en infrastruktur med tæt sammenhæng mellem storage/lagring og computing, da datamængderne her kan være så store, at data ikke nemt kan flyttes fra lager- til computing-facilitet. Et forhold, der på sigt også forventes at sprede sig til SAMF og HUM, efterhånden som disse discipliner vil kræve håndtering af meget større datamængder.

I den forbindelse efterlyser både SAMF- og SUND-forskere, at det undersøges, om data fra Danmarks Statistik, Sundhedsdatastyrelsen og andre offentlige institutioner kan overføres i krypteret form til de nationale HPC-anlæg.

Men det vurderes forskelligt, om løsninger på disse storage-problemer skal ske lokalt eller nationalt. F.eks i DI2025 (s. 23): “For nogle giver det mest mening, at behovet for lagring løses lokalt, mens andre ser et behov for fælles løsninger, blandt andet for at understøtte opbrydningen af datasiloer.” Og videre (s. 23): “Lagringsfaciliteter kan med fordel etableres nationalt, når det giver mening for at:

- Sikre kritisk masse (lokale løsninger kan blive for dyre for de enkelte projekter med mindre datamængder).
- Understøtte samarbejde på tværs af institutioner og datasiloer.
- Understøtte internationalt samarbejde med fokus på FAIR og Open Data principper.”

I DI2025 benchmarkes situationen i Danmark mod tilsvarende i hhv. Finland og Holland, og der konkluderes på side 40: “Benchmark-landene har en lang række nationale datalagringstiltag, der for især Finlands og Hollands vedkommende har fokus på at dække behovene i hele datas

livscyklus. Der er ikke lignende tiltag i Danmark, hvor det primært er overladt til universiteterne at sikre forskernes adgang til sikre lagerfaciliteter".

Alle universiteter tilbyder lokale storage-løsninger. Arbejdsgruppen vurderer at der af hensyn til forskergruppers sammensætning på tværs af institutionerne og i øvrigt forskermobilitet, vil være behov for funktioner der muliggør sikker deling og fælles brug på tværs af lokale løsninger (det gule felt i figur 4). Nogle af de eksisterende lokale løsninger kunne desuden aspirere til at blive videreudviklet som led i en national løsning. Disse er gennemgået i bilag 4.

Gruppen har identificeret langtidsbevaring er en national opgave. Langtidsbevaring er såvel en teknisk som en organisatorisk opgave. Vi vurderer at universiteterne skal have en aktiv rolle i denne aktivitet, og at løsninger kan leveres både nationalt og internationalt.

5. Forslag til teknisk infrastruktur

Gennemgangen indtil nu har ført til:

- At der identificeres en række tjenester, som udbydes nationalt. Disse kan være danske eller internationale.
- Der etableres en samlet national datamanagement-arkitektur, som bygges op som en føderation af distribuerede tjenester.
- Alle tjenester udbydes "as a service" på forskellige niveauer fra infrastruktur (Infrastructure as a Service, IaaS) til brugervendte applikationer (Software as a Service, SaaS).
- Lokale, nationale og internationale datamanagement-tjenester bindes sammen af både nationale og internationale infrastrukturkomponenter som WAYF/eduGAIN, standardiserede PID'er (DOI, ORCID m.fl.) og lignende, som leveres via standardiserede protokoller og grænseflader.
- Der skal kunne implementeres hele proces-flows.
- Der skal være en god integration til HPC og/eller computer-ressourcer.

Basis for at kunne levere tjenester lokalt, såvel som nationalt og internationalt, er en velfungerende teknisk infrastruktur. I datamanagementsammenhæng i særdeleshed frit tilgængelige og søgbare lagringsfaciliteter. Det er et nationalt anliggende, at der tilvejebringes den nødvendige lagringskapacitet, og at den overliggende logiske struktur til håndtering af data etableres.

Storage kan have mange karakteristika til forskellige formål: volumen, antal filer, opdaterings- og læsningsfrekvens, båndbredde, særlige krav til sikring, herunder adgang til data, sikkerhedskopiering, og eventuelle bevaringsaspekter. I et nationalt perspektiv er der derfor behov for at kunne tilbyde en bred vifte af datalagringsfaciliteter, baseret på forskellige storage-teknologier, bundet sammen af en eller flere servicegrænseflader, med mulighed for søgning af data på tværs af hele strukturen og gerne med udgangspunkt i allerede eksisterende løsninger.

I et teknisk perspektiv indebærer det, at der bør etableres en struktur, der binder allerede eksisterende og fremtidige datalagringsfaciliteter sammen med en tværgående struktur, der

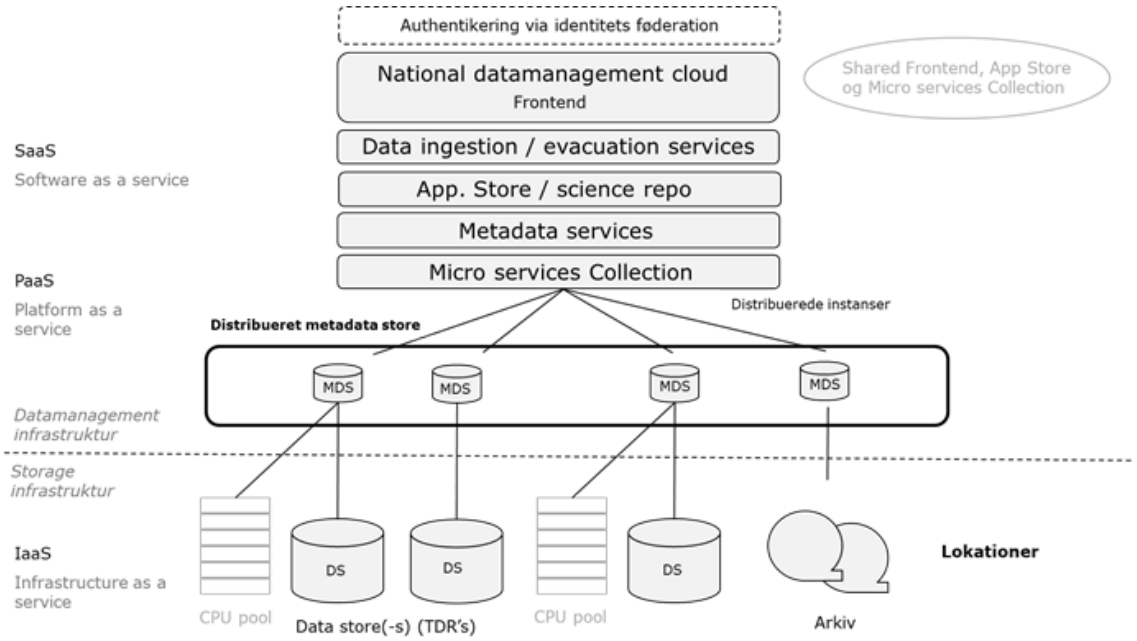
muliggør struktureret lagring og søgning, og med en overbygning, der består af service grænseflader og dataservices.

Landskabet, der er gengivet i figur 7, er baseret på en struktur, der er *distribueret* og *fødereret*.

Strukturen er distribueret over forskellige typer af datalagringsfaciliteter, der igen er baseret på forskellige teknologier.

Infrastruktur og services tænkes tilvejebragt af *flere* infrastruktur- og platformsoperatører. Disse operatører, der leverer services til den nationale back office-funktion, vil kunne være universiteternes og bevaringsinstitutionernes computing centre og it-afdelinger, eksterne leverandører og/eller partnere i EOSC.

Den samlede nationale struktur understøtter alle faser af datas livscyklus, fra data creation (i form af f.eks. parallelle filsystemer tæt knyttet til HPC-installationer) til arkiv og langtidsopbevaring.



Figur 7 - Forslag til nationalt datamanagement- og storage-landskab.

Strukturen er distribueret over flere datastores eller Trusted Data Repositories (TDRs). Det er muligt at tage afsæt i allerede eksisterende infrastrukturer, og forskelligartede behov kan opfyldes af forskellige teknologiske løsninger. Strukturen kan skaleres både ved at tilføje nye datastores, men også ved at gøre eksisterende datastores større. Ikke længere relevante teknologier kan udfases, og nye teknologier indfases sømløst. Det enkelte Trusted Data Repository kan være en del af en fagspecifik infrastruktur såvel som en generel forskningsinfrastruktur.

Et eksempel på en fagspecifik infrastruktur kan være Computerome der er fokuseret på det bioinformatiske forskningsfelt. Computerome repræsenterer et Trusted Data Repository der er designet til håndtering af de mest følsomme data. De sideordnede Trusted Data Repositories

kan således have forskellig anvendelses fokus og sikkerhedsklassifikation. Under hensyntagen til at et TDR (f.eks Computerome) har en høj sikkerhedsklassifikation, skal metadata i den overliggende metadata store kunne pålægges restriktioner mht. deres søgbarhed, såfremt disse indeholder peronfølsom information. Metadata skal sikre, at de nødvendige adgangsbetingelser til det underliggende datasæt kan overholdes.

Strukturen er fødereret og baseret på en microservice-arkitektur. En struktur baseret på microservices er til forskel fra en monolitisk struktur kendetegnet ved bedre at kunne dække forskelligartede anvendelsesmæssige behov og være rettet imod forskelligartede underliggende teknologier.

En microservice er en komponent i arkitekturen der er uafhængig af teknologi og som kan opgraderes og udskiftes uafhængigt af andre komponenter i arkitekturen. Microservices er dermed små uafhængige enheder der tilsammen udbyder strukturens overordnede service lag: Metadata services, app. Store / science repo og Data ingestion / evacuation services.

Med udgangspunkt i en microservice baseret arkitektur vil specifikke anvendelsesmæssige behov, behovet for at anvende forskellige teknologier til at opbevare data, behov for skalering, kort udviklingstid og behov for løbende opgraderinger bedre kunne tilgodeses end i en monolitisk struktur. Risiko i forbindelse med etablering og udvikling vil minimeres og den løbende vedligeholdelse blive simplificeret.

Når det tidligere er nævnt at f.eks en høj sikkerhedsklassifikation omkring et specifikt TDR afstedkommer behov for at metadata skal kunne pålægge restriktioner bl.a mht. søgbarhed vil en sådan politik omkring metadata genereringen kunne implementeres via et samspil af microservices.

Muligheden for at implementere et lag af microservice baserede politikker der regulerer dannelse og anvendelse af metadata, er afgørende for at den samlede struktur udefra kan blive set som udtryk for Data Management as a Service.

Indholdet i App. Store /science repo består typisk af forskellige containers der kan afvikles centralt eller lokalt, og på den måde federeres på tværs af strukturen. Anvendelse af forskellige virtualiserings teknologier, hvoraf containere er én, vil desuden i højere grad muliggøre at fagspecifikke løsninger udbydes baseret på en generel teknologisk platform.

Metadata skabes lokalt og lagres lokalt, men indgår i en distribueret databasestruktur, der gør søgning i et samlet nationalt metadata store muligt, såvel som potentielt også eksponerer data imod EOSC.

Data ingestion/evacuation services har til formål at sikre, at især de forskellige HPC ressource pools kan fødereres. *Data ingestion services* skal sikre hurtig og effektiv kopiering af data til HPC ressource pools, mens *data evacuation services* sikrer migrering af resultater tilbage til det lokale datastore. *Data ingestion/evacuation services* vil ikke generelt være relevante. Det vil ikke være hensigtsmæssigt at flytte større datasæt på tværs af infrastrukturen.

Der skelnes mellem storage-infrastruktur og datamanagement-infrastruktur. Datamanagement-infrastrukturen skal optimeres til at håndtere validerede data og de efterfølgende trin i datas livscyklus, mens storage-infrastrukturen (udover de validerede data)

lokalt også skal understøtte data creation og validering. Altså let adgang til diskplads, hvor data kan genereres og valideres relativt frit og ustruktureret, før de gøres til del af et valideret annoteret datasæt, der bevares i datamanagement-infrastrukturen.

I forbindelse med etablering af strukturen vil der som tidligere nævnt kunne tages afsæt i eksisterende storage-løsninger. Endvidere vil udvikling af den *nationale data management cloud* front end og *data ingestion/evacuation services* kunne tage udgangspunkt i eksisterende eScience cloud-løsninger, som er udviklet ved nogle universiteter.

Projektering og implementering af strukturen skal initieres og koordineres af Back-office funktionen (*jf. beskrivelsen i afsnit 6.2*). Arbejdspakker kan sendes i udbud hos universiteterne og bevaringsinstitutionernes computing centre og it-afdelinger samt eksterne kommercielle leverandører.

I det internationale perspektiv er det en potentielt muligt at implementere *Metadata services, microservices og metadatastore(MDS)* som en national EUDAT B2SAFE iRods node, der abonnerer på lokale obejct stores udbudt af universiteternes compute centre. Andre teknologier kan dog være grundlag for en egenudviklet løsning. iRods er dog med reference til den Hollandske Data Management og Storage infrastruktur interessant fordi teknologien er fuldt i overensstemmelse med den ovenfor beskrevne struktur. iRods rummer udover at være baseret på en microservice arkitektur, også mulighed for at modellere workflows der repræsenterer de tidligere omtalte politikker.

Desuden vil lokale B2SAFE-iRODS-noder kunne indgå i strukturen som Trusted Data Repositories.

6. Forslag til organisering af arbejde med datamanagement

Den samlede nationale datamanagement-organisering leverer teknisk infrastruktur og tjenester, kompetencer og support til forskere i Danmark.

Det er centralt at tænke de eksisterende storage- og datamanagement-faciliteter ind i en ny organisering. Flere skal opdateres og styrkes, mens andre kan spille en mere koordinerende rolle på tværs af institutioner. Organisationen består for størstedelens vedkommende af funktioner og roller, som varetages decentralt på universiteter, forskningsinstitutioner, -infrastrukturer og kommercielle udbydere. Disse suppleres af funktioner, som varetages centralt i DeiC[6]. I sit grundprincip er der tale om en distribueret organisation, bygget op omkring fælles national styring og finansiering. Denne organisation udbyder og supporterer et netværk af fødererede tjenester, som tilsammen understøtter forskeres brug, produktion, publicering og bevaring af data på tværs af fagområder og institutioner.

Organisationen understøtter principielt eksisterende og kommende politikker på området, således universiteternes allerede etablerede datamanagement-politikker såvel som eventuelle kommende nationale politikker. Strategien for nationalt samarbejde påpeger manglen på en

[6] DeiC benyttes her og fremefter som synonym for organisationen for det fremtidige nationale samarbejde om digital forskningsinfrastruktur og knytter sig ikke som sådan til den eksisterende DeiC-organisation.

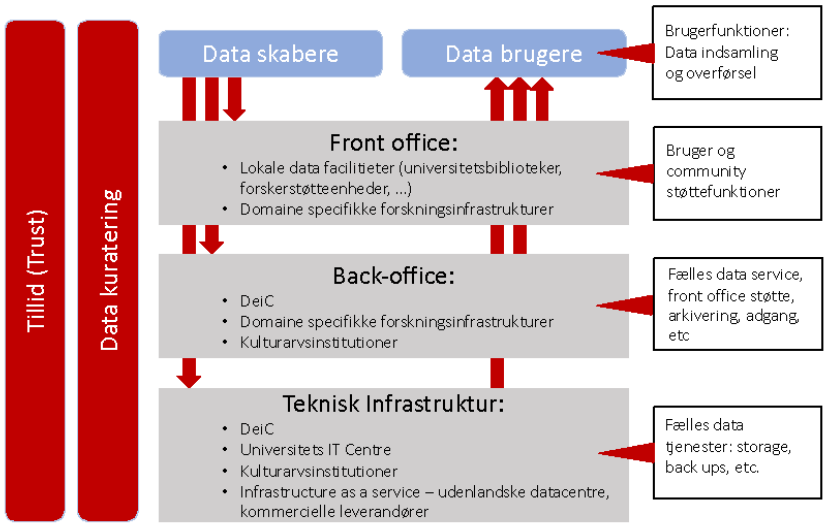
FAIR-politik i Danmark og den deraf følgende fragmentering af de lokale institutioners politikker og modenhed på datamanagement-området[7]. En eventuel kommende national politik vil være et naturligt udgangspunkt for fastlæggelsen af tjenesteporteføljen i det nationale samarbejde og vil desuden kunne fremme koordineringen af lokale politikker og dermed hjælpe til sammenhængen mellem tjenester på det lokale og det nationale niveau.

Som udgangspunkt peges derfor på,

- At al forskerrettet støtte leveres af personale på universiteterne – enten i dedikerede funktioner eller i relation til forskningsinfrastrukturer.
- Organiseringen af støttefunktioner bør gives høj prioritet ved de universiteter, hvor det ikke allerede er sket.
- At der identificeres en række tjenester, som udbydes nationalt. Disse kan være danske eller internationale.
- Der etableres en samlet national datamanagement-arkitektur, som bygges op som en føderation af distribuerede tjenester.
- Lokale, nationale og internationale datamanagement-tjenester bindes sammen af både nationale og internationale infrastrukturkomponenter som WAYF/eduGAIN, standardiserede PID'er (DOI, ORCID m.fl.) og lignende, som leveres via standardiserede protokoller og grænseflader.
- Ansvaret for levering af tjenester ligger i Back Office-organisationen (se nedenfor), mens den faktiske tekniske drift vil blive udbudt og varetaget af universiteterne, andre forskningsinfrastrukturer og eventuelt kommercielle udbydere. I forhold til langtidsbevaring skal kulturarvsinstitutionerne muligvis inddrages.
- Der etableres en forretningsafdeling som en del af Back Office, som har ansvaret for, at tjenester kommer i internationalt udbud.

Inspireret af løsningen i Holland, og konsistent med ovenstående analyse af livscyklus, foreslås organiseringen etableret omkring koncepterne *forkontor* og *bagkontor*:

[7] Strategi for nationalt samarbejde om digital forskningsinfrastruktur, 2018, s. 17.



Figur 8 En dansk version af den Hollandske model med Forkontor og bagkontor

Organisationen udgøres af tre elementer:

- Front Office
- Back Office
- Teknisk infrastruktur

Hertil kommer i en rådgivende rolle som kobles til Front Office:

- DM Forum

	Front Office	**Back Office**	**Teknisk infrastruktur**
Rolle	First level-support Markedsføring* Behovsopsamling* Kompetenceudvikling*	Levering af nationale DM-tjenester Specifikation og udbud Orkestrering og styring	Etablere og udbyde infrastruktur som service (IaaS)
Målgruppe	Forskere på institutionerne	Front Office Forskere	Back Office Tjenesteudbydere

Placering	Universiteter og forskningsinstitutioner	DeiC Nationale og internationale infrastrukturer	Udbydende institutioner, universiteter, fagspecifikke infrastrukturer
Finansiering	Lokal, den udbydende institution betaler selv	National	National

* Dele af Front Office-opgaverne varetages af DM Forum.

Den tekniske infrastruktur er allerede beskrevet i det foregående afsnit, så fokus herunder er på Front og Back Office.

6.1. Front Office

Front Office har ansvaret for den direkte tilgængelighed og support af tjenesterne til forskerne på de enkelte institutioner. Front Office er en decentral og distribueret funktion, som varetages på de enkelte forskningsinstitutioner af institutionernes egne dedikerede forskerstøtteenheder og af fagspecifikke forskningsinfrastrukturer med lokal tilstedeværelse. På flere forskningsinstitutioner kan denne funktion tænkes at slås sammen med GDPR-kontorerne.

Front Office varetager det niveau af forskerstøtte, som de lokale institutioner vælger at tilbyde, under hensyn til opfyldelsen af såvel lokale som eventuelle nationale politikker på området, f.eks. Code of Conduct. Front Office er i denne forstand ansvarlig for den praktiske anvendelse af tjenesterne til brug for lagring, deling, arkivering osv. af forskningsdata. Front Office sørger for at udbrede kendskabet til forhåndenværende datatjenester, således at disse bliver kendt og udnyttet af forskerne til rette tid og sted.

For mindre institutioner kan der muligvis opstå problemer med at dække bredden i Front Office-funktionerne. Dette kunne løses ved, at mindre institutioner kunne gå sammen om at etablere en fælles pulje af specialister, som så kan indsættes på forskellige institutioner efter behov. En model for dette kan findes i det amerikanske Data Curation Network[8].

Front Office har ansvaret for opsamling og videreformidling til Back Office af forskerbehov og for selv at identificere nye behov for datatjenester og kompetenceudvikling. Dele af denne funktion kan varetages gennem nøglepersoners deltagelse i DM Forum. Formidlingen af denne viden til Back Office skal ske gennem organiserede kanaler, det være sig som "tickets" i et udviklings-repository eller via brugergrupper eller lignende fora, som organiseres af Back Office. DM Forum kan spille en rolle i dette, og DM Forum kan stå for at gennemføre egentlige brugertilfredshedsundersøgelser efter nærmere aftale.

Front Office indgår i arbejdet med at planlægge og udføre kompetenceudvikling og erfaringsudveksling på tværs af institutionerne blandt forskere og forskerstøttepersonale. Det

[8] https://datacurationnetwork.org/

kan med fordel ske gennem DM Forum, som foreslås tildelt et budget til at fortsætte det eksisterende Train-the-Trainers-program.

De generelle forskerstøttefunktioner i Front Office finansieres fuldt ud af de institutioner og infrastrukturer, som stiller de pågældende personer og funktioner til rådighed. Front Office-funktionen referer alene til værtsinstitutionerne, men indgår i det samlede kredsløb både som aftager af tjenester og rådgivning og som leverandør af feedback, behov og krav til den samlede organisation.

6.1.1 Data Management Forum (DM Forum)

DM forum er et samarbejdsorgan, som binder aktørerne i Front Office sammen og som sikrer, at der sker en fælles kompetenceopbygning.

DM Forum har en funktion som videndelingsforum med henblik på at sikre den bedste og mest relevante støtte af forskere på datamanagement-området på tværs af institutionerne. DM Forum er i den forstand et af instrumenterne til at sikre et styrket nationalt samarbejde på datamanagement-området.

DM Forum har følgende strategiske opgaver i den foreslåede organisation:

- Bidrage til, at landskabet af lokale, nationale og internationale tjenester på datamanagement-området er kendt og forstået på institutionerne, således at de udbudte nationale tjenester udnyttes bedst muligt af forskerne,
- Bidrage til brugerinddragelse i forhold til udbudte tjenester, herunder sikre, at feedback fra brugerne og forslag til ændringer og nye tjenester løbende bliver indsamlet og tilflyder Back Office,
- Identificere behov for kompetenceudvikling blandt forskerstøttepersonale og forskere i datamanagement-relaterede emner og koordinere træningsaktiviteter til at opfylde behovene,
- Bidrage til fremtidige strategiske processer på datamanagement-området, f.eks. om FAIR data, og rådgive om implementeringen,
- Indgå med repræsentanter i Back Offices besluttende råd i forbindelse med specifikationer og udbud af tjenester,
- Indgå i arbejdet med en styrket international deltagelse og tilstedeværelse på datamanagement-området, bl.a. i en dansk RDA-node (DK-RDA).

DM Forums medlemmer udpeges af de deltagende universiteter, kulturarvsinstitutioner og KOR, og deltagelse i DM Forums møder og aktiviteter er som udgangspunkt på de deltagende institutioners egen regning.

DM Forum sekretariatsbetjenes af DeiC. DM-sekretariatet tildeles et beløb, som kan bevilges af DM Forums formand til iværksættelse af særlige aktiviteter, såsom kurser, konferencer, brugerundersøgelser, RDA-node eller andet, som skønnes værende af national betydning.

6.2. Back Office

Back Office har ansvar for leveringen af datatjenester til forskerne på de deltagende universiteter og forskningsinstitutioner. Det betyder, at Back Office påtager sig styringen af

den samlede portefølje af nationale tjenester på datamanagement-området, uanset hvem der konkret teknisk udbyder de enkelte tjenester.

Som indikeret i tabellen, kan Back Office have en sammensætning således at ud over DeiC også nationale og internationale infrastrukturer, som aspirerer til at udbydere deres tjenester til en bredere målgruppe, deltager. Vælger man denne bredere tilgang, er principper om governance meget vigtig.

For at kunne varetage styringen af den samlede portefølje er Back Office ansvarlig for *governance* for alle de tjenester, der udbydes for nationale midler. Det er Back Office, der specificerer og indkøber tjenesterne, herunder fastsætter de *tekniske krav*, som tjenesteudbyderne skal opfylde. De nærmere betingelser for udøvelse og organisering af denne governance-funktion fastlægges af DeiCs bestyrelse.

Etablering og drift af datatjenester bliver som udgangspunkt varetaget af aktører på universiteternes IT-afdelinger og e-science centre, bibliotekerne, fagspecifikke forskningsinfrastrukturer, bevaringsinstitutioner og kommercielle udbydere. I den forbindelse vil der skulle foretages udbud af etablering og drift af tjenester, der bliver besluttet iværksat. Back Office vil have brug for en funktion, man kunne kalde "indkøbskontor". Indkøbsfunktionen etableres samlet og forankres i DeiC. Principperne for udbud og de forretningsmæssige vilkår for tildeling af midler besluttes af DeiCs bestyrelse.

Tjenester, som udbydes gennem Back Office, skal være nationalt tilgængelige og udvikles konkret i henhold til de specifikationer, som Back Office publicerer (fleksibel skalerbarhed). Det betyder, at forretningsmodeller for de enkelte tjenester skal være transparente, med lige vilkår for alle nationale aftagere. Dette forhindrer ikke de medvirkende aktører i at udbyde egenudviklede tjenester nationalt efter andre ikke-godkendte forretningsmodeller og specifikationer. Disse vil blot ikke kvalificere til national medfinansiering.

Back Office er desuden ansvarlig for national koordinering af kompetenceudviklingen på datamanagement-området. På det strategiske plan skal dette formentlig ske i samarbejde med det eller de kompetencecentre, som er under opbygning på HPC-området. På det praktiske plan sker dette i tæt samarbejde med Front Office, som kan stå for planlægning og afvikling af konkrete arrangementer gennem DM Forum.

Det er Back Offices ansvar at koordinere kommunikationen både med Front Office, med decentrale og kommercielle tjenesteudbydere og med leverandørerne af den tekniske infrastruktur. Back Office varetager desuden dansk deltagelse i internationale strategiske/politiske partnerskaber på datamanagement-området, som eksempelvis RDA. Det anbefales, at de grupper, som arbejder aktivt med forskellige løsninger på datamanagement-området, involveres i de internationale projekter for at sikre, at Danmark bedst muligt udnytter de internationale resultater. Dette gælder eksisterende initiativer som EOSC projekter og europæiske infrastrukturer samt eventuelle kommende open source-communities og medlemsskaber, som måtte opstå.

Nationale og internationale Research Infrastructures (RI) kan indgå i udviklingssamarbejde på linje med nationale og internationale aktører. De skal dog kun indgå i projekter, som bidrager til en dansk løsning, og hvor man har udviklingsressourcer til at bidrage.

Back Office er det samlende led i organisationen, og de centrale funktioner finansieres via nationale midler. I det nuværende DeiC benyttes datamanagement-sekretariatet og formanden for DM Forum til en lang række konsulentlignende opgaver, rådgivning og egentligt projektarbejde. Ligesom national deltagelse i det internationale samarbejde på datamanagement-området i nogen grad varetages af DM-sekretariatet. Hvis sådanne opgaver fortsat skal løses af det nationale samarbejde, er det vigtigt, at Back Office har en tilstrækkelig bemandet og fagligt kompetent organisation til rådighed, enten centralt i DeiC eller udliciteret til en af institutionerne. Erfaringen fra den nuværende organisation peger i retning af, at der nok er brug for mere og ikke mindre af denne resurse.

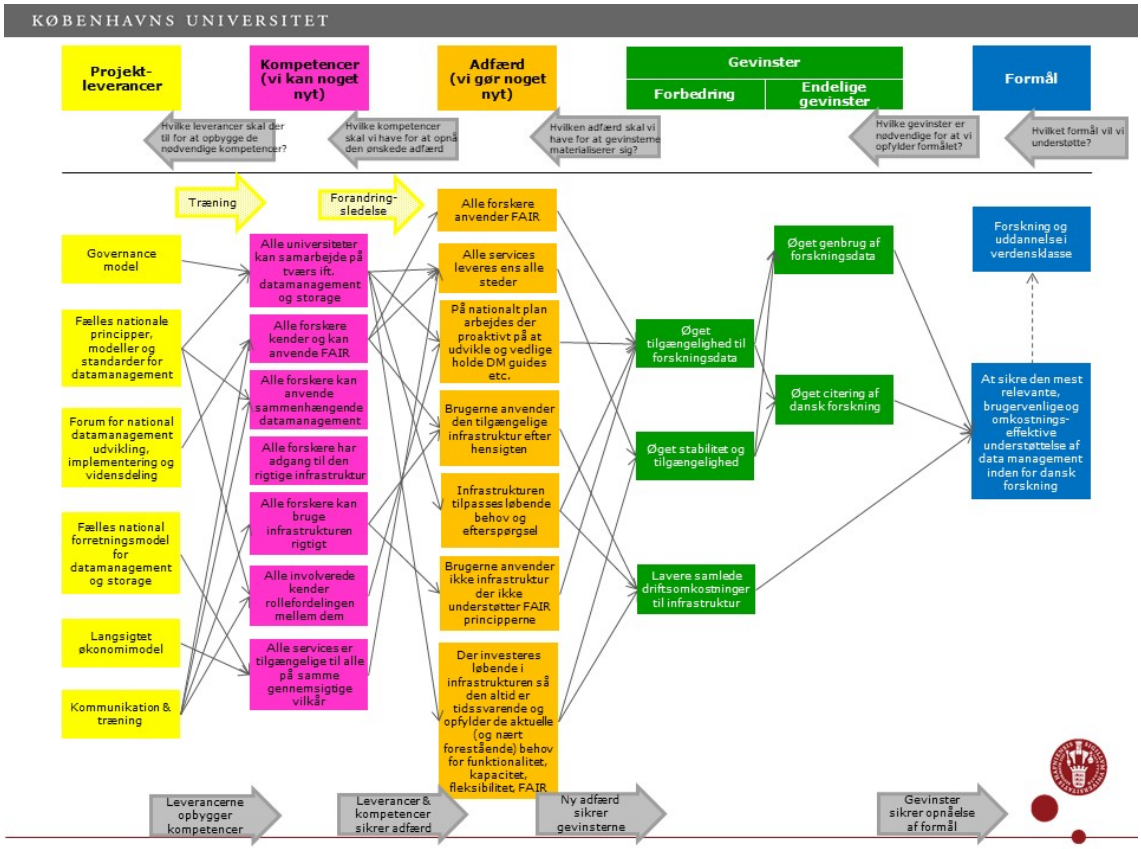
6.3. Indlemmelse/udbud af nye datatjenester

Nye tjenester og ny teknisk infrastruktur kan opstå på flere måder. En måde vil være, at Front Office opfanger opståede brugerbehov og kommunikerer disse videre til Back Office. En anden mulighed kan være, at en institution lokalt har opbygget en succesfuld tjeneste, som de mener, har potentiale til at kunne udbredes nationalt. Endelig kan der være situationer, hvor Back Office selv tager initiativ til at undersøge etableringen af en ny tjeneste, f.eks. som resultat af internationalt samarbejde eller efter ønske fra bestyrelsen om nye strategiske tiltag.

Der vil være flere faser i at afgøre, om en ny tjeneste skal opbygges og udbydes eller ej. Nogle af disse kan være af mere kommunikativ art, f.eks. workshops med mock-ups og lignende for at teste funktionaliteten hos forskere. En metode som Business Model Canvas[9] kombineret med mere formelle modeller for kapacitetsopbygning har været benyttet i DM Forum-regi til at undersøge hensigtsmæssigheden i, at en given institution udbyder en given tjeneste.

Det kan med fordel overvejes at indføre en egentlig formel model som led i beslutningsstøtten omkring indførelse og udbud af nye tjenester. Herunder ses eksempel på en sådan model fra Københavns Universitets IT-afdeling, som inkluderer et antal parametre i at opbygge en samlet vurdering af gevinsterne ved at indføre en ny tjeneste eller infrastrukturkomponent.

[9] Business Model Canvas er udviklet af firmaet Strategyzer og tilgængelig fra https://www.strategyzer.com/canvas/business-model-canvas. Denne model indgår som et element i arbejdet med Digital Curation Centre's Research Infrastructure Self Evaluation Framework (RISE), som DM Forum har afholdt træning i at bruge (http://www.dcc.ac.uk/resources/how-guides/RISE).



Figur 9 – Model fra Københavns Universitet til vurdering af gevinster ved ny tjeneste

6.4. Sammenhæng med EOSC og internationale partnerskaber

EOSC er på tidspunktet for denne rapport ikke færdigudviklet, og det kan være vanskeligt præcist at klarlægge, hvordan en dansk national infrastruktur skal pege ind imod EOSC. Her vil vi derfor pege på et begreb fra rapporten *Prompting an EOSC in Practice*: Minimum Viable Ecosystem, EOSC MVE[10]. Metaforen om et økosystem passer som udgangspunkt godt for den danske organisering og kan passende udstrækkes til at dække princippet for, hvorledes infrastruktur i Danmark kan tænkes at spille ind i en europæisk og global sammenhæng.

Som minimum sættes fokus på:

- Data og tjenester skal være **findable**, gennem brug af PID'er og eksponering af metadata via standardprotokoller (f.eks. OAI-PMH),
- Man skal benytte generiske og standardiserede AAI-tjenester, f.eks. WAYF/EduGAIN og muliggøre adgang for andre end den udbydende institution,

[10] Prompting an EOSC in practice: final report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud (EOSC), 2018, https://publications.europa.eu/en/publication-detail/-/publication/5253a1af-ee10-11e8-b690-01aa75ed71a1/language-en/format-PDF.

- Udbudte tjenester skal i rimeligt omfang understøtte FAIR-principperne, bl.a. gennem brug af standardiserede metadata og faglige vokabularer og brugen af maskinlæsbare licenser.

Herudover er det Back Offices opgave at have fokus på den videre udvikling af forretningsmodel og Rules of Participation for EOSC og sikre, at tjenester og data, der udbydes i det danske økosystem, er kompatible med disse regler og kan indgå, hvor det giver faglig og forretningsmæssig mening.

I sin reelle politiske implementering ser EOSC ud til at bygge på tre vigtige typer af aktører: de nationale infrastrukturudbydere (som DeiC), de store fagspecifikke infrastrukturer (typisk ESFRI'er) og de store eksisterende europæiske infrastrukturprojekter (som EUDAT, EGI og OpenAIRE m.fl.). Det er vigtigt at det nationale samarbejde på datamanagement-området spiller rollen som national infrastrukturudbyder i det europæiske landskab.

En anden måde at positionere det danske infrastrukturlandskab imod EOSC vil være gennem deltagelse i de europæiske projekter og partnerskaber, som bygger og indgår i EOSC. Her kan der bl.a. peges på EUDAT CDI[11], som udbyder en suite af infrastruktur og forskerrettede værktøjer, der omfatter hele datas livscyklus. DeiC er p.t. betalende medlem af EUDAT CDI og indgår i bestyrelsen, uden aktivt at udbyde nogen tjenester, som man egentlig som medlem er forpligtet til.

En aktiv deltagelse i EUDAT vil være en vej til EOSC, ikke kun med henblik på teknologi og tjenester, men også med henblik på medindflydelse gennem EUDATs plads i EOSCs styringsstruktur. En eventuel implementering af B2SAFE, som allerede foreslået for at få erfaringer med IRODS, vil således også kunne fungere som aktiv dansk deltagelse i EUDAT.

Der skal desuden peges på, at EU-Kommissionen anser Research Data Alliance (RDA)[12] som en væsentlig medspiller i dannelsen af EOSC. RDA er en medlemsorganisation med mange forskere og infrastrukturfolk som medlemmer. RDA har desuden "Organisational Members", og herunder er DeiC også medlem som organisation. Formålet med RDA er at fremme datadeling blandt forskere gennem udvikling af metoder og partnerskaber.

EU er medstifter af RDA og finansierer løbende programmer i RDA-regi, fordi de vurderes som en vigtig faglig kilde til udviklingen af EOSC. Således finansierer EU's RDA Europe 4.0-program den igangværende danske RDA-node (DK-RDA) som er forankret i DeiC, men hvor alle opgaverne er uddelegeret til DM Forums medlemmer. Det anbefales, at den nye DeiC-organisation afsætter midler til at fortsætte DK-RDA-nodens arbejde, når den nuværende EU-bevilling udløber med udgangen af maj 2020.

Generelt om det internationale samarbejde på datamanagement-området kan det siges, at Danmark deltager i en relativt begrænset mængde af disse, og at deltagelsen, med undtagelse af RDA-noden, primært har bestået i at deltage i møder. Denne form for international aktivitet

[11] https://www.eudat.eu/eudat-collaborative-data-infrastructure-cdi
[12] https://www.rd-alliance.org/

tjener til at opretholde et vist informationsniveau, men udbyttet er meget begrænset på det konkrete plan. Hvis Danmark vil have en egentlig gevinst ud af internationalt samarbejde, vurderes det, at der skal investeres i aktiv deltagelse, f.eks. i større europæiske samarbejder, i open source-communities og lignende.

Et andet internationalt samarbejde, som Danmark kunne have interesse i at deltage i gennem det kommende nationale samarbejde, er GO FAIR. GO FAIR er en bottom-up-organisation, som ønsker at gå i front omkring implementeringen af FAIR-principperne. Der har været et antal kontakter med GO FAIR og besøg i Danmark af Barend Mons og Erik Schultes, bl.a. med diskussion af konkrete projekter. Lidt ligesom RDA kan alle melde sig til de "Implementation Networks", som er kernen i aktiviteterne. Et engagement i GO FAIR kunne være et godt supplement til den danske RDA-node, DK-RDA.

6.5. Overgang/indfasning fra eksisterende nationalt landskab

Det ønskede og ovenfor beskrevne landskab af nationale datamanagement-tjenester og -infrastruktur vil ikke blive bygget op fra bunden. Udgangspunktet er, at der eksisterer både infrastruktur og tjenester lokalt og enkelte nationalt. En del af dette eksisterende udbud vurderes at have potentiale til at kunne videreudvikles og udbredes som nationale tjenester i en umiddelbar fremtid.

7. Bilag

Bilag 1 – Kommissorium

Bilag 2 – Interview med enkelte forskere ved Københavns Universitet

Bilag 3 – Opdatering og udvidelse af landeanalyse af Holland

Bilag 4 – Kort om SDU-cloud, CLAAUDIA og ERDA.