UiO : Universitetet i Oslo

Dataeksplosjonen – en stor utfordring, og en gedigen mulighet!

Rapport fra arbeidsgruppen «Lagring og deling av forskningsdata» ved Universitetet i Oslo (UiO) 11.05.2015

Phaistios Disk, et eksempel på tidlig datalagring.
The Archaeological Museum of Heraklion (cc by-sa)

Sammendrag

Forskning drives i stadig større grad av tilgangen på store datamengder. Datamengdene bare øker, og datadreven forskning har blitt en vesentlig mulighet i tilnærmet hele den vitenskapelige bredde. Beregningsvitenskap er etablert som det tredje paradigmet innen vitenskapelig metode, ved siden av teoretisk utleding og eksperimentelle studier, og nå blir dataintensiv vitenskap av mange sett på som et fjerde paradigme. Gordon Bell m.fl.(1) oppsummerer dette i artikkelen «Beyond the Data Deluge». Skal UiO være et ledende forskningsuniversitet må vi utnytte mulighetene i den datadrevne revolusjonen vi nå ser. UiOs forskere må dermed gis verktøyene og kompetansen de trenger for å være med i den fremste rekken innen datadreven forskning. «Riding the wave»(2), rapporten som virkelig satt betydningen av dataeksplosjonen på dagsorden i Europeisk sammenheng er nå 5(!) år gammel.

Som offentlig forskningsinstitusjon må UiO ha en tydelig politikk og effektiv infrastruktur for forvaltning (lagring, arkivering, deling, kuratering og gjenfinning) av forskningsdata som setter krav til alle aktører og som samtidig gjør det, ikke bare mulig, men enkelt å etterleve disse kravene. Institusjonen må stille tilveie tjenester, inklusive infrastruktur, kompetanse og kurs, mens forskerne har ansvar for håndtering av egne forskningsdata. Politikken og infrastrukturen må ta hensyn til forskningens dynamiske karakter og kunne håndtere endringer i forskernes behov i takt med teknologiutviklingen. UiO må bidra til at institusjonens forskere evner å utnytte mulighetene i dataeksplosjonen, og dermed bidra til løsninger som tillater raske og effektive søk og etterfølgende utnyttelse av forskningsdata, hvor de nå enn er produsert og arkivert.

Den foreliggende rapporten gir en overordnet beskrivelse av nåsituasjonen nasjonalt og ved UiO, og gir konkrete forslag til det videre arbeidet. Det er fire hovedtanker som må ligge som et fundament for alt videre arbeid:

i. Vi må se hele datahåndteringssyklusen i sammenheng; fra forskningsdata genereres til de gjenfinnes og gjenbrukes av andre.

ii. Vi må tenke globalt og ikke nasjonalt. Løsningene vi lager må tilfredsstille behovene til forskere i andre land.

iii. Vi kan ikke løse alle problemstillinger ved UiO, men vi må sørge for at problemstillingene løses gjennom samhandling med andre nasjonale aktører.

iv. Vi må alltid ha som ledetråd å utarbeide systemer som gjør at den enkelte forsker ser større fordeler enn ulemper ved å arkivere og dele egne forskningsdata, og utnytte andres forskningsdata. Et system tuftet på forskrifter og mer revisjonsmessig oppfølging vil neppe fungere. Vi siterer «Riding the wave»:

«Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.» (2)

Arbeidsgruppen foreslår konkret:

i. Klare retningslinjer for datahåndtering ved UiO (vedlegg til rapporten).

ii. En pilot som skal sørge for etableringen av et program for kompetanseutvikling og gode forskningsstøttetjenester.

iii. En klar arbeidsdeling (roller, ansvar og myndighet) institusjonelt, nasjonalt og internasjonalt i forhold til behovene for teknisk infrastruktur, og implisitt utviklingen av et tilbud for mellomlagring og deling av forskningsdata med metadata-beskrivelser ved UiO.

iv. At UiO medvirker til at en del sentrale problemstillinger som krever nasjonal samhandling reises og løses.

I tillegg poengterer arbeidsgruppen at det er behov for en større bevissthet rundt IT-utfordringene vi står overfor i hele organisasjonen og viser til rapportene for IT i forskning og utdanning ved UiO. Enheter bør i langt sterkere grad ha en bevisst strategi for IT i utdanning og forskning og den må ikke kun være relatert til den tekniske eInfrastrukturen. Mulighetsrommet for institusjonen og forskerne må belyses og ivaretas. Dette krever videre utvikling av IT-støttefunksjonene for undervisning og forskning. UiO bør videre vurdere å utvikle et sterkere studietilbud som sikrer kunnskapssamfunnet kandidater med spesialkompetanse innen håndtering og bruk av forskningsdata (data scientists) i tråd med anbefalingene i «Riding the wave»(2).

Et fungerende system for håndtering av forskningsdata vil kreve bred samhandling både internt ved institusjonen og med eksterne aktører. Samtidig trengs en godt forankret institusjonell politikk som er harmonisert med offentligheten og finansielle kilder. Utvikling og leveranse av tjenester, etablering og drift av infrastruktur, formidling og kompetanse-byggende tiltak, bør således i stor grad sentraliseres slik at et helhetlig og institusjonelt tilbud følger av politikken. Rådgiving- og støttefunksjoner bør i størst mulig grad være forskernære. Det er viktig at politikken samt de sentrale tjenestene harmoniseres med og utnytter at det innenfor flere fagfelt og prosjekter er etablert eksterne strukturer på internasjonalt eller nasjonalt nivå, som ivaretar forvaltning av forskningsdata etter de beste internasjonale standarder.

Ansvarsforhold må klargjøres. Vi mener at *ett* organ må ha det overordna ansvaret og samtidig være operativt sterkt nok til å sikre fremdrift totalt sett. Vi tror *eInfrastrukturutvalget, foreslått i rapporten «IT i forskning»*(3) bør ha dette samordnende ansvaret og være drivkraften for det videre arbeidet.

INNHOLD

SAMMENDRAG ***1***

1. EN GEDIGEN MULIGHET ***5***

2. DATALAGRING, ARKIVERING, DELING OG KURATERING - EN KORT INNFØRING ***6***

3. NASJONALE RETNINGSLINJER BASERT PÅ INTERNASJONALE FORPLIKTELSER ***9***

4. NÅSITUASJONEN VED UIO ***11***

5. DATALAGRING, ARKIVERING, DELING OG KURATERING – SKISSE TIL EN LØSNING ***13***
Kompetanseutvikling og støttefunksjoner *13*
eInfrastruktur og grensesnittet mellom bruker og system *14*

6. OPERASJONALISERING OG VEIEN VIDERE ***17***

7. UTFORDRINGER VI IKKE KAN LØSE ALENE ***20***

APPENDIX 1. ARBEIDSGRUPPENS MANDAT, MEDLEMMER OG ARBEID ***24***

APPENDIX 2. FORSLAG TIL P*OLITIKK OG RETNINGSLINJER:** ***25
Politikk *25*
UiOs retningslinjer for arkivering, tilgjengeliggjøring og deling av forskningsdata *26*

1. En gedigen mulighet

Begrepet dataeksplosjonen refererer til de enorme mengdene digitale data vi genererer globalt, enten direkte i forskningsøyemed eller som følge av våre digitale liv. Dataeksplosjonen er massiv, og kanskje tar vi ikke tilstrekkelig inn over oss hvor gjennomgripende den er; hvordan den er i ferd med å endre forskningen, men også samfunnet totalt sett. På én dag kan en moderne DNA-sekvenseringsmaskin lese mange milliarder deler av den menneskelige genetiske koden. I løpet av ett år genererer en slik maskin flere terrabytes data (trillioner av data enheter). Det er ikke enkelt å forholde seg til slike tall, men vi får en idé dersom vi sier at den årlige produksjonen fra en slik maskin tilsvarer informasjonen vi i dag finner i 20 biblioteker av størrelsen til US Library of Congress (2). Ett enkelt spesialisert instrument, i ett vitenskapelig underfelt og i løpet av ett år. Og dette for en instrument-type som ikke på noen måte produserer spesielt mye forskningsdata. Ved "Swedish Solar Telescope" på La Palma henter for eksempel Institutt for teoretisk astrofysikk inn 2,5 TB data pr. dag under gode observasjonsforhold. Se nå for deg det «store bildet» på tvers av alle fagfelt, over tiår og ikke minst globalt. Da får vi kanskje mer enn en idé – vi ser størrelsen på en datahåndteringsutfordring, men vi bør også se den gedigne muligheten som ligger i disse enorme mengdene med informasjon.

Eksemplet over henleder oppmerksomheten mot teknologi, naturvitenskap og medisin, men problemstillingen er ikke mindre viktig for humaniora og samfunnsvitenskap. «Store datamengder» har ulik tolkning innenfor ulike fagfelt, og begrepet «den lange halen» henviser til det enorme antallet mindre diversifiserte datasett som globalt produseres innen ulike fagfelt, innsamlinger av ulik art som enten representerer isolerte studier eller serier av studier over mange år. Et aktuelt eksempel er den pågående digitaliseringen av Nasjonalbibliotekets samlinger som åpner for nye viktige forskningsprosjekter innenfor en rekke humanistiske og samfunnsvitenskapelige disipliner. Datamengdene er også her både store og raskt voksende, og potensialet i utnyttelsen av det totale datasettet er enormt. Vi må ved UiO, nasjonalt og globalt dekke behovene til ulike fagfelt med ulike typer forskningsdata og med ulike perspektiver.

Dataeksplosjonen representerer et paradigmeskifte for forskningen(4), og det globale vitenskapelige samfunn må være proaktivt og legge forholdene til rette for utviklingen. Forskningen er en spiral hvor ny erkjennelse hele tiden komplementerer kjent viten, som så gir opphav til ny forskning, som igjen gir ny erkjennelse. Vil vi være med i forskningsfronten må vi forholde oss til et stadig høyere utviklingstempo og en stadig økende mengde med verdifulle forskningsdata. Dette fordrer dog at forskningssamfunnet kan finne og hente ut de forskningsdata de trenger raskt og effektivt, for så å sette informasjonen sammen, og starte forskningen basert på denne grunnmuren av eksisterende informasjon og kunnskap.

Vi står dermed overfor flere utfordringer. Hvordan utnytter vi den eksplosivt økende mengden globale forskningsdata? Kan vi, med riktige rammer, i større grad etterprøve andre forskeres resultater og tolkninger, og gjennom det fremtvinge en generell kvalitetshevning? Og ikke minst, dersom vi unngår unødig å duplisere allerede utførte studier; hvordan utnytter vi de frigitte ressursene? Mulighetsrommet er stort. Utviklingen understøtter f.eks. tverrfaglige, stor-skala studier knyttet til sentrale samfunnsutfordringer, som fattigdom, energi og global oppvarming. Konvergens mellom problemstillinger og disipliner muliggjøres.

2. Datalagring, arkivering, deling og kuratering - en kort innføring

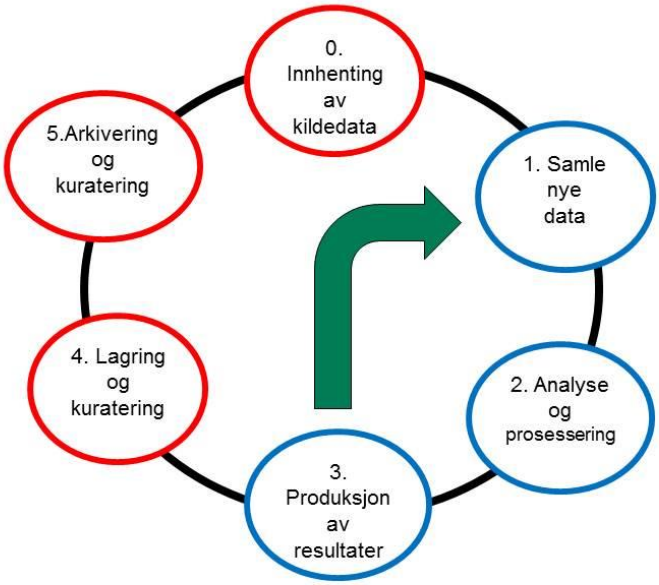
En kort innføring i sentrale begreper er påkrevd for å kunne vurdere tiltakene vi foreslår senere i rapporten. Med forskningsdata menes «*registeringer/nedtegnelser/ rapporteringer i form av tall, tekster, bilder og lyder som genereres eller oppstår underveis i forskningsprosjekter»*(5). Forskningsrådet velger å trekke et skille mellom kildedata og resultatdata. Kildedata er forskningsdata som hentes inn fra eksterne kilder og som eksisterer uavhengig av forskningsprosjektet de benyttes i. Resultatdata beskrives som «data som er generert gjennom forskning». Ved arkivering av resultatdata omdannes dermed disse til kildedata.

Vi har tatt utgangspunkt i et skjematisk forskningsdatakretsløp (figur 1), og for de fleste formål vil dette tjene som et godt bilde. I andre tilfeller vil dette enkle skjematiske bildet ikke beskrive den aktuelle problemstillingen like godt. Likevel tror vi dette er nyttig for å sette scenen.

I oppstart/planleggingsfasen for et forskningsprosjekt henter forskeren inn kildedata (markert med 0 i Figur 1). Hva har vært gjort tidligere og hva finnes av forskningsdata? Allerede her bør forskeren ta stilling til hvordan og når nye forskningsdata (resultatdata) skal deles og arkiveres. Flere finansiører krever allerede i dag en datahåndteringsplan beskrevet i prosjektsøknader. I denne tar forskeren stilling til:

i. Om forskningsdata kan deles eller om de faller inn under unntakene (se kapittel 3).
ii. På hvilke premisser deling skal foregå (inkludert *når*).
iii. I hvilke kanaler forskningsdataene skal publiseres.
iv. Hvor lang tid det er ønskelig/hensiktsmessig å lagre forskningsdataene.

Forskerne som har generert forskningsdataene vil ofte måtte ha en rimelig tidsperiode for å utnytte disse. Dette kalles førstebruksrett, og må reflekteres i datahåndteringsplanen.



Figur 1. Skjematisk forskningsdatakretsløp. Den foreliggende rapporten behandler de røde sirklene, mens de blå er ivaretatt i rapporten IT i forskning.

Ved dette er forskeren/forskerteamet klar for innsamling/produksjon av nye forskningsdata (1). Rådata som samles inn må ofte prosesseres før de kan analyseres (2). Analysene vil ofte avdekke at resultatene (3) på dette tidspunktet ikke gir svar på spørsmålene som er stilt, og det oppstår et behov for ny/endret datainnsamling (1). Det vil dermed etableres et indre forskningskretsløp for å frembringe resultater som gir svar på de opprinnelige spørsmålene. I dette forskningskretsløpet vil det være behov for lagring av forskningsdata som må kurateres fortløpende (4). Forskningsdata vil her typisk deles internt i forskergrupper, og gjerne med ekstern deltakelse. Lagringsløsninger for deling av forskningsdata og samskriving for hele forskerteam (inklusive eksterne) er dermed sterkt ønskelig. I den siste fasen i dette enkle skjematiske kretsløpet arkiveres forskningsdata for fremtiden (5). Forskningsdataene er nå klare for deling med omverdenen generelt, for eksempel via det som vi i denne rapporten kaller *gjenfinningstjenester*. Dette er verktøy/tjenester forskerne kan benytte for å gjenfinne og få tilgang til relevante globale forskningsdata.

Arkivering og deling innebærer at forskningsdata skal kunne utnyttes av andre, til nye formål. *Metadata*, den informasjonen som er nødvendig for at forskningsdataene skal kunne utnyttes av andre, må dermed følge med. Metadata vil kunne omfatte:

- Informasjon om dataformater, og om hvor, når og med hvilke instrumenter forskningsdataene ble samlet inn.
- Informasjon om f.eks. teknisk utstyr og programvare brukt, og eventuelle fagspesifikke standarder som er fulgt.
- Informasjon som gjør det mulig å finne igjen forskningsdataene, samt forstå om det eventuelt er begrensinger på gjenbruk av disse.

Samtidig må det knyttes unike identifikatorer, ofte kalt *digital object identifier* (DOI), til forskningsdataene og gjerne også til forskeren(e) som står for innsamlingen(e). Arkivering er også med på å sikre *reproduserbarheten* til studiet, at resultatene kan etterprøves. Dette er viktig i en tid hvor det i flere sammenhenger rapporteres at betydelige fraksjoner av studiene som publiseres globalt, ikke er reproduserbare.

Kuratering beskriver hele prosessen fra innsamling, til arkivering, vedlikehold og bevaring av forskningsdata for samtidig og fremtidig bruk. Et arkiv for forskningsdata uten kuratering blir som et bibliotek hvor bøker ikke beskrives og sorteres. Gjenfinning av forskningsdata blir da på sikt umulig. Kurateringsoppgaven innebærer videre å vurdere hva som skal bevares og ikke, samt sørge for at det bevarte materialet er lesbart for fremtiden. Datakuratorens/datarøkterens oppgave er å kvalitetssikre metadata og å forvalte forskningsdatasamlingene. Store deler av denne datarøktingen må skje i, eller i samarbeid med, forskningsmiljøene som genererer forskningsdataene.

Gjenfinningsløsninger som kan hjelpe forskere til å finne forskningsdata globalt, er en betydelig utfordring. Søkemotorer (for eksempel Google søk) er svært nyttige verktøy for å finne frem i ustrukturerte datasett som nettsider eller PDF-dokumenter. Det er langt vanskeligere å søke, kombinere og filtrere strukturert informasjon i databaser. Dette er en sentral datavitenskapelig utfordring, hvor UiO er verdensledende. Prof. Arild Waaler leder f.eks. et av de store Sentrene for Forskningsbasert Innovasjon, SIRIUS, med nettopp søk som den sentrale problemstillingen. Et viktig perspektiv er at det ikke finnes ett brukervennlig grensesnitt for alle søk/problemstillinger. Her er en viss grad av skreddersøm, og dermed er domenespesifikke løsninger påkrevd.

Det er også et betydelig regelverk knyttet til datahåndtering, og spesielt er det problemstillinger knyttet til personvern og opphavsrett. Prosjekter som omfattes av helseforskningsloven, personopplysningsloven, helseregisterloven eller bioteknologiloven må f.eks. ha forhåndsgodkjenning fra Datatilsynet eller Regional Etisk Komité (REK). Opphavsrettslige problemsstillinger knytter seg til hvem som eier

forskningsdata, innholdet i eksisterende databaser, og nye databaser som helt eller delvis er basert på eksisterende databaser. Det knytter seg f.eks. flere juridiske problemstillinger til bruken av helseregistre.

Gitt kompleksiteten, de mange aktørene og de store datamengdene, vil et fungerende og effektivt datahåndteringssystem kreve at både institusjonen og den enkelte forsker/forskergruppe har en klar og tydelig strategi for datahåndtering. Denne må baseres på klare retningslinjer, mens implementeringen av denne politikken vil kreve kompetansebygging, støttetjenester og en god infrastruktur med et godt brukergrensesnitt.

3. Nasjonale retningslinjer basert på internasjonale forpliktelser

Forskningsrådet presenterte i september 2014 sin politikk for tilgjengeliggjøring av forskningsdata(5). Denne kan sees i et lengre utviklingsperspektiv, og er oppsummert på Forskningsrådets nettsider:

- *I 2007 vedtok OECD "Principles and Guidelines for Access to Research Data from Public Funding". Norge har forpliktet seg til å følge opp disse retningslinjene.*
- *I de to siste forskningsmeldingene har Regjeringen understreket at den ønsker å legge til rette for økt tilgjengelighet av offentlig finansierte forskningsdata. Kunnskapsdepartementet ba Forskningsrådet etablere en politikk for åpen tilgang til offentlig finansierte forskningsdata.*
- *EU-kommisjonen anbefalte i 2012 medlemslandene å utvikle retningslinjer for åpen tilgang til forskningsdata.*
- *I Horisont 2020 er det utarbeidet en veiledning om Open Access som dekker både publikasjoner og data(10).*
- *I 2013 etablerte EU-kommisjonen Research Data Alliance (RDA) i samarbeid med NSF i USA og Australian National Data Service (ANDS). (6)*

Målsetningene for Forskningsrådets retningslinjer er:

- *Forbedret kvalitet i forskningen gjennom bedre mulighet for å bygge på tidligere arbeider og sammenstille data på nye måter*
- *Gjennomsiktighet i forskingsprosessen og bedre mulighet for etterprøvbarhet av vitenskapelige resultater*
- *Økt samarbeid og mindre duplisering av forskningsarbeid*
- *Økt innovasjon i næringsliv og offentlig sektor*
- *Effektivisering og bedre utnyttelse av offentlige midler (5)*

Det er også verdt å merke seg at Forskningsrådet ønsker å være en pådriver for bevaring og deling av forskningsdata. Dette innebærer blant annet at de skal:

- *implementere prosedyrer i søknadsbehandlingen som sikrer at relevante søknader inneholder planer for datahåndtering*
- *implementere prosedyrer i prosjektoppfølgingen som fører til at planene for datahåndtering blir fulgt av prosjektene som har fått midler*
- *videreføre praksisen med krav i sine kontrakter om at forskningsdata skal arkiveres på en forsvarlig måte i minimum 10 år(5)*

En vesentlig føring er knyttet til begrepet «åpen tilgang» (Open Access) til vitenskapelige forskningsdata. OECD-rapporten: «Principles and Guidelines for access to research data from public funding» er retningsgivende, og sier «open access to research data from public funding should be easy, timely, user friendly and preferably internet-based»(7).

Forskningsrådets policy følger også "åpen som standard"-prinsippet når det gjelder tilgang til forskningsdata. Men mens det i Horisont 2020 sin definisjon av åpen tilgang er et krav at tilgangen skal være gratis, har Forskningsrådet valgt å basere seg på at brukeren bør dekke de faktiske kostnadene knyttet til uthenting av forskningsdata. Dette er nærmere OECDs definisjon av åpen tilgang, som sier at tilgang skal gis til lavest mulig kostnad.

Det finnes flere utfordringer knyttet til åpen tilgjengeliggjøring av datasett. Forskningsrådets politikk konkretiserer noen gyldige årsaker til å begrense tilgjengeligheten:

> *Sikkerhetshensyn: I tilfeller hvor tilgjengeliggjøring av dataene kan true enkeltmenneskers eller nasjonal sikkerhet, **skal** datasettene ikke gjøres åpent tilgjengelig.*
>
> *Personsensitive data: I tilfeller hvor tilgjengeliggjøring av dataene er i strid med gjeldende regelverk for personvern, **skal** datasettene ikke gjøres åpent tilgjengelig.*
>
> *Andre juridiske forhold: I tilfeller hvor tilgjengeliggjøring av dataene strider med andre juridiske bestemmelser, **skal** datasettene ikke gjøres åpent tilgjengelige.*
>
> *Kommersielle forhold: Data som har kommersiell verdi og er generert i prosjekter der en bedrift har kontrakt med Forskningsrådet, **kan** unntas fra det generelle prinsippet om åpen tilgang. I disse tilfellene anbefales det at dataene gjøres tilgjengelig etter en periode, forslagsvis etter 3 eller 5 år.*
>
> *Andre forhold: I tilfeller hvor tilgjengeliggjøring av data får store økonomiske eller praktiske konsekvenser for dem som har generert/samlet inn dataene, **kan** datasettene unntas fra det generelle prinsippet om åpen tilgang dersom det argumenteres tilfredsstillende for dette.(5)*

4. Nåsituasjonen ved UiO

UiO er et stort breddeuniversitet og det er ingen overraskelse at det eksisterer mange forskjellige typer forskningsdata, med tilsvarende forskjellige behov for infrastruktur, tjenester og støtte. Arbeidsgruppen har gjennomført en kartlegging av nåsituasjonen på alle enheter, og den viser at det er stor variasjon selv innad ved enhetene. Samtidig finnes det mange forskere med samme type forskningsdata og lignende behov på tvers av enhetene. Dermed er det muligheter for utvikling av generiske løsninger. Kartleggingen viser at mange av UiOs forskere ikke har et bevisst forhold til arkivering og deling av forskningsdata. Til tross for at det ikke er en utbredt kultur for å dele forskningsdata åpent, er forskerne ved UiO generelt positive til å dele forskningsdata. Samtidig stiller forskerne krav om gode løsninger som er tilrettelagt for forskerne og som ivaretar forskerens interesser. Slike løsninger krever samtidig støttefunksjoner i form av oversiktlige nettressurser og ikke minst god veiledning. Det betyr at en av våre hovedutfordringer vil være kulturell; selv om det er enkelt å argumentere for den overordna tanken om deling, er en effektiv gjennomføring avhengig av at forskersamfunnet kollektivt ser seg tjent med ordningene som innføres.

Kartleggingen viser at det finnes mange eksempler på god praksis ved UiO. Det er flere forskergrupper, og til og med hele fagområder, som har gode systemer for datalagring og deling, som f.eks. innen språk, miljø og astronomi. Ikke overraskende har enheter med en gjennomtenkt IT-strategi og en godt utvikla IT-support generelt langt større bevissthet og fungerende løsninger enn andre.

Vi har også eksempler på egenutviklede generiske systemer som utmerker seg og får mye oppmerksomhet. USIT har for eksempel etablert Tjenester for Sensitive Data (TSD), en plattform for sikker håndtering av alle typer forskningsdata med lov- eller selvpålagte krav til informasjonssikkerhet. Felles for slike data er streng autentisering og autorisasjon, for å styre hvem som aksesserer hvilke data og hva som gjøres med dem. Det strenge tilgangsregimet omfatter ikke bare brukerne, men også operatørene. TSD har vært utfordrende å utvikle fordi sikkerhetsaspektet ofte er uforenlig med brukerønsker og etablerte driftsrutiner. Likevel har vi i dag en TSD-plattform som er fleksibel i den forstand at nye åpne arkiver kan inkluderes, og ved at løsningen muliggjør sikker informasjonsutveksling mellom systemer innenfor TSD og utenfor.

UiO har videre en avtale med Norsk Samfunnsfaglig Datatjeneste (NSD) som blant annet innebærer at NSD skal gjøre en forhåndsvurdering av prosjekter som skal behandle personopplysninger for å sikre at de gjennomføres i tråd med lovverket. NSD gjør også en oppfølging ved prosjektslutt for å påse at prosjektene avsluttes i tråd med det som er meldt inn og de kravene som stilles i lovverket. Denne

samarbeidsavtalen har imidlertid noen begrensninger for mulighetene til å arkivere og dele forskningsdata, og vi anbefaler derfor en klargjøring og evt. revisjon av avtalen. Et eksempel på en relevant problemstilling for mange UiO-forskere, er lagring av forskningsdata i form av lyd og video. Her påpeker Norsk Samfunns-vitenskapelig Datatjeneste (NSD) at «Dersom det er tatt lyd- eller bildeopptak (som kan identifisere enkeltpersoner) i forbindelse med prosjektet må disse også slettes/makuleres eller sladdes dersom datamaterialet skal være anonymt.»[1] Mange forskere ved UiO bruker lyd- og videodata som hovedkilde, eller som referanse for å gi mening til andre typer forskningsdata (f.eks. sensordata). Her er gjeldende praksis at det meste av opptakene må slettes etter prosjektets slutt, selv om materialet har/kan ha stor verdi for nye forskningsprosjekter. Dette er ikke optimalt, og det er generelt et betydelig behov for opplæring i relevant regelverk, støttetjenester for klargjøring og tolkning av regelverk, samt beste-praksis-løsninger for etterlevelse av regelverk. Det er i denne sammenheng også et behov for en klargjøring av ansvar og myndighet for ulike nasjonale instanser f.eks. relatert personvern.

Generelt er situasjonen ved UiO (og nasjonalt) ikke tilfredsstillende ut fra målbildet. Et fåtall av lagringssystemene som brukes ved UiO til lagring av egne forskningsdata, kan samtidig brukes til åpen eller tilgangsstyrt deling. Flertallet av lagringssystemene som brukes på UiO er interne systemer, hvor forskningsdata kun kan deles innenfor forsknings-gruppen internt på UiO. Ønsker man å dele med andre samarbeids-partnere brukes gjerne e-post eller eksterne skytjenester som DropBox. Denne formen for delings-løsning oppfyller ikke retningslinjene til Forskningsrådet eller EUs pilot for åpen tilgang til forskningsdata(5,8) (og er heller ikke ment å gjøre det).

Det er viktig å merke seg at vi i denne rapporten behandler lagring av forskningsdata. UiOs arkivsystem, ePhorte, fungerer utmerket for sitt bruk, men er ikke en løsning for forskningsdata som skal arkiveres med tanke på global tilgjengeliggjøring.

Med utgangspunkt i svarene fra kartleggingen finnes det i dag ingen generelle løsninger for å oppfylle Forskningsrådets og EUs politikk for det store flertallet av brukere ved UiO. Utfordringene er mange. Neste kapittel skisserer noen tiltak som vil bringe oss videre. Vi vektlegger her løsninger for det store flertallet av brukere, som ikke har fungerende nasjonale eller internasjonale løsninger pr i dag.

---

[1] *Epost K. U. Segadal NSD 09.03.15*

5. Datalagring, arkivering, deling og kuratering – skisse til en løsning

Nåsituasjonen ved UiO karakteriseres av stor variasjon både mellom og innenfor UiOs enheter. Utfordringene er mange, og løsningen vi skisserer kan oppsummeres i tre hovedpunkter:

i) Etablere tydelige retningslinjer for lagring og deling av forskningsdata (se kap. 6 og vedlegg) ved UiO.

ii) Understøtte kompetanseutvikling, og etablere støttetjenester for alle relevante målgrupper, og gjennom dette utvikle den kollektive systemforståelsen som muliggjør at hver enkelt forsker kan utvikle gode strategier for datahåndtering.

iii) Etablere en løsning for lagring av ulike typer forskningsdata som er lett tilgjengelig fra flere klientplattformer og som samtidig er fleksibel, sikker, langsiktig og oversiktlig. Ikke minst er det viktig å sikre en smidig overgang til arkivering i nasjonale og internasjonale arkiver. Bruker-perspektivet må stå i fokus, og det må utvikles effektive grensesnitt mot sluttbrukerne.

Evner vi dette, vil vi både kunne dele våre forskningsdata med andre, og i prinsippet kunne utnytte andres forskningsdata effektivt. For virkelig å kunne utnytte dataeksplosjonen bør det i tillegg utvikles effektive verktøy for gjenfinning av forskningsdata. Punkt ii) og iii) diskuteres videre under, mens mer konkrete forslag til tiltak ved UiO skisseres i kapittel 6. Utfordringer som krever samhandling nasjonalt beskrives i kapittel 7.

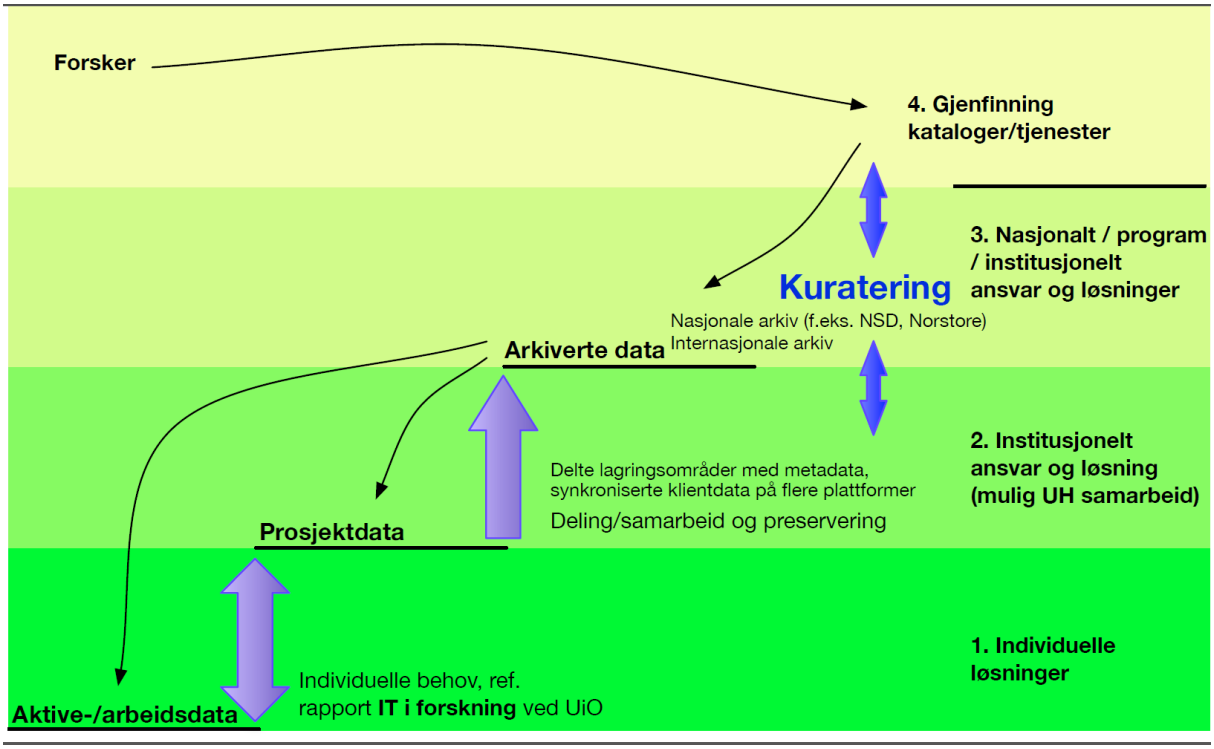
Kompetanseutvikling og støttefunksjoner

Generell kompetanseheving i organisasjonen og etablering av effektive støtte-funksjoner er påkrevd for at en gjennomtenkt UiO-politikk for forskningsdata skal kunne gjennomføres i praksis. Mange forskere mangler kunnskap og kompetanse om hvordan de skal arkivere og dele sine forskningsdata, hvordan de kan finne relevante forskningsdata i arkiver og om hvordan de skal gjenbruke andres forskningsdata på riktig måte i egen forskning. Det må tilbys kurs og opplæring. Samtidig må det etableres støttefunksjoner for en stadig mer dataintensiv forskningshverdag. Dermed trengs workshops, formelle opplæringsprogrammer, kurs i datasitering og systematisk datainnsamling, læreplaner som tar for seg utfordringene ved data-dreven forskning, etc. Konkrete tiltak er foreslått i kapittel 6. Her legges det stor vekt på strategisk utvikling av IT-støttetjenester som i økende grad involveres direkte i utdanning og forskning. Det er imperativt at dette skjer på undervisernes og forskernes premisser.

eInfrastruktur og grensesnittet mellom bruker og system

Det er nødvendig å definere og avgrense problemområdet før vi i mer detalj kan skissere en teknisk løsning som forsøker å ivareta flest mulig av våre forskeres opplevde behov for eInfrastruktur og grensesnittløsninger. Vi har valgt å definere fire ansvarsområder for forskningsdatainfrastruktur og tjenester. Denne inndelingen gjør det mulig å konkretisere tiltak og skissere en realistisk løsning på institusjonsnivå, selv om den ikke gir en stringent definisjon av hvilke forskningsdata som ligger innenfor hvert av ansvarsområdene. Inndelingen i fire ansvarsområder er illustrert i figur 2. Den sammenfaller i stor grad med datasyklusen i forskningsprosessen (se figur 1).

Forskningsprosessen (prosjektet) starter fra et forskningsspørsmål med utgangspunkt i tidligere kunnskap eller observasjoner (forskningsdata), og forskeren utformer en prosess som skal lede til ny viten. Prosessen involverer innhenting av forskningsdata, enten eksisterende forskningsdata eller nye, og videre manipulering og analyse av disse forskningsdataene. Ofte generer prosessen nye forskningsdata. Hva slags infrastruktur som er nødvendig i prosessen med bearbeiding av forskningsdata er individuelt og dynamisk. Ansvaret for dette anses derfor for å ligge hos forskeren selv eller det aktuelle forskningsprosjektet. Dette er illustrert som det nederste ansvarsområdet i figuren (nivå 1). UiO-rapporten «IT i forskning»(3) diskuterer eInfrastruktur generelt og dermed infrastruktur, tjenester og støttefunksjoner som anvendes i denne delen av datasyklusen.



Figur 2. Ansvar for løsninger og dynamikk i firedelt inndeling av «håndteringen av forskningsdata»

De to neste ansvarsområdene omfatter forskningsdata som ikke kun er tilgjengelige for en individuell forsker, men som er delte og har en grad av åpenhet. Hvor åpne og fritt tilgjengelige de delte forskningsdataene vil være, er avhengig av flere faktorer, slik som lovmessige, opphavsrettmessige eller kommersielle krav. Forskjellen mellom forskningsdata på Nivå 2 og 3 ligger i antatt eller etablert forskningsmessig verdi, og er ofte korrelert med forskningsdataenes grad av ”modenhet”.

Nivå 2 representerer en sentral løsning for lagring og deling av forskningsdata som er sikker og tilgjengelig på flere klientplattformer (brukersystemer). Ulike implementeringsløsninger finnes for-så-vidt for dette i dag, men det er flere betydelige utfordringer knyttet til dagens løsninger. Brukergrensesnitt er et stikkord. Uten et godt brukergrensesnitt blir oppgaven uovervinnelig for den enkelte forsker/forskergruppe. Dessuten er et avgjørende funksjonelt krav til løsningen at den ivaretar metadata og har hensiktsmessige tilgangsnivåer. Arbeidsgruppen mener at det er påkrevd å etablere en god infrastruktur for dette «nivå 2» ved UiO. Dette «nivået» ligger mellom de individuelle løsningene knyttet til aktiv bruk av forskningsdata (nivå 1) og etablerte nasjonale eller internasjonale systemer for arkivering av endelige forskningsdata (nivå 3). En slik ny infrastruktur har samtidig potensiale for å gi forskerne de nødvendige insentivene for registrering av metadata og dermed deling av forskningsdata.

Nivå 3 er assosiert med ferdig analyserte forskningsdata av verdi for forskersamfunnet, dvs. arkivverdige forskningsdata. Her ligger publiserte datasett som er tildelt en DOI/PID, som må være kvalitetssikret, metadataberiket og «låst» for endringer. Det finnes både nasjonale (Norstore, NSD, riksarkivet, museumsarkiver) og internasjonale programmer/tjenester og institusjoner, som ivaretar arkiveringen og tilgjengeliggjøringen av dataene (selv om det er utfordringer også her). Våre forskere kan velge de arkiveringsløsningene som er mest hensiktsmessig (i forhold til sitt fagområde og under eventuelle juridiske rammebetingelser). Samtidig må disse arkivene ha et effektivt og godt grensesnitt mot brukerne. Dagens systemer er langt fra optimale. Dette behandles videre i kapittel 7; «Utfordringer vi ikke kan løse alene».

Det er verdt å merke seg at det er en sterk sammenheng og avhengighet mellom de ulike ansvarsområdene i figuren. I forlengelsen fra individuelt eide forskningsdata er det behov for å kunne dele og sambruke forskningsdata (internt i et forskningsprosjekt, men også med eksterne nasjonale eller internasjonale samarbeidspartnere). Det er videre et behov for sikker og strukturert ”mellomlagring” for data som det ikke er avgjort om skal arkiveres for all fremtid, eller dersom det er andre forhold, som lovpålagte krav eller publikasjonskarantene som forhindrer full

åpenhet eller umiddelbar arkivering. Dette ivaretas av nivå 2 løsningen vi foreslår. Forskningsdata vil også kunne lagres på nivå 2 i påvente av kvalitetssikring, strukturering og annotering med hensiktsmessige metadata. Dette er prosesser som kan ta tid.

Gjenfinning og dermed gjenbruk av lagrede og arkiverte forskningsdata krever høy kvalitet på de assosierte metadataene. En sentral tjeneste for lagring og deling av forskningsdata ved UiO for forskningsprosjekt som ikke allerede har etablerte infrastrukturer (nivå 2), må stille krav om tilstrekkelig metadata slik at eventuell påfølgende arkivering (nivå 3) blir så enkel som mulig å gjennomføre. Her foreslår vi at det legges inn insentiver, for eksempel gratis lagring (ev. til kostpris for store prosjekt) når det er lagt inn nødvendig informasjon (metadata).

Det er viktig at den institusjonelle løsningen understøtter forskningssamarbeid og gir en tilstrekkelig sikker deling av forskningsdata. I dag deles forskningsdata med alt fra e-post til Dropbox, med tilhørende problemstillinger relatert til sporbarhet og sikkerhet. Den institusjonelle forskningslagringen bør derfor være like enkel å bruke som skyløsninger, være plattformuavhengig, og ha mulighet for sikker autentisering og tilgangsstyring. Samtidig må det være mulig å gi tilgang og dele forskningsdata med samarbeidspartnere også utenfor UiO. Autentiseringsmekanismer og andre strukturer som muliggjør dette finnes allerede i UoH-sektoren i Norge, og det er også internasjonale autentiseringsmekanismer som kan anvendes (for eksempel eduGAIN).

Vi har til nå behandlet systemer som hjelper forskeren med å gjøre egne forskningsdata tilgjengelig gjennom arkivering. Potensialet og insentivet for forskeren er på den annen side knyttet til muligheten som ligger i å utnytte andres forskningsdata. Dette krever gode gjenfinningssystemer for forskningsdata som høster metadata fra relevante arkiver og viser hvor forskningsdataene er tilgjengelige. Dette er ikke på plass i dag. Slike gjenfinningssystemer bør etableres nasjonalt, da flere små, spredte løsninger fort vil bli kaotisk, uoversiktlig og med varierende kvalitet. Vi vil understreke viktigheten av at slike gjenfinningssystemer kommer på plass, og foreslår et konkret tiltak for helsedata i kapittel 7.

6. Operasjonalisering og veien videre

Basert på nåsituasjonen, de gjeldende rammebetingelsene og ikke minst mulighetene som ligger i et fungerende, globalt forskningssamfunn som deler og utnytter forskningsdata på tvers av forskningsgrupper og landegrenser, foreslår vi følgende tiltak:

i) Etablering av klare retningslinjer for datahåndtering ved UiO. Et forslag til retningslinjer er vedlagt denne rapporten.

ii) Etablering av tilstrekkelig IT-kompetanse og -støtte i hele organisasjonen. Ved mange enheter er det behov for å utarbeide en IT-strategi, gjerne integrert i enhetens hovedstrategi, med en relatert tiltaksplan. Kompetansebehovet på «datarøkting» eller «dataarkivering» må i denne sammenhengen ivaretas. Behovet for en «datastrategi» likeså. Ved noen fakulteter, kan fakultetsnivået kanskje tenkes å dekke de underliggende enhetenes behov ved en sentralisert enhet. Arbeidsgruppens kartlegging viser imidlertid at IT-behovene i dag er en så sentral del av undervisning og forskning innen alle fagdisipliner, at behovene må synliggjøres ved budsjettering mer generelt enn i dag. I klartekst må IT-behovene synliggjøres ikke bare ved den årlige budsjetteringen av enheten, men også ved budsjettering av de enkelte forskningsprosjekter. Hva kreves i prosjektet, og hvordan dekkes de medfølgende kostnadene?

iii) Etablering av en pilot som skal sørge for etableringen av et program for kompetanseutvikling og gode forskningsstøttetjenester. Kartleggingen viser at UiO-forskere er generelt positive til å dele forskningsdata. Samtidig stiller de klare krav om gode tekniske løsninger som er tilrettelagt for forskerne og som ivaretar forskerens interesser. Utnyttelsen av disse tekniske løsningene krever samtidig god forskerstøtte i form av oversiktlige nettsider (se data.bristol.ac.uk for et godt eksempel på hvordan dette kan gjøres), opplæring og god veiledning. Støtten vil omfatte rådgivning til forskere for valg av infrastruktur, hvilke verktøy som kan brukes, gjenfinning og sitering av forskningsdata, hvilke metadatastandarder som er relevante, samt hvilke lagringssystemer som best ivaretar forskerens ønsker og samtidig oppfyller finansiørenes krav. Opplæringen må være basert på et lite antall korte opplæringsmoduler, som dekker brukerbehovene effektivt. Vi foreslår at det nedsettes en pilot som får i oppgave å utarbeide både en nettressurs og et helhetlig knippe av opplæringsmoduler, slik som:

- Datahåndtering for prosjektledere.
- Modul i oppstartstilbudet til masterstudenter i regi av UB. Her må bruk av eksisterende søkemotorer og sitering inkluderes.
- Tilsvarende eller beslektet modul for ph.d.-studenter basert på etterspørsel.
- Kurstilbud (internt og eksternt) for generell kompetanseheving blant aktuelle kursholdere på ulike nivå.

I lys av den arbeidsinnsatsen som legges ned av vår vitenskapelig ansatte, er det svært viktig at disse modulene er konsentrerte og brukerfokuserte. Vi må gi et tilbud som motiverer og ikke avskrekker. Nettbasert læring bør vurderes. Piloten må vurdere balansen mellom det generelle og generiske, og det mer spesialiserte som må foregå ved fakulteter eller ved enheter. Brukermedvirkning i piloten er påkrevd. Likeså løpende brukerevalueringer av tilbudet som utvikles.

Nettside som utarbeides må ha gode veiledere (maler, eksempler på god praksis, etc.), men må også være en støtte i forhold til juridiske aspekter som personvern og opphavsrett (se under)

iv) Etablering av en «nivå-2 løsning» for mellomlagring av forskningsdata med metadata. Denne må understøtte forskerne, og dermed ha et velutviklet bruker-grensesnitt. Mulighet for deling av forskningsdata og samskriving med kollegaer i inn og utland settes som et krav. Insentiver fungerer, og det er i det minste påkrevd at forskningsprosjekter får lagringsplass svært billig, eller til og med gratis ved å legge inn prosjektinformasjon og metadata. Det må nedsettes en arbeidsgruppe med representanter fra forskerne og IT-organisasjonen for å spesifisere tilbudet i større detalj. Kravspesifikasjonen vil danne utgangspunkt for en institusjonell satsing, f.eks. gjennom eInfrastrukturutvalget foreslått i rapporten «IT for forskning»(3).

v) UiO bør videre vurdere, i tråd med anbefalingene i EUs (2), å utvikle et sterkere studietilbud som sikrer kunnskapssamfunnet forskningsdatakyndige kandidater (data scientists); fagfolk som evner å utnytte mulighetene i et kommende velfungerende globalt system for deling av forskningsdata. UiO har et tilbud som kan fungere som en plattform for videre utvikling av et dedikert studieløp.

I tillegg er det flere utfordringer vi ikke kan løse alene. Disse utfordringene er beskrevet i mer detalj i kapittel 7.

Ansvarsforhold må klargjøres. Vi mener at *ett* organ må ha det overordna ansvar og samtidig være operativt sterkt nok til å sikre fremdrift totalt sett. Andre enheter kan og bør ha ansvar for undermengder av den totale problemstillingen. Vi tror det kommende eInfrastrukturutvalget bør ha det samordnende ansvaret og være drivkraften for det videre arbeidet som inkluderer:

i) Oppfølging av politikk og retningslinjer

ii) Implementering av institusjonell løsning (nivå 2)

iii) Kompetanseutvikling og gode forskningsstøttetjenester

7. Utfordringer vi ikke kan løse alene

Utfordringene vi ikke kan løse alene er av ulik karakter og viktighet. Viktigst er et avklart system for nasjonale arkiver, samt nasjonale løsninger for identifikatorer og metadata.

- **Nasjonale arkiver for forskningsdata.** Ansvar, arbeidsdeling, brukergrensesnitt og brukerfokus er stikkord. Hva kreves for neste generasjon Norstore? Kan NSD takle de kravene NFR stiller forskersamfunnet til arkivering av forskningsdata ved hjelp av NSD, eller er det et gap mellom kravene som stilles og tilbudet som gis? Et ryddig, helhetlig, nasjonalt system for arkivering av forskningsdata hvor internasjonale arkiver ikke kan benyttes er påkrevd.

- **Identifikatorer og metadata.** Forskningsdata skal arkiveres i systemer som kan deles med resten av verden, hvor internasjonale standarder for metadata og digitale identifikatorer som DOI for objekter og ORCID for forskeren følger datasettene. Arkivene må kurateres slik at gjenbruk sikres. Videre bør høsting av metadata gjennom åpne APIer tillates slik at dataene kan gjenfinnes av de ulike gjenfinningssystemene som måtte etableres. Dette vil også tillate etablering av systemer som genererer oversikter over bruk og gjenbruk av forskningsdataene, noe som er nødvendig for å etablere belønningssystemer for forskerne(se under). Det er en forutsetning at nasjonale utstedere av identifikatorer etableres så raskt som mulig, slik at NSD, NORSTORE og andre arkiver kan ta disse i bruk. Videre er det sterkt ønskelig at etablerte arkiv gir åpen tilgang til metadata, f.eks. gjennom APIer, og at metadataene følger relevante internasjonale standarder, f.eks. CERIF.

De to nest utfordringene er av en noe annen art fordi regelverket og de ulike aktørene eksisterer. Likevel mener vi at det er behov for en felles nasjonal tolkning/rettledning for undervisere og forskere. Dette er komplisert, og det virker i et SAK-perspektiv unødig at arbeidet utføres i parallell ved ulike institusjoner.

- **Personvern.** Forskningsaktivitet er underlagt streng regulering for å sikre personvern der dette er nødvendig. Det oppleves å være et behov for en avklaring av ansvar og myndighet for ulike instanser. Etter at helseforskningsloven trådte i kraft er det REK som gir forhåndsgodkjenning av forskningsprosjekter som faller inn under denne loven, institusjonelt internkontroll er pålagt å ha på plass for den enkelte institusjon. For forskning som ikke faller inn under helseforskingsloven, men som faller inn under

personvernloven kreves en forhåndsvurdering av NSD som personvernombud. NSD nåværende rolle og mandat i forbindelse med langtids arkivering av forskningsdata oppleves i dag som uklar og utdatert (mandatet er primært basert på tekstbaserte data). En avklaring av NSD / Personvernombudets rolle synes derfor nødvendig. Personsensitive forskningsdata kan ikke gjøres åpen tilgjengelig. Selv åpne anonymiserte data kan lede til personsensitive konklusjoner dersom slike datakilder sammenstilles på gal måte. Definisjonen av personsensitive data er imidlertid uklar og må utvides til å omfatte mulig uheldig sammenstilling av data (REK definerer videodata som personsensitive per se, mens NSD/Personvernombudet vurderer fra prosjekt til prosjekt). Visuelledata/videodata utfordrer gjeldende regelverk og det er behov for en bredere forståelse av sensitive data. Det kan være hensiktsmessig å være i dialog med Datatilsynet og NEM for å sikre en felles forståelse som følger den teknologiske utviklingen**.**

- **Opphavsrett.** Det må avklares hvordan forskningsmateriale som er underlagt opphavsrett kan benyttes, arkiveres og deles, f.eks. skjønnlitterære tekster, avisfotografier, TV-reportasjer og konsertopptak. Her er det også ofte snakk om flere nivåer av opphavsrett: komponister har rett til verk, musikere har rett til sin fremføring, plateforlag har rett til utgivelsen. Mange av disse gjelder også i lang tid etter at de involverte personene er døde, og det er forskjellige regler i ulike land. Hvordan skal vi håndtere dette på de ulike tilgangs- og delingsnivåene i systemene som utvikles?

De to siste utfordringene vi vil fremheve er av en annen karakter. Behovet for belønningssystemet som diskuteres til slutt er ikke like prekært, til dels utfordrende og kan med fordel utsettes noe i tid.

- **Gjenfinningsløsninger.** Det er gjennom gode gjenfinningssystemer for forskningsdata, deling og gjenbruk muliggjøres, og en strategi for tilgjengeliggjøring/utvikling av slike systemer er påkrevd. Generelt vil domenespesifikke løsninger være løsningen i uoverskuelig fremtid. Disse gjenfinningsløsningene kan like gjerne være internasjonale som nasjonale, men det er likevel fristende å foreslå et nasjonalt system for søk i strukturerte helsedata. Dette da Norge er i en unik posisjon ift helseregistere. UiO har allerede sterk kompetanse innen dette feltet, og et prosjekt i samarbeid med OUS/HSØ vil kunne styrke begge institusjoner.

- **Belønningssystem.** Det er ønskelig/nødvendig å belønne forskning og forskere som praktiserer god forskningsdata-håndtering og gir andre tilgang til sine forskningsdata. Mye forskningsdata tilgjengeliggjøres når studiene publiseres, men det er ønskelig med et system som premierer produksjon og tilgjengeliggjøring av gode forskningsdata. Her kan vi merke oss at det er i ferd med å etableres et antall Data tidsskrifter hvor det er mulig å publisere datasett og metadata. Slik publisering vil kunne gi forfatterne/dataeierne kreditering i form av publikasjonspoeng

Referanser:

1. Bell G, Hey T, Szalay A. Beyond the Data Deluge. Science. 2009 Mar 6;323(5919):1297–8.

2. High level expert group on scientific data. Riding the wave - How Europe can gain from the rising tide of scientific data - Final report of the High Leverl Expert Group on Scientific Data A submission to the European Commission [Internet]. European Union; 2010 [sitert 2015 Mar 23]. Tilgjengelig fra:: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

3. Arbeidsgruppe for IT i forskning ved Universitetet i Oslo. IT i forskning ved Universitetet i Oslo - Rapport fra arbeidsgruppe for IT i forskning ved Universitetet i Oslo (UiO). Oslo: Universitetet i Oslo; 2015 Jan 37 s.

4. Hey T, Tansley S, Tolle K, editors. The Fourth Paradigm: Data-Intensive Scientific Discovery. 1 edition. Redmond , Washington: Microsoft Research; 2009. 284 s.

5. Forskningsrådet. Tilgjengeliggjøring av forskningsdata - Policy for Norges forskningsråd. Norges forskningsråd; 2014.

6. Norges forskningsråd. Forskningsdata skal deles - Norges forskningsråd [Internet]. forskningsradet.no. [sitert 2015 May 6]. Tilgjengelig fra: http://www.forskningsradet.no/no/Nyheter/Forskningsdata_skal_deles/1254000298821?lang=no

7. Gurria A. OECD Principles and guidelines for access to research data from public funding [Internet]. OECD; 2007 [sitert 2015 Mar 9]. Tilgjengelig fra: http://www.oecd.org/dataoecd/9/61/38500813.pdf

8. European Commission. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 [Internet]. European Commission; [sitert 2015 Jun 5]. Tilgjengelig fra: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

9. CERIF: the Common European Research Information Format http://www.eurocris.org/cerif/main-features-cerif

Vedlegg:

Bakgrunnsmateriale for dokumentet,
Kartleggingen ved fakultetene utført av arbeidsgruppen:
https://www.usit.uio.no/om/organisasjon/uav/itf/saker/forskningsdata/kartlegging/

Til slutt vil vi rette en takk for nyttige og gode innspill til rapporten fra interne vitenskapelige ansatte og eksterne ansatte fra Kunnskapsdepartementet og Forskningsrådet.

Appendix 1. Arbeidsgruppens mandat, medlemmer og arbeid

Mandat:

- Kartlegge eksisterende tjenester og praksis, og forskernes behov for lagring og deling av forskningsdata ved UiO.
- Utvikle forslag til prinsipper og retningslinjer for lagring og deling/tilgjengeliggjøring av forskningsdata som ivaretar vitenskapelige ansattes rettigheter.
- Anbefale løsninger som samsvarer med finansiørenes økende krav til lagring og deling, og forskernes egne behov.

Medlemmer i arbeidsgruppen «Deling og lagring av forskningsdata»:
Leder Prodekan for forskning Svein Stølen, MN
Senioringeniør Torben Leifsen, MN
Førsteamanuensis Tor Endestad, SV
Senioringeniør Torgeir Christiansen, UV
Overingeniør Tore Miøen, OD
Instituttleder Alexander Refsum Jensenius, HF
Senioringeniør Espen Uleberg, KHM
Forskningssjef Fridtjof Mehlum, NHM
Rådgiver Katrine Ore, MED

Sekretariat:
Seksjonssjef Hans Eide, USIT(Seksjon for IT i forskning)
Hovedbibliotekar Live Kvale, UB(Realfagsbiblioteket)
Rådgiver Margaret Fotland, AF(Seksjon for forvaltning av forskning og utdanning)

Juridisk bistand:
Seniorrådgiver Einar Noreik, AF(Seksjon for forvaltning av forskning og utdanning)

*Appendix 2. Forslag til p*olitikk og retningslinjer:

Politikk

Universitetet i Oslo ønsker å forvalte forskningsdata[2] etter de mest krevende internasjonale standarder, og gjennom dette støtte utviklingen av et globalt forskningssamfunn hvor forskningsdata deles bredt. Dette skal bidra til:

- forbedret kvalitet i forskningen gjennom bedre mulighet til å bygge på tidligere arbeider og sammenstille forskningsdata på nye måter
- gjennomsiktighet i forskingsprosessen og bedre mulighet for etterprøvbarhet av vitenskapelige resultater
- økt samarbeid og mindre duplisering av forskningsarbeid
- økt innovasjon i næringsliv og offentlig sektor
- effektivisering og bedre utnyttelse av offentlige midler

UiO vil legge til rette for at ansatte og studenter enklest mulig skal kunne følge det til enhver tid gjeldende regelverk. Dette innebærer at UiO må ha klare retningslinjer for datahåndtering, gode kompetansehevende opplæringstilbud og understøttende nettsider, samt effektive støttetjenester for eInfrastruktur. Hele organisasjonen, samt relevante eksterne samarbeidspartnere/finansiører, må samarbeide for å implementere god praksis under de rammebetingelsene som er satt av lovverket og finansiører.

Forskningsdata skal være:

a. nøyaktige, fullstendige, ekte og pålitelige
b. identifiserbare, gjenfinnbare, og tilgjengelige
c. sikre og trygt lagret, enten sentralt ved egen institusjon eller i nasjonale/internasjonale arkiver i forhold til krav som stilles
d. vedlikeholdt i samsvar med juridiske og forskningsetiske forpliktelser
e. i stand til å bli tilgjengeliggjort for andre i tråd med relevante etiske prinsipper for deling forskningsdata.

[2]Med forskningsdata menes registreringer/nedtegnelser/rapporteringer i form av tall, tekst, bilde, lyd, video som genereres eller oppstår underveis i forskningsprosjekter.

Forskningsdata skal lagres/arkiveres så lenge de er av verdi for forskeren og et bredere forskningsmiljø, og så lenge som angitt av finansiøren, patentbestemmelser, lovgivning, embargokrav og andre myndighetskrav. Den minste lagringsperiode for forskningsdata og registreringer er tre (3) år etter publisering/offentliggjøring. I de fleste tilfeller vil forskningsdata beholdes lenger enn minstekravet på tre år. Generelt bør forskningsdata tilgjengeliggjøres på et tidligst mulig tidspunkt, men etter en førstebruksrettsperiode for forskerteamet selv.

Når forskning er støttet av en kontrakt/avtale/stipend som inneholder spesifikke bestemmelser om eierskap, oppbevaring av og tilgang til forskningsdata, vil bestemmelsene i denne avtalen/kontrakten ha forrang.

Hvis forskningsdata skal slettes eller tilintetgjøres, enten fordi den avtalte perioden for oppbevaring er utløpt eller på grunn av juridiske bestemmelser, bør dette gjøres i samsvar med alle juridiske og etiske prinsipper, samt finansiørenes og samarbeidspartnerenes krav, og med særlig hensyn til konfidensialitet og sikkerhet.

UiOs retningslinjer for arkivering, tilgjengeliggjøring og deling av forskningsdata

1. Forskningsdata skal lagres/arkiveres på en sikker måte
    a. Dataene skal lagres i sikre arkiver, enten sentralt ved egen institusjon eller i nasjonale/internasjonale arkiver
2. Forskningsdata skal gjøres tilgjengelig for videre bruk
    a. Forskningsdata skal gjøres tilgjengelige for alle relevante brukere, under like vilkår, så fremt det ikke er juridiske, etiske eller sikkerhetsmessige grunner til ikke å gjøre det (se under)
3. Forskningsdata bør gjøres tilgjengelig på et tidlig tidspunkt
    a. Dataene som ligger til grunn for vitenskapelige artikler bør gjøres tilgjengelig så tidlig som mulig, og aldri senere enn ved publiseringstidspunkt
    b. Andre data som kan være av interesse for annen forskning, bør gjøres tilgjengelig innen rimelig tid, og aldri senere enn tre år etter endt prosjekt
4. Forskningsdata skal utstyres med standardiserte metadata
    a. Metadataene skal gjøre andre i stand til å søke etter og ta i bruk dataene
    b. Metadataene skal følge internasjonale standarder
    c. Metadata skal inkludere en beskrivelse av datakvaliteten

    d. Dersom de data som utgjør grunnlaget for en publikasjon er utvalgt fra et større datasett, skal det store datasettet enten offentliggjøres eller beskrives
    e. Dersom observasjoner er fjernet fra datasett, skal prosedyren for eksklusjon beskrives og begrunnes
5. Forskningsdata bør utstyres med lisenser for tilgang, gjenbruk og videredistribusjon
    a. Lisensene bør være internasjonalt anerkjente
    b. Lisensene bør legge så få begrensninger som mulig på tilgang, gjenbruk og videredistribusjon av dataene
6. Forskningsdata bør gjøres fritt tilgjengelig, men reelle kostnader til distribusjon bør dekkes
    a. Metadata skal gjøres tilgjengelig uten kostnad og publiseres slik at de kan høstes maskinelt og brukes i søk etter forskningsdata
7. Forskningsdata bør utstyres med en langtidsplan
    a. Det bør utarbeides en plan for hvordan data som er vurdert til å ha verdi på lang sikt, skal forvaltes
    b. De vitenskapelige ansatte bør ha et bevisst forhold til hvordan forskningsdata som er vurdert til ikke å ha langsiktig verdi, skal forvaltes, eventuelt destrueres etter en viss tid

Forskerne har ansvar for å administrere forskningsdata i henhold til de prinsipper og krav som er gitt ovenfor. Dette betyr at de må utvikle og dokumentere klare prosedyrer for innsamling, lagring, bruk, gjenbruk, tilgang og oppbevaring eller ødeleggelse av forskningsdata i forbindelse med sin forskning. Dette skal inkludere ansvarsdeling i samarbeidsprosjekter med andre institusjoner. Informasjonen skal beskrives i en datahåndteringsplan. Juridiske rammebetingelser og krav fra finansiører skal ivaretas.

Hovedprinsipp for åpen tilgang

UiOs politikk følger "åpen som standard"-prinsippet når det gjelder tilgang til forskningsdata. UiO vil derfor bidra til at forskningsdata i utgangspunktet skal gjøres åpent tilgjengelig, men at det gjøres unntak for data som ikke kan eller bør gjøres tilgjengelig (se under). Tilgang skal gis til reell kostnad for tilgjengeliggjøringen. Unntakene omfatter:

Sikkerhetshensyn

- I tilfeller hvor tilgjengeliggjøring av dataene kan true enkeltmenneskers eller nasjonal sikkerhet, skal datasettene ikke gjøres åpent tilgjengelig.

Personsensitive data

- I tilfeller hvor tilgjengeliggjøring av dataene er i strid med gjeldende regelverk for personvern, skal datasettene ikke gjøres åpent tilgjengelig.

Andre juridiske forhold

- I tilfeller hvor tilgjengeliggjøring av dataene strider med andre juridiske bestemmelser, skal datasettene ikke gjøres åpent tilgjengelige.

Kommersielle forhold

- Data som har kommersiell verdi og er generert i prosjekter der en bedrift har kontrakt med UiO, kan unntas fra det generelle prinsippet om åpen tilgang. I disse tilfellene anbefales det at forskningsdataene gjøres tilgjengelig etter en gitt periode, forslagsvis etter tre eller fem år.

Andre forhold

- I tilfeller hvor tilgjengeliggjøring av data får store økonomiske eller praktiske konsekvenser for dem som har generert/samlet inn dataene, kan datasettene unntas fra det generelle prinsippet om åpen tilgang dersom det argumenteres tilfredsstillende for dette.