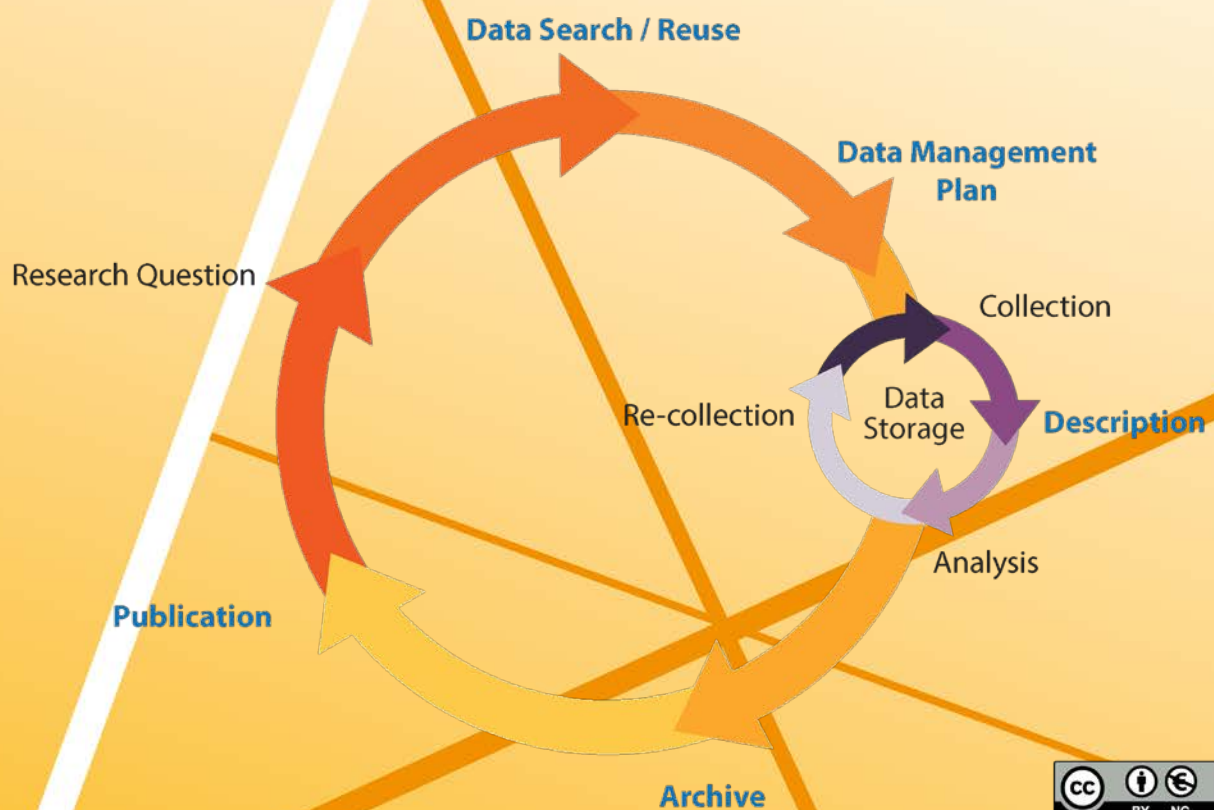


Pilotprosjekt infrastruktur for forskningsdata ved UiT

2016

Leif Longva*, Universitetsbibliotekar (Universitetsbiblioteket),
Espen Thijssen Amundsen, Overingeniør (Avdeling for IT),
Roy Dragseth, Seksjonssjef (Avdeling for IT),
Randi Østhus, Seniorrådgiver (Avdeling for forskning og utviklingsarbeid),
Stein Høydalsvik, Faggrupeleder (Universitetsbiblioteket)
* korresponderende forfatter – leif.longva@uit.no

The Research Data Management Lifecycle



Adapted original source:
The University of California, Santa Cruz,
Data Management LibGuide, Research Data Management Lifecycle, diagram,
viewed May 2, 2016 at <<http://guides.library.ucsc.edu/datamanagement>>

UiT The Arctic University of Norway – 2016

Septentrio Academic Publishing

<http://septentrio.uit.no/>

Septentrio Reports, number 4, 2016

ISSN: 2387-4597

DOI: <http://doi.org/10.7557/7.3784>

How to cite this report: <http://doi.org/10.7557/7.3784>

Licensee UiT The Arctic University of Norway

This Open Access report is licensed under a Creative Commons Attribution 4.0 International

License: <http://creativecommons.org/licenses/by/4.0/>



Innhold

Sammendrag.....	4
Om Pilotprosjekt infrastruktur for forskningsdata ved UiT.....	5
Resultatmål.....	5
Tidsramme	5
Organisering.....	5
Ressursramme.....	5
Metode.....	5
Prosjektdeltakere	5
Erfaringer fra prosjektet.....	6
Forskerdeltakelse - kommunikasjon og samarbeid	6
Organisering.....	7
Teknisk implementasjon	7
Støttetjenester for forskerne.....	9
Forslag til forvaltningsmodell for forskningsdata	11
Organisering.....	11
Roller og ansvar	11
Utvikling av UiTs policy for forvaltning av forskningsdata	12
Samspill med andre aktører i forvaltningen av Forskningsdata.....	12
Forslag til implementering i full skala – et åpent tilbud til alle ved UiT	14
Tekniske krav for skalerbarhet	14
Støttetjenester – Tjenesteutvikling – hva mangler?	15
Markedsføring.....	15
Tidsplan med milepæler.....	16
Vedlegg.....	17
Invitasjon til forskere om å delta i pilotprosjektet	17

Sammendrag

Pilotprosjekt infrastruktur for forskningsdata ved UiT ble vedtatt oppretta av IT-direktør og Bibliotekdirektør 25.9.2015 som et fellesprosjekt mellom ITA og UB. Avdeling for forskning og utviklingsarbeid ble invitert og har deltatt på status- og planleggingsmøter i prosjektet.

Hensikten med prosjektet var å få erfaring med tilrettelegging av teknisk infrastruktur for et institusjonelt arkiv for forskningsdata, kartlegge behov for kompetansebygging og støttetjenester for forskere som skal bruke en slik tjeneste, foreslå en forvaltningsmodell for forskningsdata ved institusjonen og beskrive hva som må til for å iverksette en slik tjeneste i full skala.

Arbeidet har bygget på «best practice» fra andre institusjoner (primært i England) og på en «adopt, adapt, develop»-modell, dvs. at en bruker ressurser som finnes allerede, tilpasser om nødvendig, og har nyutvikling som en siste utvei. Vi mener dette har vært en vesentlig grunn for resultatet. Rett nok betyr en slik tilnærming at ikke alt blir slik vi selv ville ha valgt å gjøre det, hvis vi hadde startet med blanke ark. Men hadde vi startet med blanke ark så hadde vi heller ikke vært i nærheten av målet på dette tidspunkt. Og er løsningen bra nok for Harvard så kan den sikkert også fungere ok for UiT Norges arktiske universitet.

Prosjektet kom ikke i mål innen utgangen av den tentative prosjektperioden som var satt til 31.3.2016, primært på grunn av merarbeid med integrasjon av Feide og problemer med migrering av eldre data til ny versjon av programvaren. Men prosjektet har likevel vært svært kompakt i tid og fungert godt som et samarbeidsprosjekt mellom de to enhetene, – og prosjektet har levert på alle resultatmålene som ble satt opp.

Det er levert og demonstrert en løsning for et institusjonelt arkiv for åpne forskningsdata. Løsningen er kvalitetssikret og har en definert forvaltningsmodell på plass slik at den kan iverksettes som en operativ tjeneste for institusjonen nå ved prosjektavslutning. En rekke opplæringstiltak for forskerne er også igangsatt og mer er under utvikling av UB. Planlagt lansering er satt til 1. september, men tjenesten er allerede operativ.

Videre er det beskrevet hvordan tjenesten kan skaleres opp til stort volum, hvilke støttetjenester som må være tilgjengelig for forskerne, og ikke minst forslag til hvordan institusjonen kan utvikle og iverksette en policy for forvaltning av forskningsdata.

Tilbakemeldingene fra forskerne som har deltatt med data til pilotprosjektet har gjennomgående vært svært positive og de har vist både interesse og en nysgjerrighet på hva dette betyr for den enkelte forsker.

Pilotprosjektet skulle bygge på UBs erfaring med åpne data fra arbeidet med dataarkivet TROLLing – The Tromsø Repository of Language and Linguistics, og har i liten grad sett på eller diskutert løsninger for data som ikke kan gjøres åpent tilgjengelig.

Om Pilotprosjekt infrastruktur for forskningsdata ved UiT

Pilotprosjektet ble opprettet på oppdrag fra IT-direktør og Bibliotekdirektør som et fellesprosjekt med avtalte deltakere fra ITA og UB. Avdeling for forskning og utviklingsarbeid ble invitert til å delta på status- og planleggingsmøtene i prosjektet, og har der deltatt med seniorrådgiver Randi Østhus.

Resultatmål

Prosjektet hadde følgende resultatmål:

- Forslag til tekniske løsninger og nødvendige støttetjenester for forskerne
- Demonstrere en generell arkivløsning basert på UBs erfaring med Dataverse-plattformen
- Forslag til implementering i full skala (hvordan komme i gang)
- Forslag til forvaltningsmodell for forskningsdata ved UiT

Tidsramme

Pilotprosjektet var planlagt gjennomført innenfor en kort periode (halvt år) for å sikre framdrift og nødvendig oppmerksomhet fra de involverte i prosjektet og fra de to enhetslederne. Tentativ prosjektperiode ble satt til 1.10.15-31.3.16.

Organisering

Pilotprosjektet har blitt gjennomført på oppdrag fra IT-direktør og Bibliotekdirektør som et fellesprosjekt med avtalte deltakere fra ITA og UB. Det daglige arbeidet har vært ledet av en koordinator fra hver av de to avdelingene. Koordinatorene har hatt flere status- og planleggingsmøter med de prosjektansvarlige på UB og ITA. AFU har også deltatt på flere av disse møtene.

Ressursramme

Prosjektforutsetningen var at ITA og UB begge stilte med nødvendige ressurser for gjennomføring av prosjektet, innenfor godkjent ambisjon og de rammer som ble avtalt. Eventuelle direktekostnader i pilotprosjektet, som f.eks. reiser eller anskaffelser, skulle dekkes av den enkelte enhet.

Metode

- Arbeidsformen skulle være faste, korte, felles arbeidsmøter for de involverte og individuelle arbeidsoppgaver med avtalte frister.
- Der det fantes grunnlag for det, så skulle arbeidet bygge på 'best practice' fra andre institusjoner, identifisert gjennom litteraturstudier, tjenestebeskrivelser og direkte erfaringsinnhenting.
- Løsningsforslag skulle ta utgangspunkt i en 'adopt, adapt, develop'-modell, dvs. at en bruker ressurser som finnes allerede, tilpasser om nødvendig, og har nyutvikling som en siste utvei.
- Samle inn og teste ut et avgrensa, men 'representativt' utvalg av forskningsdata fra forskere som er interessert i å delta i utviklinga av en arkivløsning.

Prosjektdeltakere

Seksjonssjef Roy Dragseth (prosjektansvarlig ITA), overingeniør Espen Thijssen Amundsen (koordinator ITA), senioringeniør Lars Slettjord (prosjektdeltaker ITA), faggruppelider Stein Høydalsvik (prosjektansvarlig UB), universitetsbibliotekar Leif Longva (koordinator UB), senioringeniør Obiajulu Odu (prosjektdeltaker UB), universitetsbibliotekar Philipp Conzett (prosjektdeltaker UB), førstebibliotekar Helene N. Andreassen (prosjektdeltaker UB), førstebibliotekar Lene Østvand (prosjektdeltaker UB) og førstebibliotekar Lars Figenschou (prosjektdeltaker UB).

Erfaringer fra prosjektet

Forskerdeltakelse - kommunikasjon og samarbeid

Det ble sendt ut en åpen invitasjon pr. e-post (vedlegg 1) til forskere ved de ulike fakultetene, fra «deres» kontaktperson (fagansvarlig) på UB. I tillegg ble samme e-post sendt direkte til forskere som tidligere hadde vist interesse for arkivering og publisering av forskningsdata.

Til sammen kom det inn 14 svar på henvendelsen, med følgende fakultetsfordeling: Helsefak (1), TMU (1), NT-fak (2), BFE-fak (2) og HSL-fak (8). Noen hadde spørsmål de ønska svar på, noen hadde kommentarer til initiativet, noen var nysgjerrige og ville gjerne delta, men hadde ikke data tilgjengelig for arkivering nå, – og noen svarte positivt og hadde data de kunne bidra med. Ut fra tilgjengelig tid i prosjektet, graden av tilrettelegging for arkivering av dataene, type data, samt tekniske avklaringer rundt dataene, så valgte vi å jobbe videre med følgende data i pilotprosjektet:

Datasettbeskrivelse	Forsker – organisasjon	Kommentar
Omfatter lingvistiske datasett og statistiske programmer/-koder for språkanalyser fra arkivet TROLLing - The Tromsø Repository of Language and Linguistics	TROLLing er et internasjonalt, tematisk dataarkiv med egen styringsgruppe som har ansvar for arkivets policy knyttet til vitenskapelig innhold og tematisk avgrensning. UB har ansvar for kvalitets- og relevanskontroll, kuratering av data og brukeradministrering.	Oppretta juni 2014 etter initiativ fra HSL, som arkiv for lingvistiske data og statistikkode. Arkivet er open access, noe som betyr at innholdet er åpent og fritt tilgjengelig for alle. Flyttes nå til UiT Open Research Data og inngår som en egen samling i den nye tjenesten.
Ornitologiske feltdata fra observasjoner og undersøkelser over en 30 års periode.	Rob Barrett, Professor, Seksjon for naturvitenskap, Tromsø museum	Data samla i et antall Excel-filer med et større antall ark i hver fil. Behov for noe kuratering, men oversiktlige og godt ordna data for arkivering.
Datafiler fra pilotering av implisitt assosiasjonstest (single category IAT – SC- IAT). Testen måler assosiasjonsstyrke mellom en kategori og evaluering av positiv eller negativ valens.	Gerit Pfuhl, Førsteamanuensis, Institutt for psykologi, Helsefak.	Har sendt oss data i form av iqdat-filer og en wordfil (beskrivelse).
Omfatter forskningsdata/ bakgrunnsdata for en publikasjon.	JoLynn Carroll, Professor II, CAGE – Centre for Arctic Gas Hydrate, Environment and Climate, Institutt for geologi, NT-fak	Datafiler i excelformat, pluss noen bilder/illustrasjoner. Mottatt readme-fil "containing information on how the data were interpreted and how they were derived (methods)."
Omfatter annoterte ordbøker, tekstkopis, programsript for tekstanalyse og andre typer språkteknologisk ressursar for ulike samiske språk og andre ur-	Giellatekno, Senter for samisk språkteknologi, Institutt for språk og kultur, HSL-fak.	Har sendt oss filer i XML-format og andre åpne format. Mangler noe metadata/dokumentasjon.

De øvrige data som kom inn til pilotprosjektet vil bli jobbet videre med etter avslutta prosjekt, som del av forvaltning av tjenesten.

Tilbakemeldingene fra forskerne har gjennomgående vært svært positiv og interessert, koblet med en nysgjerrighet på hva dette betyr for den enkelte forsker. Utfordringen med nye tjenester er alltid å nå ut med informasjon til de tjenesten er retta inn mot. Men i dette tilfellet hvor en ny tjeneste skal vokse fra et nullnivå, så kan det være en fordel for organisasjonen og støttetjenestene rundt dataarkivet at bruken i starten øker gradvis, slik at brukererfaringene kan bidra til å forbedre tjenesten.

Organisering

I oppdragsbeskrivelsen står det at dette skulle være et fellesprosjekt med deltakere fra ITA og UB og at det daglige arbeidet skulle ledes av en koordinator fra hver av de to avdelingene. Erfaringen er at gjennom arbeidet i prosjektet så har det blitt tydelig for deltakerne både fra UB og ITA at kvaliteten på tjenesten forutsetter engasjement og deltakelse fra begge enhetene, og at forståelsen av kompleksitet og behovet for støttetjenester har vokst betydelig i løpet av prosjektet. UB og ITA har tatt ansvar for ulike oppgaveområder, men har involvert den andre parten for å sikre god sammenheng med systemet totalt og hvordan dette kan forvaltes videre. Avdeling for forskning og utviklingsarbeid ble invitert til å delta på fellesmøtene mellom de to koordinatorerne og de to prosjektansvarlige ved UB og ITA.

Forvaltning av en moden og velfungerende tjeneste vil forutsette et tett og koordinert samarbeid mellom UB og ITA, dvs. mellom en teknisk infrastruktur som er tilpasset behovet og støttetjenester som opplæring, rådgivning og kuratering overfor forskerne.

Teknisk implementasjon

Ut fra de resultatmål som var satt for prosjektet så har ITA utført følgende teknisk tilrettelegging:

- Installert programvaren Dataverse
- Etablert sikkerhet rundt løsningen
- Implementert Feide pålogging
- Dokumentert installasjon
- Etablert løsning for backup av systemet

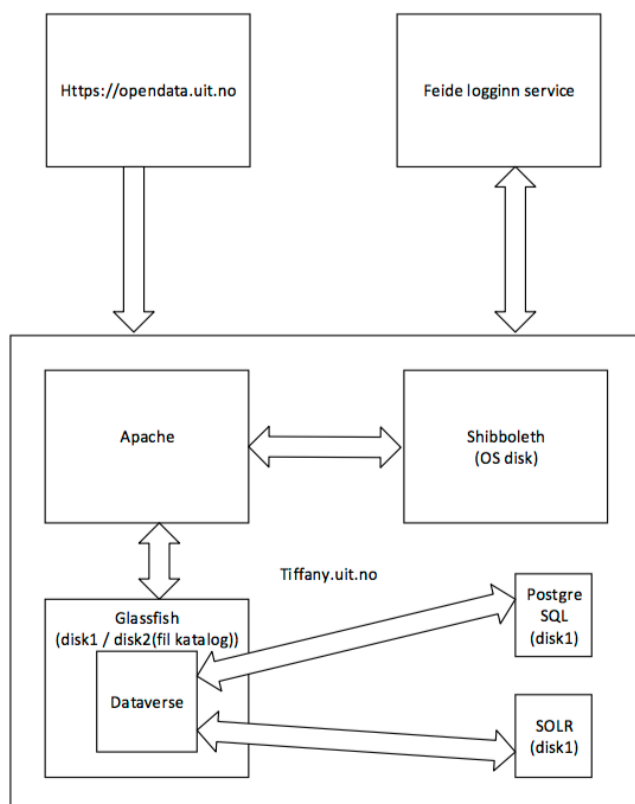
I tillegg er det sett litt på hvilke tekniske støtteverktøy som finnes eller kan tilrettelegges for forskere som har behov for å konvertere datafiler til formater som egner seg til langtidsbevaring. Men dette er noe som må løses utenfor prosjektfasen.

a) Dataverse på produksjonsplattform

Konfigurasjonen av Dataverse er basert på en antakelse om at bruken av UiT Open Research Data vil øke gradvis over tid, i samsvar med erfaringer fra UB sin tidligere installasjon av TROLLing). UiT har begrensa erfaring med denne tjenesten når det gjelder bruk og trafikk i stor skala. Den første produksjonsfasen må derfor brukes til bli kjent med hvordan systemet takler belastning (brukere), og erfare hvor mange som tar i bruk dette tjenestetilbudet. Erfaringsinnhenting må gjøres i samarbeid mellom ITA og UB. Slik som arkitekturen på systemet er satt opp fra starten av, så skal det lite arbeid til for å løse kapasitetsutfordringer ved maskinens ytelse og lagringsbehov, – i alle fall opp til et visst nivå.

Gjeldende systemarkitektur baserer seg på at alle deler av tjenesten kjører på en sentral maskin. Å bruke en maskin for alle komponenter har sine fordeler og ulemper. Fordelene i denne fasen av tjenesten er at

vi mener en slik løsning dekker godt antatt forventet behov ett til to år frem i tid. Ulempen er at tjenesten ikke kan være ute av drift i arbeidstiden, og er sårbare i en eventuell feilsituasjon.



Arkitektur av tjenesten

Som autentiseringsmetode (påloggingsmetode) er det valgt å bruke Feide tjenesten (<https://www.uninett.no/feideinnlogging>) som er levert av Uninett. Dette er en felles tjeneste for sikker identifisering i utdanningssektoren. Det er blitt bestemt at i første fase skal tjenesten kun tilby pålogging til ansatte ved UiT igjennom Feide (kan utvides til å tilby pålogging til andre utdanningsinstitusjoner hvis ønskelig). Ved å bruke Feide forenkler en brukerhåndteringen i systemet, og UiT-brukerne er allerede godt kjent med påloggingsmetoden fra andre systemer ved UiT.

Systemet har også en inngang for direkteregistrerte bruker som kommer fra andre institusjoner. I første omgang gjelder dette brukere av TROLLing, siden denne tjenesten er åpen for brukere fra hele verden. Men løsningen kan også være aktuell for andre grupper med medlemmer utenfor UiT.

ITA har utarbeidet en god og kvalitetssikret dokumentasjon på installasjonen. Dette er gjort ved å sette opp flere testinstanser av Dataverse, der gjeldende dokumentasjon har vært benyttet som oppskrift. Denne dokumentasjonen vil bli gjort tilgjengelig på et delt område som både UB og ITA har tilgang til.

Tjenesten inngår i vanlig driftsrutiner ved IT avdelingen. Dette innebærer at tjenesten blir overvåket med verktøy som ITA benytter. Vedlikehold som medfører nedetid skal varsles til brukerne god tid i forveien.

b) Migrering av TROLLing til produksjonsversjon

TROLLing er et internasjonalt dataarkiv for lingvistiske data og statistikkode som ble lansert 18. juni 2014, og er et samarbeid mellom HSL og UB. I rammen for pilotprosjektet ble det forutsatt at prosjektet skulle bygge på samme systemplattform som for TROLLing, – systemet Dataverse fra Harvard. På den måten kunne vi også oppnå en integrasjon mellom løsningen for TROLLing og den nye tjenesten UiT

Open Research Data, hvor TROLLing kan inngå som en egen samling (sub-Dataverse). Imidlertid så viste det seg svært vanskelig å få migrert TROLLing-data fra dagens løsning til den nye UiT-installasjonen på grunn av en stor systemoppgradering for den nye versjonen av Dataverse. En slik migrasjon var heller ikke godt nok testa av Harvard og har bidratt til noe av forsinkelsen for prosjektet. Men med god bistand fra Harvard-teamet så har vi nå en løsning som vi også vil gjøre tilgjengelig for andre (internasjonale) installasjoner av Dataverse. Den nye installasjonen av TROLLing har også gitt økt funksjonalitet og kvalitet til TROLLing-data, bl.a. så får alle datasett nå en unik DOI som referansenøkkel, og flere muligheter for å bli høsta av søketjenester.

c) Tilknytning til ID-tjeneste for forskningsdata

UB har inngått avtale med DataCite om rett til å tildele en unik identifikator i form av den internasjonale identifikatoren DOI (Digital object identifier) til datasett arkivert i UiT Open Research Data. Gjennom denne avtalen oppnår vi fordeler for de arkiverte datasett som f.eks.:

- inkludering av våre data i søketjenesten DataCite, noe som gir økt synlighet
- en sikker og varig referanse for datasett til bruk i siteringer og for andre kildehenvisninger
- et kvalitetsstempel i seg selv ved å være arkivert i et DataCite-godkjent arkiv

BIBSYS er nasjonalt bindeledd til DataCite, og for 2016 så dekker departementet avgiften for norske tildelinger av DOI på datasett. Men deretter må institusjonene basere seg på at de selv dekke denne kostnaden. Erfaringer fra Danmark tyder på at det årlige beløpet vil ligge på rundt kr. 20 000.

Støttetjenester for forskerne

Erfaringene her bygger først og fremst på UBs erfaringer fra TROLLing (to år) og kommentarer og erfaringer fra institusjoner i andre land.

Kartlegging av «tingenes tilstand» rundt forskningsdata som ble gjort i regi av Universitetsdirektørens arbeidsgruppe våren 2015 mot fakultetene, viste tydelig at det er manglende kunnskap om og forståelse av behovet for arkivering og langtidsbevaring av forskningsdata. En viktig forutsetning for det videre arbeidet må derfor være å styrke kunnskapen og bevisstheten om de krav og føringer som gjelder for datahåndtering, utarbeide gode veiledninger for forskerne, og gi tilbud om opplæring. I punktene nedenfor sier noe om hva dette kan og må resultere i.

a) Databehandlingsplaner

En databehandlingsplan er et dokument som beskriver hvordan datamaterialet i et prosjekt skal håndteres, fra innsamlingen av dataene til etter prosjektslutt. Ved å skisse opp et slikt helhetlig bilde i forkant av prosjektet kan en avdekke og eventuelt unngå utfordringer som er knyttet til oppbevaring av datamateriale. Hensikten er å sikre datamaterialet, ikke bare underveis i prosjektet, men også for fremtidig gjenbruk. Krav om slike planer inngår i retningslinjene fra Forskningsrådet og for prosjekter i Horizon 2020. Til nå er det få forskere som har erfaring med utvikling av databehandlingsplaner, og derfor også få som kan planlegge og beskrive håndteringen av forskningsdata for arkivering og langtidsbevaring på en god måte.

b) Kuratering – arkivverdige filformater

Kuratering beskriver hele prosessen fra innsamling, til arkivering, vedlikehold og bevaring av forskningsdata for samtidig og fremtidig bruk. Begrepet brukes særlig om vedlikehold og oppdatering av dataformater for å gjøre dem relevante for videre bruk over tid. Dette er et arbeid som UB har håndtert for TROLLing i hele driftsperioden og som også har vært en helt nødvendig oppgave overfor de forskerne som har kommet med data for pilotprosjektet. I noen tilfeller har dette vist seg å være et omfattende arbeid, spesielt når det skal gjøres og dokumenteres i ettertid (ref. punkt a over). Det er utarbeidet en

guide over filformater som tilfredsstillt kravet om langtidsbevaring og råd om verktøy og fremgangsmåter for å få konvertert sine data til arkiverbare formater.

ITA har sett på støtteverktøy som kan utføre effektiv konvertering av filer. En mulig løsning kan være at brukere laster opp et antall filer (gjerne Office filer) fra en webside og kan deretter laste ned innholdet som en ferdig Pdf/A (godkjent pdf arkiveringsformat). Konklusjonen er at ITA *kan* utvikle dette.

c) Kuratering – metadatabeskrivelse

Dette omfatter kvalitetssikring av nødvendig informasjon om et datasett for at forskningsdata skal kunne anvendes av andre og i ettertid. Slike metadata beskriver dataenes opprinnelse (hvor, når og hvordan forskningsdataene ble samlet inn) og informasjon som gjør det mulig å finne dataene, samt forstå om det er begrensninger på gjenbruk av disse. I TROLLing-sammenheng har metadatabeskrivelsen blitt gjort av den enkelte forsker, men kvalitetssikret av UB før publisering.

d) Sluttbrukerlisens på data

Det å gjøre data tilgjengelig uten å spesifisere bruksvilkår kan virke mot sin hensikt. Forskningsrådets anbefaling er derfor at det alltid knyttes en lisens til forskningsdata og at denne lisensen bør være internasjonalt anerkjente og legge så få begrensninger som mulig på dataene. Det finnes ikke noe enkelt svar på spørsmålet om hvilken lisens en bør velge for sine data, men den mest brukte lisensen internasjonalt når det gjelder forskningsdata er Creative Commons Zero (CC0). Denne er den lisensen som gir maksimal mulighet for deling og gjenbruk. Det som for noen kan være kontroversielt med denne lisensen er at den ikke *krever* at opphavsperson skal oppgis. Det finnes mange gode grunner¹ for at den likevel bør benyttes. For UiT Open Research Data vil CC0 være standard lisens med følgende tillegg:

Basert på fellesskapsnormene våre og akademisk skikk og bruk forventer vi at en viser til opphavskilden når en bruker data fra UiT Open Research Data. Vennligst bruk referansen som er generert for det aktuelle datasettet i UiT Open Research Data.

e) Opplæring

UB tilbyr allerede dette semesteret ulike kurs i håndtering av forskningsdata, fra utarbeidelse av databehandlingsplaner til konkret opplæring om bevaringsverdige filformater, beskrivelse av forskningsdata, sitering og deling av forskningsdata og bruk av dataarkiv for langtidsbevaring. Målgruppe denne våren er PhD-studentene som har fått tilbudet [Take control of your PhD journey: Manage your research data according to best practice](#), – et tilbud som har blitt svært godt mottatt.

Parallelt med dette så utarbeides det kortere opplæringstilbud som er mer rettet inn mot aktive og erfarne forskere. Opplæringsprogrammene vil ikke være begrenset til arkivering i UiT Open Research Data, men se på generelle problemstillinger uavhengig av hvilke arkiv som benyttes.

¹ Cornell University – Research datamanagement, Introduction to intellectual property rights in data management <http://data.research.cornell.edu/content/intellectual-property>

Forslag til forvaltningsmodell for forskningsdata

En operativ og pålitelig tjeneste forutsetter tydelig organisering med klart definerte roller og ansvar, og et mandat eller en oppdragsbestilling fra institusjonen. Det vil alltid være en fordel at forvaltning og drift av tjenester utføres av noen som har dette som ansvarsområde, noe som gjelder både for UB og ITA i UiT-sammenheng, hvor UB forvalter en felles kunnskapsinfrastruktur og ITA en felles IT-infrastruktur for UiT Norges arktiske universitet.

Organisering

I oppdragsbeskrivelsen lå det en oppfordring om å bygge på andres erfaringer, – en «best practice»-tilnærming. Nå er det ikke så mange institusjoner i vårt nærrområde som har etablert en operativ tjeneste på dette, så datagrunnlaget er ganske tynt. Men både Universitetsdirektørens arbeidsgruppe og dette prosjektet har sett på hvordan University of Bristol² har organisert noe som virker å være en velfungerende tjeneste. Deres *University of Bristol Research Data Service* bygger på et samarbeid mellom *IT Services* og *The University Library*, med enheten for *Research and Enterprise Development* som en strategisk partner. Dette er en institusjon som ikke er så veldig forskjellig fra UiT i størrelse og faglig bredde og som derfor kan være nyttig å lære av.

I denne omgang er derfor forslaget at det videre arbeidet og forvaltningen av forskningsdata organiseres som et felles samarbeid mellom UB og ITA, med avklarte roller og ansvar, og med AFU som partner for strategiske saker og institusjonell policy på området.

Roller og ansvar

Forvaltning av tjenesten er et felles ansvar for UB og ITA, men med følgende særskilte, overordnede ansvarsområder:

- UB er systemeier for det institusjonelle dataarkivet
- ITA har primæransvar for teknisk drift av systemet

Gjennom dette samarbeidet skal enhetene legge til rette for effektiv, forsvarlig og framtidsetta forvaltning av forskningsdata, – i dialog med fakultetene og AFU.

Videre er disse ansvarsområdene definert:

ITA har ansvar for følgende oppgaver som inngår i vanlig drift:

- Applikasjonsdrift, herunder konfigurasjonsretting av selve Dataverse (ikke innhold).
- Integrasjon av Feide i systemet (ikke støtte på brukerhåndtering, dette anses som innholdsdrift)
- Være teknisk støtte for applikasjon, siste instans hvis UB ikke får rettet problemer selv. ITA skal normalt ikke gi støtte direkte til brukerne.
- Oppgradering av basisprogramvare på server
- Generell drift for server/servere Dataverse kjører på

UB har ansvar for følgende oppgaver:

- Innholdsdrift, herunder brukerhåndtering og registrerte data fra brukerne
- Direkte støtte til brukerne
- Mindre systemendringer (blir dokumentert av ITA)

² University of Bristol Research Data Service, <https://data.bris.ac.uk/>

Videreutvikling av tjenesten må avklares og avtales med begge enhetene og håndteres som et utviklingsprosjekt der det anses hensiktsmessig. Dette kan f.eks. være :

- Utvikling av tilleggstjenester
- Oppdatere Dataverse til en ny hovedversjon
- Skalere løsningen for å tilfredsstille nye behov

Utvikling av UiTs policy for forvaltning av forskningsdata

Det første universitetet i England (og trolig ett av de første i Europa) som vedtok en egen policy for forvaltning av forskningsdata var The University of Edinburgh (UoE) som i mai 2011 fikk styrebehandlet dette. Deres policy har senere blitt mal for en rekke universiteter i England. UoEs viktigste argumentene for en policy var:

- Finansieringskildene ber institusjonene ta ansvar for resultatdata fra forskningen som de finansierer.
- Økt mobilitet blant forskerne og spesielt blant forskningseliten medfører betydelig større risiko for tap av data, og spesielt innen fagområder hvor store datavolum blir generert
- Ansvar for forvaltning av forskningsdata kan ikke ligge på den enkelte forsker alene. Det umuliggjør forvaltning av f.eks. lange tidsserier innen klimaforskning. Institusjonen må sørge for at forskerne har et arkiv for langtidsbevaring av data og med støttetjenester for metadata og kuratering.
- En policy for forvaltning av forskningsdata må balansere ansvaret for dette mellom institusjon og forsker.

Denne beskrivelsen gjelder også i dag og den gjelder for UiT Norges arktiske universitet.

Gjennom pilotprosjektet har vi etablert den første delen av en tjeneste for forvaltning av forskningsdata, – et institusjonelt arkiv for åpne data. Et naturlig neste trinn vil være å igangsette et arbeid som skal munne ut i et policydokument om forvaltning av forskningsdata ved UiT Norges arktiske universitet.

Et første tiltak i den retning kan være en uttalelse eller et vedtak om at UiTs politikk følger av Forskningsrådets "åpen som standard"-prinsipp når det gjelder tilgang til forskningsdata. UiT vil derfor legge til rette for at nødvendige infrastruktur- og støttetjenester er til stede for forskerne, og med det stimulere til at forskningsdata i utgangspunktet gjøres åpent tilgjengelig, men at det gjøres unntak for data som ikke kan eller bør gjøres tilgjengelig.

Samspill med andre aktører i forvaltningen av Forskningsdata

I en nasjonal kartlegging i 2015 blant 36000 forskere fra 24 universitet og forskningsinstitutter i Østerrike svarte forskerne at internasjonale fagspesifikke dataarkiv og institusjonelle dataarkiv ble mest benytta ved arkivering, men at også internasjonale tverrfaglige arkiv og nasjonale fagspesifikke arkiv ble benytta, om enn i betydelig mindre grad enn for de to første.

Behovet ved arkivering av data vil også variere enormt, fra forskeren som får krav om at grunnlagsdata for en publikasjon skal være arkivert og siterbare før en artikkel kan publiseres, og til store datavolum fra måleinstrumenter. Og krav om åpenhet eller høyeste grad av sikkerhet vil kunne være bestemmende for valg av dataarkiv. En fleksibel modell som tar hensyn til alle ulike behov, men som også ivaretar det ansvaret som institusjonen og den enkelte forsker har når det gjelder forvaltning av forskningsdata, vil måtte tilby og/eller godkjenne flere ulike tjenester for dette, fra institusjonsbaserte til nasjonale og internasjonale. Eksempler på dette er et åpent institusjonelt arkiv som UiT Open Research Data, i et

nasjonalt arkiv som NSD eller i et internasjonalt, tematisk arkiv som f.eks. ELIXIR - en pan-Europeisk infrastruktur for arkivering og analyser av biologisk informasjon. Men det følger ikke av dette at det bør være helt opp til den enkelte forsker hvor data arkiveres. Institusjonen bør ha en vurdering/godkjenning av eksterne arkiv for bl.a. å sikre at tilgang og rettigheter til data ikke overføres til kommersielle aktører. En tilsvarende ordning har vært gjennomført mellom anna av University of Bristol.

Innholdet i et lokalt arkiv kan høstes av sentrale arkiv og på den måten bygge gjenfinningstjenester på nasjonalt eller globalt nivå. Blant annet så vil nye NSD *kunne* høste nødvendige data fra et forskningsprosjekt som er arkivert i UiT Open Research Data, hvis disse også skal kunne gjenfinnes hos NSD. Dette er en infrastrukturmodell som er praktisert med stor suksess for de institusjonelle vitenarkivene (som Munin ved UiT). UiT Open Research Data bygger på felles metadatastrukturer og teknologi for høsting, noe som gjør at innholdet (primært metadata) kan bli spredt til en rekke søketjenester og dermed få maksimal synlighet.

Forslag til implementering i full skala – et åpent tilbud til alle ved UiT

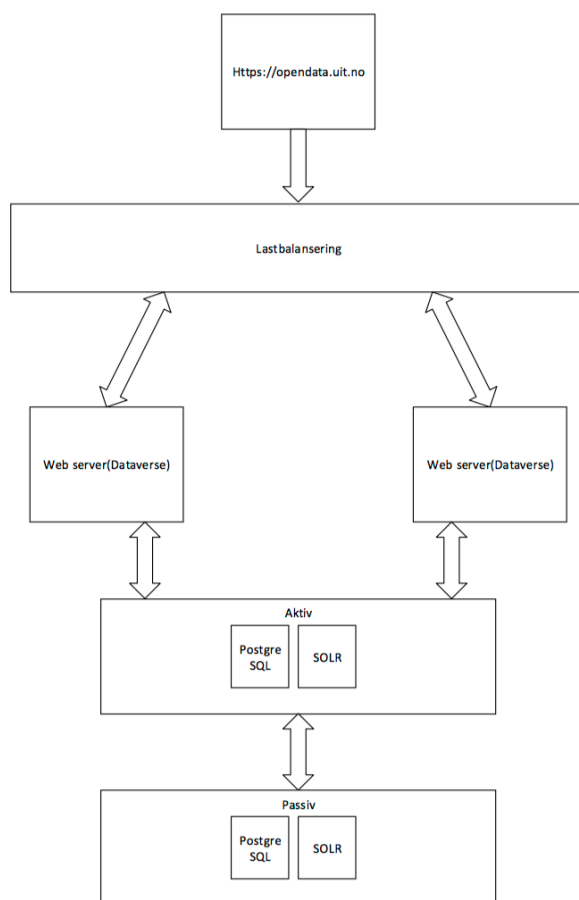
Når pilotprosjektet nå avsluttes med en nesten produksjonsklar tjeneste i *UiT Open Research Data*, så er den videre planen å fjerne «nesten» og gjøre den til en operativ tjeneste for forskere ved UiT Norges arktiske universitet. UB og ITA vil ta ansvar for tjenesten slik dette tidligere er beskrevet under Roller og ansvar. Men før vi lanserer tjenesten så er det noen tekniske endringer som må gjøres, tilsvarende ønsker vi å gi tjenesten en grafisk oppgradering, og vi må ikke minst sørge for at opplæringsstilbud og støttetjenester er på plass for forskerne.

Tjenesten er operativ nok fra dag en til at alle som har et konkret behov for å få arkivert sine data og som oppfyller forutsetningene for åpen arkivering, kan ta kontakt for hjelp til å komme i gang med arkivering av data. Men vi går ikke ut med aktiv markedsføring av tjenesten før vi har det vi vil kalle en operativ tjeneste (se tidsplan nedenfor).

Tekniske krav for skalerbarhet

I en første fase med få brukere og datasett så kan dagens tekniske løsning fungere godt. Det er umulig å si noe sikkert om hvor lenge dette vil vare (og vi *ønsker* den kortest mulig), men vi planlegger med at dette er status ett år fram i tid etter lansering. Deretter må tjenesten overføres til en mer langsiktig og robust teknisk plattform.

For at løsningen skal kunne skales opp fra dagens løsning til en ny og mindre sårbar løsning, kreves det i hovedsak flere virtuelle datamaskiner som systemet kan fordeles på og større kapasitet til datalagring. Neste nivå vil derfor være å øke antall maskiner fra en til minimum fire (se figur).



Forslag til ny arkitektur

Ut fra erfaringer med tjenesten UiT Open Research Data det kommende året så vil ITA kunne igangsette et prosjekt som utreder alle tekniske sider ved en ny arkitektur for tjenesten. Foreløpige tanker er at dette vil bety at en setter opp den nye plattform parallelt med dagens løsning, og deretter migrerer produksjonsdata over til ny plattform og kjører videre i produksjon der. Nærmere kartlegging og detaljavklaringer må gjøres innenfor et prosjekt for oppskalering. Et slikt prosjekt kan gjøres av interne utviklere på ITA sammen med Dataverse-kompetanse på UB og med bistand fra Dataverse-miljøet ved Harvard.

Støttetjenester – Tjenesteutvikling – hva mangler?

Det er et stort behov for mer kunnskap om og økt forståelse av arkivering og langtidsbevaring av forskningsdata. En viktig forutsetning for å få flere til å ta forvaltningsansvar for egne data er å styrke kunnskapen og bevisstheten om de krav og føringer som gjelder for datahåndtering, utarbeide gode veiledninger for forskerne, og gi tilbud om opplæring.

UB har igangsatt et opplæringstilbud nå i vår (2016) rettet inn mot PhD-studentene, som gir en første innføring innen temaet. Det er ønskelig at kurset kan bli tellende for studentene, for dermed å motivere flere til å gjennomføre kurset, og det er etablert en dialog om dette med High North Academy.

Det er ønskelig at det ved lansering også finnes spesialiserte, korte opplæringstilbud som er retta inn mot aktive og erfarne forskere, og som dekker ulike faser av livssyklusen til forskningsdata. Dette er noe UB nå jobber med å få ferdig. Aktuelle og sentrale tema er:

- **Databehandlingsplaner** – Hva er det? Hvordan oppfylle krav om dette fra finansieringskilder som Forskningsrådet og Horizon 2020? Støtteverktøy?
- **Håndtering av datafiler** – Om navngivning, organisering i mapper, håndtering av versjoner, og konvertering til filformater for langtidsbevaring, er noen av temaene som går på å legge et godt grunnlag for arkivering, helt fra starten av innsamlingen av data.
- **Beskrivelse av datainnhold** – hvordan beskrive innholdet slik at det kan forstås og gjenbrukes av andre enn forfatter og i ettertid. Hva er generelle metadata om et datasett og hvordan beskrive det på en standardisert form slik at egne data kan inngå i en større samling av data og kan søkes opp fra generelle søketjenester.
- **Sluttbrukerlisens** – Hvorfor knytte en lisens til data? Hvilken lisens skal jeg velge?
- **Arkivering av data** – Hva betyr det å arkivere data? Hva er forskjellen mellom lagring og arkivering? Hvor kan jeg arkivere? Hvilke krav stiller finansiererne til resultatdata fra forskning som de har bidratt til å finansiere? Hvilke krav kan jeg bli møtte med fra utgiver ved publisering?
- **Publisering av data** – Hva betyr det? Hvordan sitere et datasett? Tidsskrifter for publisering av data? Om deling av data på en sikker måte.

I tillegg til opplæringstiltak innenfor de ulike temaer så er det behov for veiledningsbeskrivelser på nett og enkle guider på papir som kan fungere som huske-/sjekklister for den enkelte.

Og ikke minst så har UB er korps på 15 fagansvarlige som er i tett dialog med og oppsøker «sine» forskermiljøer, og som der kan informere, avklare, svare på spørsmål og bistå forskerne innen forvaltning av forskningsdata.

Markedsføring

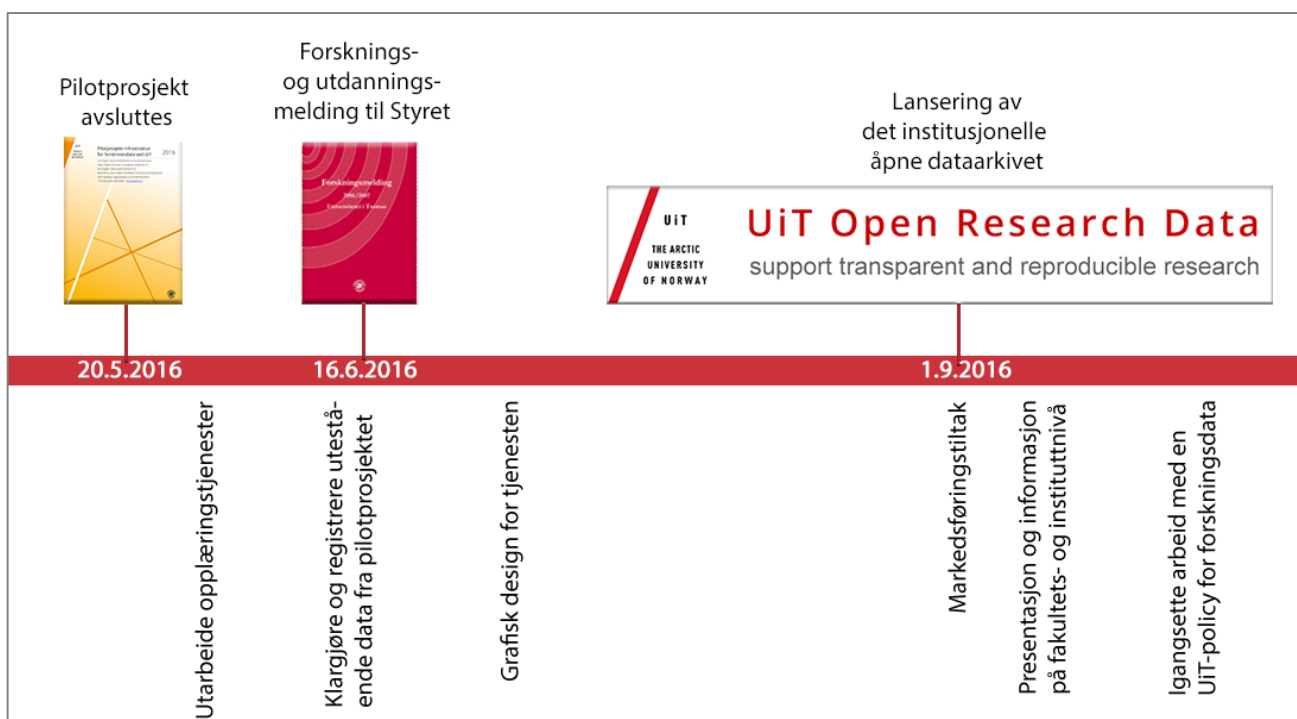
Når tjenesten UiT Open Research Data lanseres til høsten, så vil det være naturlig å gi den oppmerksomhet i form av nyhetsmelding på UiT-web, info i UiTs forskningsportal og på relevante nettsider for UB og ITA. I tillegg satser UB på å bruke alle våre kontakter og kontaktpunkter mot fagmiljøene til å spre informasjon om tjenesten. En egen kampanje retta inn mot seniorforskere som har

et ønske om å få arkivert data som de har liggende, er noe som kan være mulig å gjennomføre i løpet av høsten, men for øvrig så er det ønskelig det første året at veksten er gradvis og knytta til motiverte forskere. Det gir tjenesten en mulighet til å få luka ut feil, mangler og problemer, som alltid vil finnes i nye systemer.

Tidsplan med milepæler

Nedenfor er det identifisert tre viktige milepæler for utviklingen av UiT Open Research Data fra en eksperimentell tjeneste med TROLLing som pilotarkiv, til et institusjonelt arkiv for UiT Norges arktiske universitet for åpne data. I tillegg ligger det flere handlinger på tidsplanen som skal bidra til en implementering i full skala.

- Den første milepælen er nådd 20.5.2016 med en vellykket avslutning av pilotprosjektet. UB og ITA har vist at de med felles innsats har fått etablert en teknisk løsning for et åpent dataarkiv, og bygget nødvendig kompetanse til å gi opplæring og veiledning til forskerne, slik at dette kan stå fram som en fullverdig tjeneste for institusjonen.
- Den andre milepælen er 16.6.2016 og er kanskje noe på siden av pilotprosjektet og tjenesten UiT Open Research Data, men likevel svært viktig for implementering i full skala. For første gang vil Styret få seg forelagt en Forsknings- og utdanningsmelding hvor det står et eget kapittel om forskningsdata *Forskningsdata ved UiT – policy, infrastruktur og støttetjenester*. Dette vil øke oppmerksomheten på temaet, bidra til legitimitet i diskusjoner og i opplæring og formidling av informasjon om viktigheten av god forvaltning av forskningsdata. Så er det lov å håpe på en bonus i form av at det også kommer et tydelig budskap fra Styret om behovet for en UiT-policy på området, som ytterligere kan gi et løft.
- Og den tredje milepælen er naturlig nok selve lanseringen av tjenesten UiT Open Research Data. Den er satt til 1.9.2016. Det betyr ikke at UB og ITA er ferdig med å utvikle og legge til rette for en best mulig tjeneste for institusjonens forskere og inviterte forskerkolleger, men at da er minimumspakken klar for full lansering, inkludert nødvendige støttetjenester. Deretter vil den videre utviklingen skje innenfor løpende forvaltning av tjenesten.



Vedlegg

Invitasjon til forskere om å delta i pilotprosjektet

Det ble sendt ut en åpen invitasjon pr. e-post (nedenfor) til forskere ved de ulike fakultetene, fra «deres» kontaktperson (fagansvarlig) på biblioteket. I tillegg ble samme e-post sendt direkte til forskere som tidligere hadde vist interesse for arkivering og publisering av forskningsdata.

Til sammen kom det inn 14 svar på henvendelsen. Noen hadde spørsmål de ønska svar på, noen hadde kommentarer til initiativet, noen var nysgjerrige og ville gjerne delta, men hadde ikke data tilgjengelig for arkivering nå, – og noen svarte positivt og hadde data de kunne bidra med. Av de positive svarene så ble følgende 9 valgt ut til å bearbeides videre i pilotprosjektet:

Forsker	Forskningsdata	Kommentar
Lisbet Holtedahl, Professor Visuelle kulturstudier, Institutt for arkeologi og sosialantropologi, HSL-fak	Et omfattende arkiv over feltarbeid i flere Afrikanske land, samt i Nord-Norge over en 45 års periode. Inneholder egne noter og skrivelier, litteratur, billeder, lydbåndopptak og film. Noe er digitalisert, men det meste er ikke det.	Mottatt filer med film og lydspor. Må ha en litt mer spesifikk oppfølging for å definere metadata og en readme-fil.
Merete Andersen, Professor engelsk språkvitenskap, Institutt for kultur, litteratur og språkvitenskap, HSL-fak og Kristine Bentzen, Førsteamanuensis nordisk, engelsk, Institutt for kultur, litteratur og språkvitenskap, HSL-fak	Data består av 70 lydfiler (wav) og 70 transkripsjonsfiler (clan), hvorav kun de siste er anonymisert.	Vi må få tak i filene (clan) for å se på dem.
Prof. Trond Trosterud, Forsker Giellatekno – Samisk språkteknologi, Institutt for kultur, litteratur og språkvitenskap, HSL-fak og Ciprian-Virgil Gerstenberger, Overingeniør Giellatekno, Institutt for kultur, litteratur og språkvitenskap, HSL-fak		"Jeg kommer til å lage en pakke med noe dataer og deres metadata i CMDI format snart. Så sender jeg lenker til dokker. Ciprian."
Dr. Jannike Falk-Andersson, Forsker ressursøkonomi, Norges fiskerihøgskole, BFE-fak	Data fra en spørreundersøkelse rundt forvaltning av kongekrabbe – folks syn på ulike alternative forvaltningsstrategier.	Det meste OK, men anonymiseringsbehov. Har bedt henne om å gjøre anonymiseringene og sende oss nye filer.
JoLynn Carroll, Professor II, CAGE – Centre for Arctic Gas Hydrate, Environment and	Omfatter forskningsdata/bakgrunnsdata for en publikasjon. Datafiler i	Mottatt readme-fil "containing information on how the data were interpreted and how they

Climate, Institutt for geologi, NT-fak	excelformat, pluss noen bilder/illustrasjoner.	were derived (methods)"
Chris Hall, Professor, TGO – Tromsø Geofysiske Observatorium, NT-fak	Ionosonde data fra Tromsø Geofysiske observatorium. Både enkelte ionogrammer - grafikk-filer (siden 1980), og analyserte data siden 1935. Den siste blir en kombinasjon av skannet rapporter (pdf) 1935-1976 og ASCII siden 1977 – ca. 300GB tilsammen.	Eksempelfiler mottatt.
Rob Barrett, Professor, Seksjon for naturvitenskap, Tromsø museum	Ornitologiske felldata fra observasjoner og undersøkelser over en 30 års periode.	Har sendt oss en del excelfiler som må gås gjennom. Legges inn på sharepoint.
Gerit Pfuhl, Førsteamanuensis, Institutt for psykologi, Helsefak		Har sendt oss to iqdat-filer og en wordfil. Må gås gjennom. Legges inn på sharepoint.
Kari Anne Bråthen, Førsteamanuensis AMB, BFE.	Observasjonsdata innen planteøkologi, fra Finnmark og Frankrike, og analyseresultater.	Vi mangler nærmere beskrivelse (readme) og klargjøring av datafilene (rådata, i tillegg til analyseresultater fra R-script)

E-post til forskerne

Til forskere ved / To researchers at xx fakultet
(for English, please scroll down)

UB og ITA starter nå et felles pilotprosjekt om arkivering av forskningsdata. Denne eposten er en invitasjon til å bidra med dine data i pilotprosjektet, som etter planen skal være avsluttet 1. april 2016.

Hvorfor dette er nyttig for deg og for oss:

Pilotprosjektet skal avklare behov og finne gode løsninger for håndtering og arkivering av forskningsdata. Dette vil være til stor nytte for deg som forsker. Arkivering og tilgjengeliggjøring av forskningsdata vil bety økt synlighet av din forskning, og følgelig økt sannsynlighet for siteringer. Videre er en god infrastruktur for håndtering av forskningsdata i økende grad viktig i søknader om forskningsfinansiering, og gjennom deltakelse i pilotprosjektet vil du opparbeide deg nyttig kompetanse på temaet. Har du innsamlede eller genererte forskningsdata, og ønsker å delta, kan du ta kontakt med meg eller med UBs prosjektkoordinator: Leif Longva (leif.longva@uit.no), tlf. 7764 6255.

Litt mer om pilotprosjektet:

Tilgjengeliggjøring av forskningsdata er et svært aktuelt tema. Stadig flere forskningsfinansieringskilder stiller krav om arkivering av data som innsamles eller genereres i forskningen de finansierer. I tillegg blir det mer og mer vanlig at tidsskriftene ber om arkivering av data som ligger til grunn for artiklene de publiserer. Bakgrunnen for dette er at forskningsresultatene lettere skal være etterprøvbare, og at det vil effektivisere forskningen ved at det blir lettere å bygge videre på tidligere forskning. For å gjøre det lettere

for våre forskere å imøtekomme slike krav, lanserer UiT nå dette pilotprosjektet. Piloten skal danne grunnlag for å etablere en best mulig tjeneste for forvaltning og arkivering av forskningsdata fra UiT. Pilotprosjektet skal avklare hvilke behov forskerne har rundt dette, og definere hvilke støttefunksjoner og teknisk infrastruktur en slik tjeneste bør inneholde.

Pilotprosjektet vil begrenses til kun å omfatte forskningsdata som kan være åpent tilgjengelige, og persondata og andre sensitive data faller således utenfor i denne omgang. Forskningsdata som kan gjøres åpent tilgjengelige om noen få år kan imidlertid inkluderes i prosjektet. De som ønsker å delta må være tilgjengelige for spørsmål og avklaringer i de seks månedene pilotprosjektet varer. Om interessen for å delta blir stor, må prosjektgruppen forbeholde seg retten til å velge ut hvilke datasett som skal inngå i pilotprosjektet.

Mange takk for samarbeid og hjelp med dette.

[english]

The Library and the IT department at UiT is currently launching a pilot project on archiving of research data. Through this e-mail we invite researchers at UiT to take part in our pilot project, scheduled to end by 1 April 2016.

Why this is beneficial for you and for us:

The main objective of the pilot is to identify the needs of the researchers, and find good solutions for managing and archiving research data. This will be most beneficial for you as a researcher, as archiving and open access to the data means increased visibility of your research, and thus an increased likelihood of being cited. Furthermore, a good infrastructure for managing your research data is more and more important in applications for research funding, and by participating in the pilot project you will gain valued knowledge on the issue.

If you have collected or generated research data, and you wish to contribute to the pilot, please contact me directly, or the project coordinator at the library: Leif Longva (leif.longva@uit.no), tel. 7764 6255.

A little more about the pilot project:

Access to research data is a topic of high current interest. More and more research funders ask for the collected or generated data from the research they fund to be archived. Furthermore, an increasing number of journals require archiving of the data on which the research articles they publish are based. The motivation behind this is that access to the data will make research more easily verifiable, and that research more efficiently may build on previous research. In order to help our researchers meet these new requirements, UiT is now launching this pilot project. The pilot will enable us to establish a best possible local service for managing and archiving research data. The pilot aims to determine the needs of the researchers on this issue, and which support services as well as technical infrastructure this service will need to include.

The pilot project will be limited to open research data only. Data identifying individuals or other sensitive data thus cannot be included. Research data that can be made openly available in a few years from now, may, on the other hand, be included. If you want to contribute, you must agree to be available for questions and clarifications during the six months duration of the project. Note that in the case we get many possible contributors, the project group reserve the right to decide which data to include and not. Thank you for your cooperation and help.

Mvh / Yours,
«fagansvarlig UB»

