Integrationen mellan lärosätets egen lagring och SND (DORIS)

– *Ett förslag på arbetsgång*

Datum: 2022-01-14

Version: 9

Svensk nationell datatjänst
snd.gu.se
031-786 10 00
snd@gu.se
SND, Göteborgs universitet,
Box 463, 405 30 Göteborg

Innehåll

Inledning

SND har tagit fram ett förslag på hur arbetsgången kan se ut från det att en forskare placerar sitt forskningsmaterial i lärosätets egen lokala lagring fram till det att materialet går att söka fram i SND:s forskningsdatakatalog, och antingen direkt laddas ner eller utlämnas efter prövning. Först ges en förenklad genomgång av arbetsflödet och därefter följer en teknisk beskrivning av integrationen mellan lärosätets egen lagring och SND:s system (DORIS).

Det här dokumentet är till för DAU-medarbetare vid en organisation som ingår i SND-nätverket. Första delen av dokumentet riktar sig till alla medarbetare i en DAU-grupp och kräver inga tekniska förkunskaper. Den andra delen av dokumentet är främst till för IT-medarbetare. Det är viktigt med ett nära samarbete mellan DAU-gruppen och organisationens IT-avdelning, och IT-medarbetare kan med fördel ingå i DAU-gruppen.

Det här dokumentet är ***inte*** en mall för hur arbetet med lagring av forskningsdata ska hanteras vid ett lärosäte. Det är heller inte en lösning som kan användas rakt av, utan är just ett övergripande förslag på hur ett arbetsflöde för att tillgängliggöra forskningsdata kan se ut, med lagringen som huvudfokus. Det är sedan upp till varje lärosäte att plocka valda delar och anpassa dem efter de behov som finns. Dokumentet beskriver inte DORIS eller arbetsflödet i DORIS (t.ex. hur man som DAU granskar metadata och data). Information om hur en databeskrivning skapas och hanteras i DORIS hittar du bland annat på SND:s hemsida https://snd.gu.se/sv och i DAU-handboken https://dhb.snd.gu.se.

SND finns som stöd för DAU:en, både under processen att ta fram rutiner för dataflöde och efter att lagringen är etablerad. Vi kan också hjälpa till med frågor som rör förslag på mappstruktur, paketstrukturer och filmanifest med till exempel BagIt[1] samt andra delar som gäller lagringen.

Förklaring av begrepp

DORIS (SND:s DataORganiserings- och InformationsSystem)

DORIS är SND:s digitala dokumentationssystem. Det används för att forskare ska kunna beskriva forskningsprojekt och göra forskningsdata sökbara och tillgängliga i SND:s forskningsdatakatalog. https://doris.snd.gu.se

Lokal lagring

Med detta menas:

- **lagringslösning**: det eller de IT-system som organisationen använder för lagring av filer. I det här dokumentet avses den lagring som har koppling till DORIS. Detta kan vara både en molntjänst och/eller en lokal lagringsplattform hos organisationen.

[1] https://en.wikipedia.org/wiki/BagIt (2021-02-26)

- **göra data tillgängliga**: i det här dokumentet avses data som finns beskrivna i SND:s forskningsdatakatalog och antingen kan laddas ner direkt eller nås via förfrågan (gäller data som innehåller sekretessbelagd information, eller som av andra skäl inte kan ligga öppet tillgängliga, exempelvis på grund av storlek) från forskningsdatakatalogen.

Forskare

Används i det här dokumentet som samlingsbegrepp för den eller de personer som delar data via SND:s forskningsdatakatalog genom DORIS. Förutom organisationens anställda forskare kan det också vara doktorander eller andra medarbetare, som exempelvis data managers.

Lagring

Alla lärosäten har informationstillgångar där forskningsdata utgör en stor del. Eftersom lagringsfrågan till stor del är en resursfråga så hanteras den olika på olika lärosäten. Faktorer som kan påverka val av lagringslösning är:

- Lärosätets storlek
- Existerande system och andra upphandlade IT-lösningar
- IT-resurser
- Typ av data och lagring utifrån informationsklassning (t.ex. ställer sekretessbelagda data högre krav på säkerheten än data som inte omfattas av sekretess)
- E-arkivering och hur den integrerar med andra lagringslösningar
- Lärosätet kan ha behov av fler än en lokal lagringslösning
- Molntjänster för datalagring

Det finns fler påverkande faktorer än de som nämns ovan, men oavsett hur lagringen hanteras lokalt på ett lärosäte är tanken att den ska kunna kopplas till SND:s system DORIS och forskningsdatakatalogen via ett API[2]. Forskningsdata lämnar därmed aldrig lärosätet, men kan ändå laddas ner eller förmedlas via SND:s katalog.

Förslag på arbetsgång

Här följer ett förslag på arbetsgång som sträcker sig från det att forskaren laddar upp data på lärosätets utsedda lagringsyta, till att DAU granskar data och slutligen till att data görs tillgängliga i SND:s katalog, antingen via direkt nedladdning eller utlämnas efter prövning. Arbetsgången beskrivs i olika steg, men det är inte självklart att det just är i den här ordningen arbetet sker. Arbetsgången

[2] API, *Application Program Interface*, är ett slags protokoll som används för att program ska kunna prata med varandra. Man kan likna det vid en slags tolk, som kommunikationen går genom.

påverkas exempelvis av när i forskningsprojektet som en forskare får åtkomst till en lagringsyta och vilken/vilka lagringslösningar som används.

Steg 1: Forskaren får tillgång till en lagringsyta av sitt lärosäte

Oavsett hur lagringen av forskningsdata hanteras på lärosätet behöver det finnas en lagringslösning på plats där DAU:en har (eller kan ges) rättigheter att läsa och kurera forskningsdatafiler som ska delas via SND:s forskningsdatakatalog. Innan data delas bör både forskare och DAU ha skrivrättigheter till filerna, men efter att katalogposten publicerats ska endast läsrättigheter finnas.

För forskaren är det enklast om lärosätet rutinmässigt tillhandahåller en lagringsyta till alla anställda forskare och som går att använda under hela forskningsprocessen. Är en sådan lösning inte möjlig behöver den forskare som vill dela data via SND:s forskningsdatakatalog få tillgång till en yta i samband med att data ska göras tillgängliga. Lärosätet beslutar vilka kontaktvägar som ska finnas för detta, hur och var kontakt tas, exempelvis görs det direkt i DORIS eller via det egna lärosätets medarbetarportal/intranät.

DAU:en (eller någon annan funktion) ser till att det skapas en lagringsyta för forskaren eller forskargruppen, eller så ges forskaren tillgång till en mapp på en redan befintlig lagringsyta. Forskarens lagringsyta finns antingen lokalt på det egna lärosätet eller i en molnbaserad lagringstjänst (som exempelvis SUNET:s tjänst eller någon annan lösning). Ytan kan vara helt tom eller ha en mappstruktur som är bestämd av lärosätet. Forskarens lagringsyta registreras i en indextjänst (se nedan under *Teknisk beskrivning*) så att DORIS kan hitta den och därefter kan data delas via SND:s forskningsdatakatalog. När en lagringsyta väl har skapats och registrerats i indextjänsten behöver den inte registreras på nytt utan kan därefter användas för att dela projektets data.

Tänk på:

- Forskaren behöver få tydlig information om att en särskild lagringsyta krävs och hur den tilldelas och hur den kan/ska användas.
- Fler än en forskare kan behöva få tillgång och rättigheter till lagringsytan.
- DAU:en behöver kunna granska filer som lagts på lagringsytan, och behöver alltså som minst läsrättigheter för filerna.
- Det kan vara en god idé att standardisera hur lagringsytor skapas och utformas, och att exempelvis ha tydliga rutiner för behörighetsadministration. Här har DAU:ens IT-medarbetare en viktig roll.
- Det behöver finnas en lokal strategi för hur data ska hanteras utifrån ett långtidsperspektiv.
- Filerna bör struktureras på ett konsekvent och begripligt sätt. Det är särskilt viktigt för publicerade versioner av dataset. SND har ett tagit fram ett vägledningsdokument om hur en mapp-/ filstruktur bör utformas.

- Ett snabbt, automatiserat flöde förenklar för både forskare och DAU, men DAU:ens kontroll över data minskar.
- En komplicerad lösning där forskaren själv måste kontakta DAU kan avskräcka forskare från att vilja dela data.
- En lärosätespolicy för öppen vetenskap och delning av data kan vara bra att ha på plats.

Steg 2: Filer laddas upp

En forskare som har fått rättigheter till en lagringsyta kan därefter ladda upp forskningsdata som ska delas via katalogen. Det görs via ett användargränssnitt, till exempel Nextcloud, eller via en nätverksenhet. Om man vill går det bra att använda ett annat användargränssnitt. Beroende på vilka rättigheter som finns för ytan så skulle DAU-medarbetare teoretiskt sett kunna hjälpa forskaren med att ladda upp filer. Rättighetshanteringen är någonting som kräver lite eftertanke och är upp till varje enskild organisation att avgöra. Ett alternativ är till exempel att rättigheter hanteras per fil istället för att gälla hela ytan, och att tillgång/läsrättigheter för DAU aktiveras i samband med att filerna pekas ut i DORIS.

Ytan där filerna hamnar tillhör lärosätet och de enda som bör ha tillgång till filerna är forskaren/ forskargruppen och DAU. Än så länge har forskaren rättighet att ta bort eller ändra i filer på sin lagringsyta, men när den färdiga databeskrivningen (metadata och data inkluderat) har publicerats i SND:s forskningsdatakatalog sker en rättighetsförändring hos filerna så att de endast kan läsas och inte längre redigeras, varken av forskaren eller DAU. Det säkerställer att filerna inte kan ändras efter publicering, eftersom de då har tilldelats en DOI[3]. Läs mer om detta under *Steg 6*.

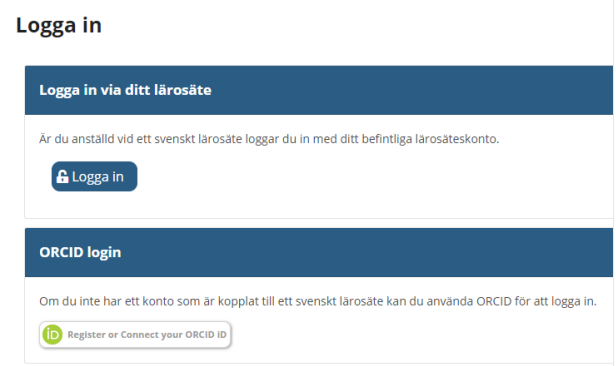
Tänk på:

- Forskare kan vilja dela stora dataset, med många och/eller stora datafiler, och därför kan det behövas kapacitet för att hantera stora datamängder.
- Data kan innehålla skyddsvärd information. Lagringsytan måste uppfylla en säkerhetsklassning som matchar de data som lärosätet producerar och rättigheter till ytan bör begränsas så att endast de som verkligen behöver ha åtkomst har det.
- Det behövs ett användargränssnitt för filhantering som kan användas av forskare och DAU.

Steg 3: En databeskrivning påbörjas i DORIS

Forskaren loggar in på SND:s webbplats med sin SWAMID-inloggning (samma inloggningsuppgifter som till det egna lärosätets system) och påbörjar en databeskrivning i DORIS.

[3] DOI, *Digital Object Identifier*, är en persistent identifierare (PID) som tilldelas varje version av ett dataset. PID:ar är unika och beständiga digitala referenser som gör det möjligt att hitta, citera och återanvända digitalt material.



Steg 4: Filer väljs ut

När forskaren pekar ut filer (alltså väljer vilka filer som ska delas via forskningsdatakatalogen) så kommunicerar DORIS med den lagringsyta där forskaren har laddat upp data. För att DORIS ska hitta till rätt lagringsyta vid rätt lärosäte går kopplingen via ett API som sitter som en “grindvakt”. API:et kommunicerar med både DORIS och den lagringslösning lärosätet har valt. Vilken lagringslösning som används har ingen betydelse förutom att den måste kunna ”prata” med SND:s API. När API:et har säkerställt att forskaren är den han/hon säger sig vara och har behörighet till just den här ytan får forskaren se sina filer på ungefär samma sätt som man ser filer i Utforskaren på datorn. Forskaren kan inte se att kontrollen görs utan upplever det som att kopplingen hittas automatiskt.

När forskaren har valt de filer som ska delas och känner sig färdig med databeskrivningen skickas den vidare till DAU:en för granskning. DAU:en får då information om vilka filer som kräver granskning.

Tänk på:

- Det kan finnas fler filer på lagringsytan än de som ska delas. Forskaren behöver därför vara noggrann med vilka filer som väljs/pekas ut.
- Forskaren kan ha påbörjat en databeskrivning i DORIS innan data finns på en lagringsyta. Filer kan utan problem läggas till på lagringsytan under tiden metadata färdigställs.
- Rätt personer behöver ha behörighet till rätt ytor.

Steg 5: DAU granskar

Så snart forskaren har skickat in sin databeskrivning får lärosätets DAU ett mejl om att det finns en ny databeskrivning som behöver granskas i DORIS. DAU:en granskar metadata och data och tar ställning till om data kan vara direkt nerladdningsbara eller inte (rätt tillgänglighetsnivå). Om data eller metadata behöver kompletteras eller korrigeras på något sätt meddelar DAU:en forskaren. Behövs en komplettering skickas databeskrivningen tillbaka till forskaren för att åtgärdas. Forskaren har då möjlighet att uppdatera metadata och välja nya filer. När både forskare och DAU är nöjda med metadata och data kan databeskrivningen publiceras i SND:s forskningsdatakatalog.

Tänk på:

- Om forskaren kan radera eller ändra filer på den egna lagringsytan innan databeskrivningen har publicerats är det viktigt att DAU:en informerar forskaren om vikten av att filer inte ändras eller flyttas under granskningsprocessen. Det bästa är om läs- och skrivrättigheter för filer automatiskt följer databeskrivningens status.
- Alla DAU-medarbetare som behöver kunna granska data behöver ha behörighet att läsa data- och dokumentationsfiler.

Steg 6: Data går att hitta i forskningsdatakatalogen

När databeskrivningen är publicerad i SND:s forskningsdatakatalog kan de data som forskaren valt ut nås antingen via direkt nedladdning eller via förfrågning om data genom DAU, beroende på vilken tillgänglighetsnivå som har valts för databeskrivningen. På sikt kommer de metadata som anges i DORIS även överföras till lagringsytan för att lagras tillsammans med data. För att möjliggöra olika e-arkivlösningar kommer DORIS i framtiden även att stödja arkivmetadata. Ett e-arkiv skulle då till exempel kunna hämta data från lagringsytan in i arkivet, alternativt att man bestämmer att data, tillsammans med tillhörande metadata, ska struktureras på ett sådant sätt att det kan fungera som en del av ett e-arkiv.

Tänk på:

- Varken forskaren eller DAU ska, från och med publicering, ha rättighet att ändra eller ta bort publicerade filer. Eventuellt kan en *superuser*[4] ges rätt att ändra filer, t.ex. vid gallringsbeslut. Om en katalogpost behöver avpubliceras behöver det finnas särskilda rutiner för det. SND kan bistå med råd kring eventuell avpublicering. Då DOI sätts på data behöver en så kallad, graveyard-sida finnas kvar med information om varför data inte längre finns tillgängliga.
- Andra opublicerade filer på samma lagringsyta kan fortfarande redigeras.
- Behöver något ändras i en publicerad datafil måste det skapas en ny version av filen som läggs på lagringsytan och kopplas till DORIS. Det här medför att man publicerar en ny version av datasetet. Varje ny version av ett dataset tilldelas en ny DOI.

Steg 7: Leverans av data

Det är framförallt innehållet i datafilerna som avgör om data kan vara öppet tillgängliga eller om det krävs någon form av prövning innan de kan lämnas ut.[5]

[4] Superuser: en användare som har högsta tänkbara behörighet i ett system. Funktionen finns inte i dagsläget.
[5] Mer om SND:s tillgänglighetsnivåer för data finns att läsa här: https://snd.gu.se/sv/hitta-data/forskningsdatakatalogen/tillganglighetsnivaer-hos-snd

För att filer ska kunna vara direkt åtkomliga för nedladdning i SND-katalogen krävs att de finns lagrade på en lagringsyta med möjlighet att skapa länkar direkt till filerna.

Data som innehåller skyddsvärd information kan inte vara direkt nedladdningsbara i forskningsdatakatalogen. De behöver levereras på något annat sätt än via en direktlänk. Tänk på att någon form av prövning alltid måste göras innan den här typen av data kan lämnas ut.

SND:s API-lösning omfattar inte utlämnande av skyddsvärda data som nås via förfrågan. Däremot finns stöd i DORIS för att ange ett meddelande i samband med att en förfrågan godkänns, för att t.ex. informera om åtkomst till datafilerna.

Tänk på:

- Data som innehåller skyddsvärd information måste delas på ett sätt som inte strider mot gällande lagstiftning, t.ex. bör säkra krypterade tjänster användas. Direktiv för detta kan eventuellt anges i lärosätets IT-säkerhetsregler/policys eller informationssäkerhetsregler.
- Även om datafiler inte kan delas öppet i katalogen behöver tillhörande dokumentationsfiler finnas direkt nedladdningsbara i katalogposten.
- Filer som av praktiska skäl inte kan delas genom direkt nedladdning, till exempel på grund av storlek eller komplex fil-/mappstruktur, kan behöva andra metoder för nedladdning.
- De data som är arrangerade i en komplex mappstruktur eller innehåller väldigt många individuella filer kan packas i till exempel en zip-fil för att bibehålla mappstruktur och förenkla nedladdning.

Teknisk beskrivning

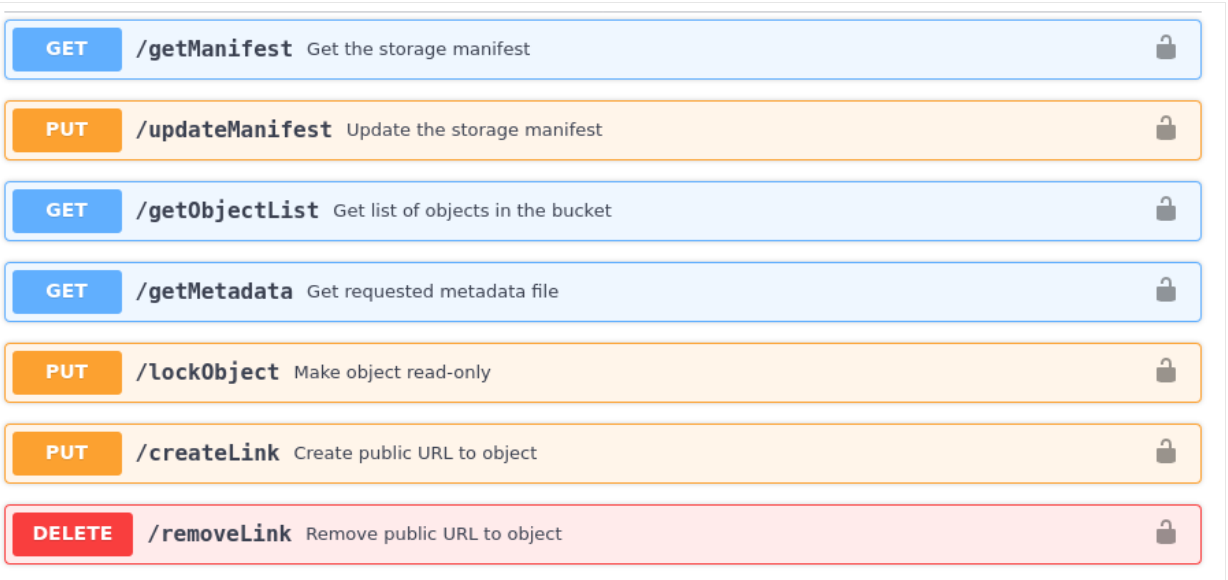
Här följer en kortare teknisk beskrivning av det föreslagna arbetsflödet som avslutas med ett flödesschema. I första hand riktar sig informationen nedan till DAU-medarbetare med olika slags IT-kompetenser.

Lagrings-API:et är fortfarande under aktiv utveckling och kommer få en formaliserad dokumentation och versionsnummer. För mer uppdaterad information följ länkarna till GitHub.

Metadata-API:et

Serverapplikation som är kopplad till lagringen. Applikationen ansvarar för att tillhandahålla metadatamanifest och objektslista för varje lagringsyta (eller "lagringsbucket"). Har funktioner för att låsa publicerade filer och skapa publika länkar för öppna data."

https://github.com/SUNET/s3-metadata-api/



Indexservern

Tjänst dit metadata-API:et skickar metadata-manifest som indexeras. DORIS kan sedan ställa frågor till indexservern för att hitta metadata för en användares lagring.

https://github.com/SUNET/metadata-index

Lagringsfrontend

Gränssnitt som forskare och DAU kan använda för att hantera data som lagras, t.ex. Nextcloud.

Metadatamanifestet

Ett json-ld-dokument som innehåller grundläggande metadata om lagringen.

https://github.com/snd-sweden/data-storage-information-interface

Exempel:

```
{
    "@context": {
        "mm": "https://raw.githubusercontent.com/SUNET/metadata-
manifests/master/schema/mm.jsonld"
    },
    "@id": "192.0.2.0/getManifest?bucket=research-data-2019",
    "mm:publisher": "SUNET",
    "mm:creator": "kalkyl@example.edu",
    "mm:identifier": "3f2a775e-f33a-461e-a49c-b9450c7da5",
    "mm:rightsHolder": "example.edu",
    "mm:manifest": [{
        "@id": "META-DATA/dc.xml",
        "mm:schema": "dct"
    }]
}
```

Objektslistan

Lista över filerna i en bucket. Exempel:


```
    [
        {
          "identifier": "9e3bb2e3-443c-4763-97f3-a25eb9c8639d",
          "name": "projektet2021/data/readme.pdf",
          "dateCreated": "2020-09-15T07:16:46.329Z",
          "dateModified": "2021-01-15T07:20:46.978Z",
          "contentSize": 123,
          "contentUrl": "https://s3.example.com/2021/data/readme.pdf",
          "encodingFormat": "application/pdf",
          "immutable": false,
          "checksum": [{
            "checksumValue":
"39ae639d33cea4a287198bbcdca5e6856e6607a7c91dc4c54348031be2ad4c51",
            "checksumAlgorithm": "checksum_algorithm_sha256"
          }]
        },
        {
          "identifier": "eae959d3-a17d-42af-9510-44624213efca",
          "name": "projektet/2021/data/data.csv",
          "dateCreated": "2020-09-18T07:16:46.329Z",
          "dateModified": "2020-09-18T07:16:46.329Z",
          "contentSize": 76912,
          "contentUrl": "",
          "encodingFormat": "text/csv",
          "immutable": true,
          "checksum": [{
            "checksumValue": "45ef90d63b90328915a561a1c88b2945",
            "checksumAlgorithm": "checksum_algorithm_md5"
          }]
        }
    ]
```


Sammanfattande beskrivning av flödet

1. En lagringsyta/bucket skapas åt forskaren. Exakt hur forskaren går tillväga för att få en yta kan se olika ut, men ett alternativ är att forskaren får ansöka om/beställa lagringen via någon webbportal.

2. Manifestet skapas och lagras på forskarens yta, i en katalog som heter “.metadata”. Manifestet kan skapas av samma portal eller verktyg som i steg 1. SUNET har tagit fram ett hjälpverktyg för att skapa manifest när det gäller en s3-lösning: https://github.com/SUNET/s3-mm-tool

3. Metadata-API:et upptäcker genom sin koppling till underliggande lagring att en ny lagringsyta har skapats, detta kan ske antingen genom någon slags signal, eller att applikationen regelbundet kollar om någon ny yta har skapats.

4. Metadata-API:et registrerar manifestet hos Indexservern genom att göra ett */register*-anrop till Indexservern.

5. Forskaren laddar upp data via en lagringsfrontend, till exempel Nextcloud.

6. Forskaren påbörjar en databeskrivning i DORIS.

7. DORIS skickar då en fråga med forskarens ID (eppn från SWAMID) till Indexservern som svarar med URL till samtliga manifest som är kopplade till forskarens ID.

8. DORIS frågar sen respektive Metadata-API om objektslista, *getObjectList*, för varje manifest. Alla objekt presenteras i DORIS gränssnitt så att forskaren kan peka ut vilka filer som databeskrivningen gäller.

9. När databeskrivningen är färdig i DORIS och ska publiceras skickas ett */lockObject*-anrop för varje objekt som ska publiceras. Då blir filerna skrivskyddade så att publicerade datafiler inte kan ändras i efterhand. I de fall filerna ska vara direkt nedladdningsbara skickas också ett */createLink*-anrop för varje objekt. Länken publiceras tillsammans med databeskrivningen i SND:s forskningsdatakatalog.

Flödesschema