Vägledning för filhantering och mappstruktur

Datum: 2021-10-26

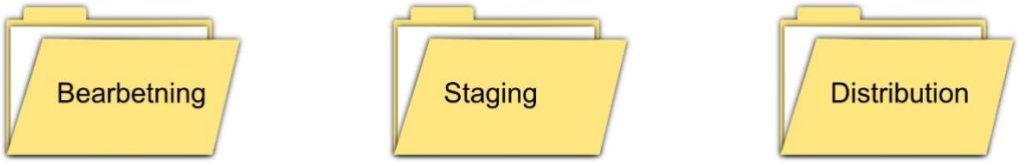
Version: 1

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

Lärosäten och forskande organisationer inom SND-nätverket kommer allt närmare egen lokal lagring för forskningsdata med koppling till DORIS och forskningsdatakatalogen via lagrings-API:et. Därför är det också viktigt att fundera på hur filer ska hanteras i lagringen. I det här dokumentet går vi igenom SND-kontorets filhantering, vilken kan fungera som underlag för hur filer kan hanteras lokalt. Det är viktigt att poängtera att filhanteringen är organisationernas egna ansvar, och att dokumentet endast erbjuder ett vägledande exempel.

Filstrukturen inom SND CARE

Här ges en beskrivning över hur filstrukturen ser ut inom SND CARE[1]. Grunden i filstrukturen utgörs av tre huvudmappar: *Bearbetning*, *Staging* och *Distribution*.



Bearbetning: Har en direktkoppling till formuläret i DORIS och för varje ny databeskrivning som skapas tillkommer här en mapp med databeskrivningens ID. Det är i bearbetningsmappen som arbetet med att granska data, konvertera filformat och förbereda filer för tillgängliggörande sker.

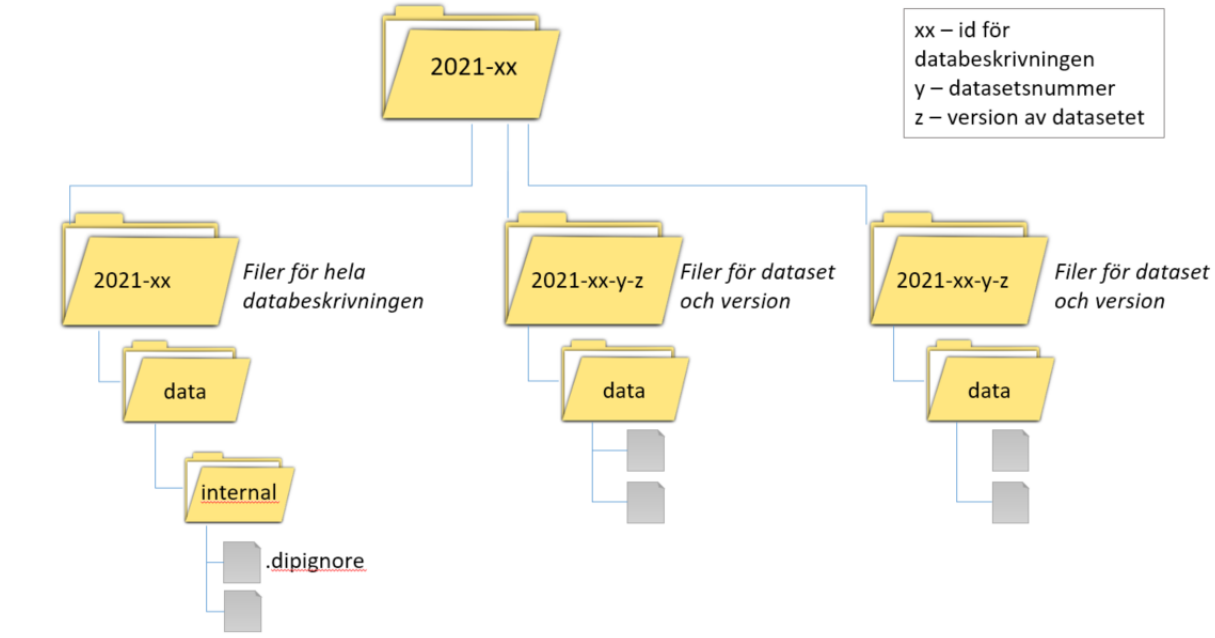
Staging: När en databeskrivning har publicerats i SND:s forskningsdatakatalog flyttas hela filpaketet automatiskt till Staging-mappen. Enligt SND-kontorets rutiner finns också en särskild yta avsedd för långtidslagring av data. När filpaketen ligger i Staging-mappen visar detta att filerna tillhör en publicerad databeskrivning och att filpaketet är redo att flyttas till ytan för långtidslagring. Denna flytt sker manuellt.

Distribution: Delen av lagringen som har en direktkoppling till SND:s forskningsdatakatalog. Här finns en kopia av de filer som ska vara direkt åtkomliga i katalogposten.

Bearbetning

När en databeskrivning påbörjas i DORIS skapas automatiskt en mapp för denna i filstrukturen. Mappen tilldelas automatiskt samma namn som databeskrivningens ID i DORIS. I denna huvudmapp finns underliggande mappar: en för sådant som rör hela databeskrivningen och en eller flera mappar för filer som är kopplade till dataset och version (se bild nedan). I mappen/mapparna som tillhör datasetsversionerna läggs aktuella datafiler och dokumentationsfiler.

[1] SND CARE är SND:s CoreTrustSeal-certifierade repositorieverksamhet.



Mapp 2021-xx: Heter samma sak som huvudmappen och kan används för filer som gäller för hela databeskrivningen, till exempel interna dokument som inte ska delas tillsammans med datafilerna men som ändå behöver sparas tillsammans med materialet. För att filerna inte ska synas i forskningsdatakatalogen använder SND ett system med *.dipignore-filer*. En .dipignore-fil är en tom textfil som markerar att innehållet i den mapp den ligger i inte ingår i ett förmedlingspaket och därmed inte ska visas upp i forskningsdatakatalogen.

Mapp 2021-xx-y-z: Här ligger filer som hör till ett specifikt dataset och en specifik version. Det kan vara både datafiler och dokumentationsfiler. Vilken tillgänglighetsnivå som valts i DORIS avgör om filerna för datasetet (*2021-xx-y-z*) är synliga i SND-katalogen eller inte.

Mapparna som beskrivs ovan innehåller alla en undermapp som heter ”data”. Vid publicering skapas automatiskt en BagIt ”bag” för respektive undermapp, i bilden ovan mapparna 2021-xx och 2021-xx-y-z (en viss fördröjning kan förekomma). På sidan 6 kan du läsa mer om vad BagIt är och hur det kan användas. Mappen “data” måste finnas med för att BagIt ska fungera. Det finns inga krav på att filerna i datamappen måste vara strukturerade på ett visst sätt, utan de kan ligga direkt i mappen eller i undermappar – beroende på vad som passar bäst för det aktuella datasetet.

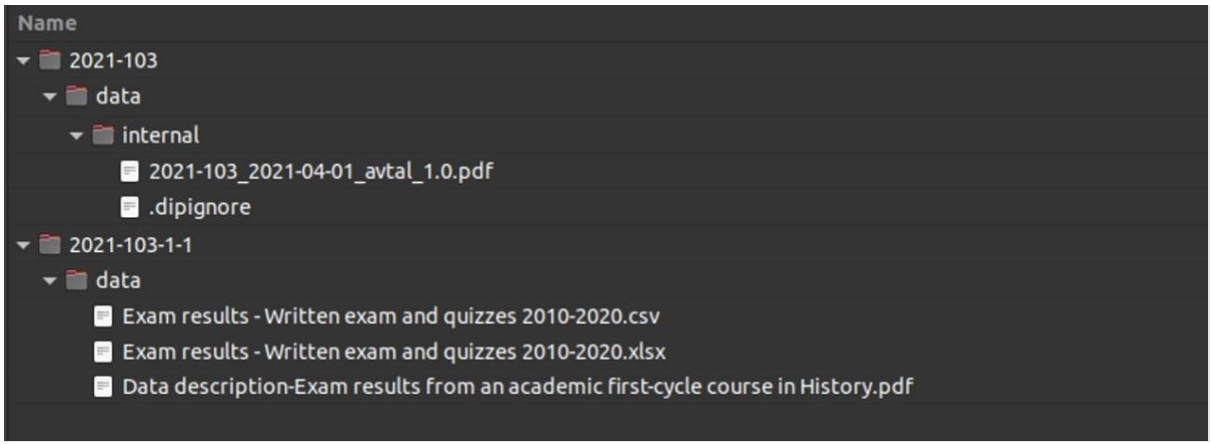
BagIt-filerna läggs direkt i respektive mapp som ”bagas”. ”Bagen” innehåller checksummor för filerna och viss grundläggande metadata (datum, filstorlek, antal filer, identifierare). Checksummor används för att säkerställa integriteten hos en fil, till exempel efter att den har överförts från en lagringsenhet till en annan, eller för att se så att den inte har manipulerats. Att använda BagIt är också bra om man ska lägga över filerna till exempelvis ett arkivsystem.

Filhantering genom arbetsprocessen

Nedan följer en beskrivning med exempel på hur filer hanteras när en databeskrivning granskas hos SND CARE.

Under redigering/inkommande data

För alla påbörjade databeskrivningar i DORIS skapas automatiskt en grundstruktur för filerna enligt mappstrukturen i bilden ovan. Antalet mappar för datasetsversioner kan variera beroende på antalet beskrivna dataset i DORIS.



Bilden visar filstrukturen för en inkommen databeskrivning.

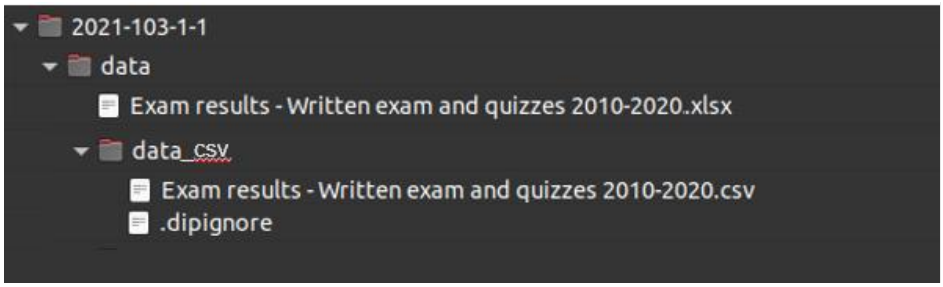
Filer som laddas upp för ett dataset (vilket sker i avsnitt 5 i DORIS:s formulär) läggs automatiskt i mappen som i exemplet ovan heter *2021-103-1-1 > data.*

Under granskning

I granskningen av data- och dokumentationsfiler ingår att se till att filerna finns sparade i format som lämpar sig för långtidslagring och tillgängliggörande. Lämpliga format för långtidslagring är enligt SND format som 1) är vanligt förekommande, 2) kan läsas av många olika datorprogram, 3) är väl dokumenterade och 4) är icke-proprietära eller har öppen källkod. Vilka format som är lämpliga för tillgängliggörande beror på typ av data och forskningsområde. Ofta används samma format som forskaren skickar in data i. Mer information om val av filformat finns bland annat i DAU-handboken[2].

Eftersom filformat för tillgängliggörande och långtidslagring kan vara olika kan det uppstå situationer där filer finns i flera kopior i olika format. Då görs en bedömning om filerna ska delas i båda formaten eller inte. Faktorer som kan väga in är exempelvis antal filer och filstorlek.

[2] DAU-handboken: https://dhb.snd.gu.se/



I exemplet i bilden är xlsx-filen synlig och direkt åtkomlig i katalogposten. Kopian i csv-format syns inte i katalogen men finns för att säkra långsiktigheten. Observera att detta enbart är exempel, att dela både en excel-fil och en csv-fil kan vara att föredra.

För databeskrivningar med data som ska vara fritt tillgängliga så visas/delas alla filer i datasetsmappen (*2021-xx-x-y*) automatiskt i forskningsdatakatalogen. Om några filer inte ska synas i katalogen, exempelvis kopior i andra format, behöver en underliggande mapp för dessa skapas manuellt. Mappen ska då också innehålla en .dipignore-fil.

Det finns fall där SND-kontoret bistår forskaren med viss bearbetning av de datafiler som skickats in. Om det är betydande ändringar som görs sparas originalfilerna så att man kan gå tillbaka och se vad som ändrats. Sparade originalfiler placeras i en egen mapp, med en .dipignore-fil.

OBS! Det är viktigt att arbetet med datafiler är färdigt innan man går vidare med publicering. Om något behöver ändras efter publicering medför det att en ny version av datasetet måste publiceras.

Publicering

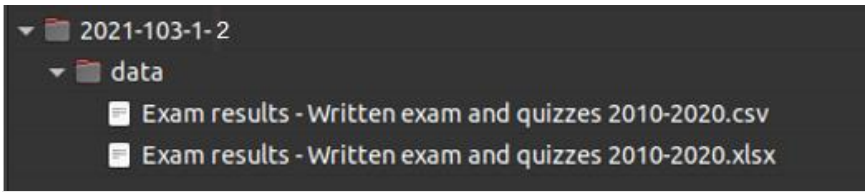
När granskningen av en databeskrivning är klar publiceras katalogposten manuellt i DORIS och blir då sökbar i SND:s forskningsdatakatalog. Följande sker när en databeskrivning publiceras:

- Hela filpaket flyttas (automatiskt) från *Bearbetning* till *Staging.*

- Att filpaket ligger i *staging*-mappen visar att filerna är redo för långtidslagring. Genom en manuell process flyttas filerna till en permanent lagring. På SND-kontoret är det endast ett fåtal forskningsdatarådgivare med särskild behörighet som gör detta.
- BagIt skapas för varje filpaket. En viss fördröjning kan förekomma som man kan behöva ta hänsyn till när filer ska flyttas till den permanenta lagringen.
- De filer som ska vara direkt åtkomliga via katalogposten kopieras automatiskt till mappen *Distribution*.

Uppdatering av publicerad databeskrivning

När en publicerad databeskrivning sätts under redigering för att uppdateras skapas en ny tom mappstruktur med databeskrivningens ID under *Bearbetning*. Nya filer som laddas upp i DORIS läggs i nya mappar med nytt versionsnummer, ex. 2021-103-1-2.



Exempel på filstruktur när en ny version av data laddats upp i DORIS.

Granskning av uppdaterad databeskrivning

Filhanteringen för granskningsprocessen vad gäller uppdateringar av publicerade databeskrivningar genomförs på samma sätt som för nya databeskrivningar.

OBS! Det är även här viktigt att filhanteringen är klar innan publicering.

Publicering av uppdaterad databeskrivning

Vid publicering av en uppdaterad post sker samma process som vid publicering av nya data-beskrivningar; filpaket flyttas automatiskt från *Bearbetning* till *Staging*, BagIt skapas och filer kopieras till *Distribution*.

Eftersom filer redan förekommer både på *Staging* och *Distribution* gäller följande:

Distribution

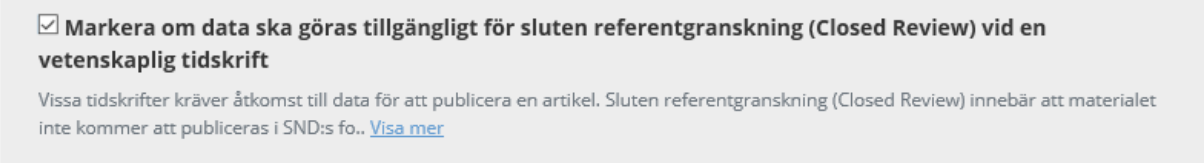
- Om en ny fil med likadant namn som en redan befintlig fil läggs till, ersätts den befintliga filen automatiskt.
- Om en ny fil med ett annat namn läggs till, adderas den till den/de fil/-er som redan ligger i mappen. Om den/de tidigare filen/-erna inte ska synas i katalogen så behövs en manuell hantering (detta bör göras av IT så att det blir rätt i katalogen).

Staging

- Ny fil med likadant namn som en befintlig fil läggs till och får en ny ändelse på filnamnet, exempelvis “(1)”. BagIt-informationen uppdateras.
- För överföring till den permanenta lagringen behöver hänsyn tas till att en ny ”bag” har skapats, vilket kräver manuell hantering och kontroll.

Filhantering för databeskrivningar under sluten referensgranskning (*closed review)*

Om forskaren har markerat att data ska göras tillgängliga för sluten referensgranskning (closed review) hanteras filerna som vanligt, men utan att katalogposten publiceras. Under granskningen går det dock att se filerna genom en förhandsvisning av katalogposten. Datafiler som tillhör databeskrivningar med tillgänglighetsnivån *begränsad åtkomst* visas däremot inte upp i förhandsvisningen.



Om data ska göras tillgängliga för sluten referensgranskning (closed review) kan forskaren markera detta i formuläret i DORIS.

BagIt

BagIt är en metod för att paketera data med grundläggande metadata och checksummor för att kunna verifiera datafilerna. Specifikationen för BagIt finns publicerad här: https://tools.ietf.org/html/rfc8493

BagIt är helt och hållet baserat på filer och håller ingen information i externa databaser vilket gör backup och överföring mindre komplext. Bagger är ett populärt verktyg för att skapa och verifiera “bags” via ett grafiskt gränssnitt. Bagger hittar du här: https://github.com/LibraryOfCongress/bagger

Struktur på en bag för en datasetsversion

SND CARE paketerar varje datasetsversion som en “bag”. En bag har en katalog för data där samtliga filer som hör till datasetsversionen ligger. Underliggande kataloger rekommenderas för mer komplexa dataset som kräver uppdelning och gruppering av filerna. Dokumentation som beskriver data exempelvis kodböcker, beskrivningar, frågeformulär och readme-filer kan också läggas till här.

Bilden på nästa sida visar en struktur på en bag för en datasetsversion. I bilden är ”2021-42-1-1” namnet på katalogen för datasetversionen, där ”2021-42” står för databeskrivningens identifierare och ”1-1” visar att det är första datasetet i version 1.

Alla data ligger i katalogen ”data” med valfri struktur med underkataloger. Filen ”manifest-md5.txt” innehåller md5-checksumma och relativ sökväg till varje fil i "data"-katalogen. Det är också möjligt att använda andra checksummor som till exempel sha-256.

Filen ”bagit.txt” beskriver vilken version av BagIt som används och vilken teckenkodning som används i ”bag-info.txt”. Filen ”bag-info.txt” innehåller i sin tur grundläggande metadata om själva datasetet i form av nyckel: värde. Här finns bland annat DOI med.

```
2021-42-1-1/
  |-- data
  |    |-- collection
  |    |    \-- images
  |    |         \-- img-1.png
  |    |-- data.csv
  |    |-- introduction.pdf
  |    |-- readme.txt
  |
  |-- manifest-md5.txt
  |      ae3b31e65c50ba30ef203d046026d5b4 data/collection/images/img-1.png
  |      93dfcaf3d923ec47edb8580667473987 data/data.csv
  |      a3dcb4d229de6fde0db5686dee47145d data/introduction.pdf
  |      b5fd664ba7c28a798365d454d651b5ee data/readme.txt
  |-- bagit.txt
  |      BagIt-Version: 0.97
  |      Tag-File-Character-Encoding: UTF-8
  \-- bag-info.txt
         External-Identifier: https://doi.org/10.1000/182
         Bag-Group-Identifier: 2021-42
         Source-Organization: Example organisation
         External-Description: Example description
```

Bilden visar en struktur på en ”bag” för en version av ett dataset.