Bilaga 1

Guide till god praxis för datahantering

Den här guiden är till för att:

- Presentera användbara datastrategier inom Heritage Science och föreslå lämplig projekt- och arbetsflödesdokumentation.
- Assistera kort- och långsiktig planering: identifiera praxis som är avgörande för bevarande, tillgängliggörande och återanvändning av data enligt FAIR-principerna.

Innehåll

Projektets livscykel

Planera för skapandet av digitala data

Tänk igenom följande punkter i projektets inledningsskede:

- Vilken typ av output kommer projektet troligen att generera? Inte bara dataset, utan även fotografier, tabeller och diagram är forskningsdata om de används för tolkning och resultat.
- Vilka filformat bör du lagra och dela data i för att säkerställa långsiktig användbarhet?
- Gör en bedömning av vilken typ och nivå av löpande dokumentation över output och beskrivande metadata som kommer att behövas.

Projektdokumentation

- Bestäm var projektets dokument och data ska lagras, sprid det inte över fler platser än nödvändigt.
- Ha en tabell eller databas med uppgifter om projektet där output registreras efterhand. Det är ett bra underlag vid tillgängliggörande. Sparar tid vid sammanställning av det som ska laddas upp och hjälper tredje part att förstå din data (se *Tillgängliga data* nedan).

Namnge och organisera filer

Att följa god praxis för att namnge och organisera sina filer gör det mycket lättare att hitta rätt data, inte bara för dig utan, för dina samarbetspartners och senare även för andra som vill återanvända din data. Följ gärna dessa rekommendationer:

- Var konsekvent när du namnger filer inom ett projekt. Särskilt de som skapats med samma metod och/eller för samma analysobjekt.
- Filnamn ska vara deskriptiva. Undvik generiska namn som ”Test 1”, ”Bild 1” osv.
- Använd understreck (första_studien_projektnr), bindestreck (första-studien-projektnr) eller kamelnotation (FörstaStudienProjektnr).
- Undvik specialtecken i filnamnet som \ / ? : * ” > < | : # % ” { } | ^ [ ] ` ~ æÆ øØ åÅ äÄ öÖ …
- Filnamn på data som exporterats till ett öppet format ska vara identiskt med namnet på original-filen, så att den går att identifiera.

God praktik för dataset (databas, kalkylfil)

- Första raden är rubriker, med variablernas namn
- Variabelnamnen i rubrikerna får inte innehålla specialtecken, mellanslag eller börja med en siffra. Vid behov, använd understreck, bindestreck eller kamelnotation (se ovan).
- En kolumn = 1 variabel
- En rad = 1 observation/prov
- En cell = 1 värde
- Använd så långt det går standardiserade termer från kontrollerade vokabulärer för både variabler och observationer, så att din data blir förståelig för andra användare.
- Använd inte specialtecken (se ovan) i celler med observationsdata.
- Det går bra att ha celler som innehåller text med specialtecken, t.ex. i kolumner som förklarar och kommenterar observationer, men skilj på celler med observationsdata respektive fritext.
- Använd internationellt datumformat: YYYY-MM-DD (till exempel 2022-07-13)
- Skapa en ReadMe.txt fil som förklarar eventuella förkortningar eller mindre kända termer, anger måttenheter, samt om det finns en kontrollerad vokabulär med definition.

Råd och tips för FAIR data

Sökbara data (Findable)

För att andra ska kunna hitta din data så är det av stor vikt att du beskriver den på ett konsekvent och begripligt sätt. Det är bra att göra det på flera sätt: dels med hjälp av den metadata som används för att beskriva filerna i samband med uppladdning på en plattform, dels i en fil som dokumenterar projektets output (se Projektdokumentation ovan).

Beskrivande metadata säkerställer att din data går att hitta av både människor och maskiner. Därför är det bra att använda standardiserade ämnesord från kontrollerade vokabulärer och i bästa fall kombinera dessa med beständiga länkar till definitioner (så kallade auktoriteter).

Det går bra att blanda termer och länkar från olika vokabulärer.

Använd tabellerna i dokumenten "Kontrollerade vokabulärer för Heritage Science" och "Auktoriteter för Kulturarvslaboratoriet" som stöd.

Följande uppgifter bör förekomma i beskrivande metadata:

- Metod/instrument
- Material
- Ämnesområden (t.ex. Heritage Science, Art History etc)
- Organisation (ansvarig, medverkande – t.ex. samlingsförvaltaren)
- Projekt ID (namn och/eller projekt/diarienummer)

Andra vanligt förekommande uppgifter inom Heritage Science:

- Objekt typ (t.ex. mynt, möbel, svärd)
- Objekt ID (identifierare för föremål, byggnad eller lämning)
- Geografi (land, region och/eller plats)
- Kulturell/historisk kontext (tidsperiod, stil, kultur)
- Personer (associerade med materialet – t.ex. skapare, avbildad, omnämnd)

Upphovspersoner bör helst också ha globalt unik identifierare, så det blir enkelt att hitta mer som de publicerat. Genom att registrera dig på ORCID får du en personlig unik identifierare: https://orcid.org/

> **Zenodo-tips**: *Keywords* är fritext, där kan du skriva in vilka ord som helst som du tror att någon intresserad av just denna data skulle kunna söka på. *Subjects* är termer som har länkar till en auktoritetspost i en kontrollerad vokabulär. Där kan du även ange unika identifierare för de objekt som analyserats.
>
> Under *Related Identifiers* kan du lägga till länkar till publikationer eller andra uppladdningar som hör ihop med de uppladdade filerna, och förklara hur de relaterar till varandra. Det hjälper både sökmotorer och mänskliga användare.

Tillgängliga data (Accessible)

När ett projekt avslutas ska relevanta filer bevaras på ett enhetligt sätt, så att det lätt går att hitta dem oavsett vilka som producerat den och om de finns kvar i verksamheten eller ej. Beroende på verksamhet kan detta handla om en server med backup, ett e-arkiv eller ett digitalt repositorium. Filer och eventuella mappar ska namnges på enhetligt och begripligt sätt (se ovan).

Det är viktigt att tänka igenom vilka data som ska sparas och vilka av dessa som ska tillgängliggöras så tidigt som möjligt i projektet. Allt som bör bevaras behöver inte tillgängliggöras, men tillgängligheten bör inte begränsas i onödan. Nya metoder och verktyg kan möjliggöra för framtida forskare att bättre analysera redan existerande data. Därför kan det också vara viktigt att spara och dela rådata som inte har genomgått ändringar.

En kompromiss kan vara att inte ladda upp all bevarad data på en plattform, men att däremot ladda upp en tabell med projektdokumentation som redovisar vilken typ av output som projektet skapat och vad filerna heter. Med bra metadata blir det en form av tillgängliggörande. Data som fungerat som direkt underlag till publicerade tolkningar och resultat bör tillgängliggöras så öppet som möjligt. Principen är ”så öppet som möjligt” – det ska med andra ord finnas tydliga motiv för att *inte* tillgängliggöra.

En arkiveringsstrategi bör inkludera att under projektplanerings- och datainsamlingsstegen identifiera lämpliga filformat för både bevarande och återanvändning. Adekvat dokumentation och metadata ska skapas för att underlätta leveranser och stötta administration under projektets gång.

Forskningsdata behöver alltså inte göras öppet tillgänglig omedelbart. Det viktiga är att publicera metadata som beskriver innehållet, där det framgår vilka eventuella begränsningar som finns för åtkomst t.ex. embargo fram till en viss tidpunkt, eller om man behöver kontakta producenten. De allra flesta digitala repositorier har sätt att begränsa åtkomsten vid behov.

Sätt gärna i system att ladda upp filer löpande under en projektets gång, med eller utan embargo. På så vis behöver inte allt samlas ihop och laddas upp när projektet avslutas. Fördelen med detta är dessutom att både projektdeltagare och peer-reviewers vid behov kan få åtkomst till data på ett effektivt sätt.

> **Zenodo-tips**: En sak som kan vara bra att tänka på är att det är svårt att radera filer från Zenodo (eller repositorier generellt). Det går utmärkt att ladda upp en ny *version* och att ändra beskrivningar, men inte att få något bortplockat. Detta är för att säkerställa att ingen hänvisar till data i publikationer, som sedan tas bort. Du kan dock förbereda en uppladdning Zenodo och vänta med själva publicerandet till dess du är säker på att det är den slutgiltiga versionen.

Kompatibla data (Interoperable)

Det bästa sättet att säkerställa att data blir interoperabel, att det går att kombinera data från olika källor, är genom användning av standardiserade termer och stavningar i dina dataset, vilka går att hitta i kontrollerade vokabulärer. Auktoriteter (globalt unika identifierare) kan också inkluderas för att hänvisa till viktiga källor. Du ska även använda dig av god datapraxis generellt. Se ”god praktik för dataset” ovan för mer detaljerad handledning.

Det går bra att blanda termer och länkar från olika vokabulärer. Använd tabellerna i dokumenten “Kontrollerade vokabulärer för Heritage Science” och “Auktoriteter för Kulturarvslaboratoriet” som stöd.

Återanvändbara data (Reusable)

Öppna format

Att välja ett öppet filformat är viktigt för att försäkra att din data kommer att vara läsbar även i framtiden. Vissa filformat är mer optimala än andra när det gäller att möjliggöra långsiktig läsbarhet:

- Icke-proprietära (formatet ägs inte av ett företag)
- Öppen källkod, med dokumenterad internationell standard
- Använder standardiserad teckenkodning, företrädelsevis Unicode (till exempel UTF-8)
- Icke komprimerade, filen innehåller all data och den har inte komprimerats.

Library of Congress har en omfattande guide för rekommenderade format:
https://www.loc.gov/preservation/resources/rfs/

Svensk Nationell Datatjänst har också en händig guide:
https://snd.gu.se/sv/hantera-data/guider/att-valja-filformat

En del instrument som används inom Heritage Science saknar lämpligt öppet filformat där all information bevaras. I dessa fall föreslås att tillgängliggöra dels originalfilerna (även om formatet inte motsvarar ovanstående kriterier), dels data exporterat till öppna filformat så att så mycket som möjligt av informationen ska vara tillgänglig långsiktigt. I de fall där det enkelt går att omvandla data till ett rekommenderat filformat och få med all information är det självklart att föredra vid publicering.

Licenser

För att data verkligen ska vara användbar måste det finnas uppgifter om eventuella villkor och begränsningar i samband återanvändning. Detta ska göras med beskrivande metadata som följer standarder, så att det är både maskin- och människoläsbart. Creative Commons-licenserna är internationellt etablerade och rekommenderas i de flesta fall: https://creativecommons.org/

CC-BY är den vanligaste licensen för forskningsresultat. Det betyder att det är fritt för andra att använda informationen på olika sätt, till exempel att kombinera data från flera olika undersökningar eller att använda sig av bilder på olika sätt, så länge som de *citerar källan* i samband med publicering.

Mer restriktiva licenser som CC-BY-NC (*non-commercial*) eller -ND (*non-derivative*) bör *inte* användas för forskningsdata. Publicering i vetenskapliga tidskrifter eller för förlag kan räknas som kommersiell användning, eller till och med presentation inför en publik. *Non-derivative* betyder att inget får ändras, vilket gör det omöjligt att kombinera data från olika källor eller ens att publicera del av en bild.

Den som oroar sig för missbruk av sina data ska komma ihåg att det finns regelverk för god etik och praxis inom forskning. Att du då har publicerat din originaldata är det bästa beviset du har om du utsätts för ohederlig återanvändning eller stöld av data. Du får också ett större genomslag om andra kan använda sig av dina resultat och *citera* dig.