SND
Svensk nationell datatjänst

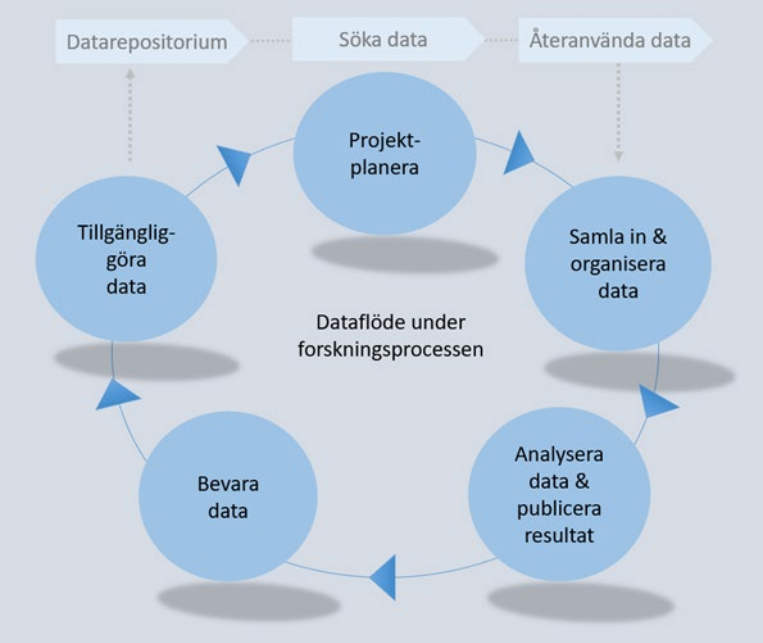
Principer för dokumentation

Pass 5: Dokumentation av forskningsdata

BAS Online 2021-01-20

Den första delen i pass fem fokuserar på övergripande principer för dokumentation under forskningsprocessen.

Ett begrepp som är vanligt att använda när man pratar om datahantering eller data management är *data life cycle*, som på svenska kan benämnas som datalivscykeln. Under pass 4, som handlade om metadatastandarder, presenterades DDI-alliansens modell av datalivscykeln. Medan den modellen utgick från dataflödet och visade de olika steg som data befinner sig i, har modellen i bilden nedan istället sin utgångspunkt i forsknings-processen. Modellen synliggör hur datamaterialet används och hanteras under forskningsprocessens olika faser och hur data som tillgängliggörs kan återanvändas i sekundärforskning. I verklig-heten är forskningen naturligtvis mer detaljerad och komplex än den modell som visas. I samtal med forskare kan dock en bra utgångspunkt vara att reflektera kring vad som är relevant att dokumentera under de olika faserna av forskningsprocessen. Låt oss nu titta lite närmare på modellens olika delar.



Under projektplanering skriver forskaren ansökan om forskningsmedel och förbereder sådant som måste vara klart inför projektets start. Det kan exempelvis innefatta planering om forskningsprojektets upplägg och de frågeställningar som ska besvaras, vilka metoder som ska användas, hur data ska samlas in och hanteras.

När projektets finansiering är klar kan processen fortskrida och planeringen kan gå vidare mer detaljerat. Data som senare ska analyseras samlas in och organiseras på olika sätt. T.ex. kan datamängden behöva rensas och struktureras i en databas.

Under analysfasen arbetar forskaren aktivt med att analysera, tolka och publicera resultat. Det är inte ovanligt att flera forskare arbetar i ett forskningsprojekt och studerar olika delmängder av datamaterialet. Rapporter sammanställs och resultat publiceras på olika sätt, såsom i vetenskapliga tidskrifter.

När analyserna är klara och resultaten är publicerade ska sedan det forskningsmaterial som tagits fram under processen förberedas för långtidsbevaring och tillgängliggörande. Här ingår

bland annat att se över vilka filformat som är lämpliga att använda för långtidsbevaring. I enlighet med arkivlagen ska forskningsmaterial arkiveras vid det egna lärosätet. Allmänna handlingar vid universitet och högskolor ska som huvudregel bevaras och det är i princip förbjudet att förstöra allmänna handlingar om det inte föreligger lagstöd för det, det vill säga att man har rätt att gallra. Om data ska lämnas till ett datarepositorium behöver också data och tillhörande dokumentation göras iordning.

Tillgängliggörande av data sker oftast i slutet av ett forskningsprojekt men kan även inträffa tidigare, t.ex. i samband med att resultat publiceras i någon tidskrift som har krav på öppen tillgång till data. Kraven ökar allt mer på att offentligt finansierad data ska göras tillgängliga för andra att forska vidare på. Generellt kan man säga att data bör vara så öppna som möjligt och så stängda som nödvändigt, då det t.ex. kan finnas juridiska skäl som begränsar tillgängligheten. I samband med att data tillgängliggörs är det viktigt att all dokumentation som är nödvändig för att återanvända materialet också följer med. *Vad* som är relevant att dokumentera skiljer sig mellan olika forskningsprojekt och olika ämnesområden, men gemensamt för alla är att det kommer finnas saker att dokumentera under varje fas i processen.

Att systematiskt dokumentera sitt forskningsmaterial är väsentligt för att data senare ska kunna publiceras, vara sökbart, användas av andra och citeras korrekt. All den information som behövs i samband med tillgängliggörande kan vara svår att återskapa i efterhand och det är därför viktigt att redan från start dokumentera materialet väl. Finansiärer som har krav på att data ska göras tillgängliga vill också säkerställa att materialet dokumenteras tillräckligt, vilket är en av anledningarna till att de kan kräva en datahanteringsplan som bland annat innefattar beskrivning om hur data ska dokumenteras.

Det finns flera skäl till att dokumentera väl under forskningsprocessen. För forskaren själv är det ett sätt att skapa ordning och reda. Att minnas detaljer kring hur data har kodats, vad olika förkortningar betyder, eller komma ihåg vad som skiljer olika versioner av data från varandra är inte särskilt hållbart. Ett väldokumenterat forskningsprojekt bidrar till ökad kontroll och forskaren har tillgång till den information som behövs, när den behövs. Dokumentation är inte bara viktigt för forskaren själv utan även för att andra ska förstå vad som har gjorts, t.ex. vid granskning av publikation, eller för att kunna verifiera eller replikera resultat. För att möjliggöra återanvändning av data är dokumentation om forskningsprojektet och dess data en förutsättning. Det finns även lagstiftningar som ställer krav på dokumentation, till exempel offentlighets- och sekretesslagen, dataskyddsförordningen och etikprövningslagen. Dessutom kan det finnas riktlinjer och regler om dokumentation från finansiären eller lärosätet.

En väldokumenterad datasamling innefattar information som syftar till att vara fullständig och självförklarande för framtida användare. Om man redan från början tänker på att datamaterialet ska kunna förstås och användas av andra så ökar förutsättningen för att det kommer dokumenteras tillräckligt väl under processens gång.

Att komma igång med dokumentationen

Att dokumentera det som sker under forskningens olika faser kan kännas tidskrävande och det kan därför vara bra med några råd om hur man enklare kommer igång med arbetet. De här råden kan ses som exempel på sådant som kan lyftas fram i kontakt med forskare:

- Ett gott råd är att börja dokumentera så tidigt som möjligt och sedan göra det konsekvent under projektets gång. Skapa gärna rutiner som underlättar och förenklar arbetet.

- Fundera över vilken information som andra behöver för att förstå datamaterialet. Det som kan verka självklart för en själv är inte alltid det för andra.
- Utöver den eller de filer där datamaterialet finns organiserat är det lämpligt att skapa en separat fil som innehåller information om datamaterialet. För varje datafil som skapas kan man t.ex. skapa en kodbok eller variabellista som beskriver de variabler eller enheter som ingår i filen. Om datamängden består av många filer, t.ex. bilder, bör det framgå om och hur dessa är relaterade till varandra.
- Om data planeras att göras tillgängliga via ett datarepositorium eller någon liknande service kan man tidigt i processen kontakta dem för att få råd om dokumentation och metadata.

Dokumentation på olika nivåer

Dokumentation om forskningsprojektet och dess data sker på olika nivåer. Dokumentation på *projekt- eller studienivå* handlar om information som är av mer övergripande karaktär för projektet såsom frågeställningar, syfte, metoder och urvalsprocess.

På *fil-dataset-databasnivå* beskrivs t.ex. information om datafilen, hur flera filer förhåller sig till varandra, vilket format de har, eller vilken version som en specifik fil har och vad som skiljer den från andra versioner.

Det som anges på *variabel/objektsnivå* är beskrivningar av de olika variabler eller objekt som data består av. Namn på de variabler som ingår i ett dataset är till exempel inte tillräcklig information för att andra ska kunna förstå och återanvända datamaterialet. Om data är insamlade via frågeformulär är frågeformuleringen också viktig att ha med för respektive variabel. Om data istället innefattar mätvärden insamlande via fysiska mätningar under en hälsokontroll kan uppgifter som tid, plats eller särskilda instrument som använts vara relevanta. Generellt kan man säga att ju mer detaljerad nivå som

metadata dokumenteras på desto mer kommer metadata att skilja sig mellan olika ämnesområden och typ av data.

Exempel på dokumentation

För att konkretisera vad dokumentation kan innebära under forskningsprocessen kommer du att få se några exempel på vad som kan behöva dokumenteras under de olika faserna, och olika dokument/verktyg som forskare kan använda som stöd för att strukturera dokumentationen. Observera att det inte är en fullständig förteckning som passar alla typer av datamaterial eller ämnesområden.

Innan du får se exempel får du möjlighet att själv reflektera. Under varje fas uppmanas du att pausa presentationen för att skriva ner 3 saker som är relevanta att dokumentera, och gärna föreslå något exempel på dokument eller verktyg som kan användas som stöd. När du skrivit ned det här kan du starta presentationen igen. Då kommer du att få se några exempel som vi sammanställt. Därefter går du vidare till nästa fas. Vi börjar med den första fasen som är projektplanering. Pausa presentationen och skriv ner 3 saker som är relevanta att dokumentera, och föreslå gärna något dokument eller verktyg som kan användas som stöd.

I samband med planering och uppstart av forskningsprojektet bör en mängd saker dokumenteras. Information om forskningsprojektet kan exempelvis inbegripa frågeställningar, syfte, metoder, plan för datainsamling och hantering av data. Ansvarsområden, beslut som fattas och andra typer av löpande händelser är också sådant som är relevant att dokumentera. Hur forskaren strukturerar all den information som produceras skiljer väldigt mycket, men oftast skriver forskaren i den här fasen en forskningsbeskrivning (kan även kallas forskningsplan eller protokoll) som innefattar många av de nämnda delarna. Andra exempel på dokument/verktyg som kan

användas för att dokumentera är till exempel datahanteringsplan eller loggbok.

Pausa nu presentationen och skriv ner 3 saker som är relevanta att dokumentera i fasen när data samlas in och organiseras, och ge gärna förslag på vad för dokument som kan användas som stöd.

Under den här fasen sker oftast många saker som behöver dokumenteras löpande. Det kan vara anteckningar som relaterar till datainsamlingen, men även beslut som fattas och kontakter som tas. När data väl samlats in är dessa viktiga att beskriva, också med hänseende till hur data har rensats och bearbetats. Vidare är det relevant att beskriva metoder såsom urvalsmetod, insamlings-metod, och eventuella verktyg eller instrument som används. När forskningsprojektet kommit igång brukar rutiner och principer för datahantering att arbetas fram, vilka bör skrivas ner. Det kan handla om principer för versionering, mappstruktur och filnamn etc. Exempel på dokument/verktyg för sådan dokumentation är loggbok, fältanteckningar, kodbok/variabellista, statistikfil, teknisk rapport, metodbeskrivning, datahanteringsplan.

Nu går vi vidare till nästa fas, som handlar om att analysera data och publicera resultat. Pausa presentationen och skriv ner 3 saker som är relevanta att dokumentera och ge gärna förslag på vad som kan användas för dokument som stöd.

Några exempel på sådant som bör dokumenteras under den här fasen är en överblick över analysarbetet, såsom frågeställningar, medförfattare, beskrivning av de data som används för analys, vilka analyser som utförs, hur olika filer hänger ihop, och eventuellt komplettering eller korrigering av olika aspekter kring datahantering. Exempel på dokument/verktyg för sådan dokumentation är analysplan, kodbok, variabellista analysloggbok, program-

kodfiler/syntaxer/loggfil, README-fil, databasschema, datahanteringsplan.

Pausa nu presentationen och skriv ner 3 saker som är relevanta att dokumentera under den fjärde fasen, bevara och förbereda data, och ge gärna förslag på vad som kan användas för dokument som stöd.

Här är det exempelvis relevant att skapa en lista över datafiler och dokumentation som ska långtidsbevaras, både sådant som ska lämnas för arkivering vid lärosätet men även om material ska lämnas till ett datarepositorium. Det är också viktigt att se över huruvida några åtgärder behöver göras med datamaterialet innan det ska tillgängliggöras, och dokumentera de beslut som fattas. Som stöd för att dokumentera kan man använda sig av någon slags loggbok för att ange löpande anteckningar och komplettera datahanteringsplanen med aspekter kring bevarande och tillgängliggörandet. Om det finns något dokumentationsverktyg som passar för datatypen kan det användas för att föra samman olika metadata inför bevarande och tillgängliggörande.

Till sist kan du nu reflektera över vad som är relevant att dokumentera i samband med att data ska tillgängliggöras. Pausa presentationen och skriv ner 3 saker som är relevanta att dokumentera och ge gärna förslag på vad som kan användas för dokument som stöd.

Här är det framför allt beslut och rutiner kring tillgängliggörande som är relevant att dokumentera, med hänseende till vad som ska tillgängliggöras, var, när det skall ske, och vilken åtkomst materialet ska ha. Finns t.ex. restriktioner, embargo, åtkomstvillkor eller licens så behöver det skrivas ner. Som stöd för att dokumentera kan man även här använda en loggbok, samt fylla på datahanteringslanen. Som framgår av bilden kan datahanteringsplanen användas kontinuerligt under hela forskningsprocessen.

Vilken dokumentation som är relevant för en sekundäranvändare av datamaterialet varierar mellan olika projekt. En del av det som produceras kommer inte vara relevant medan andra delar är helt väsentliga. En forskningsstödsenhet, såsom DAU, kommer att fylla en viktig funktion gentemot forskarna, dels för att ge råd och stöd gällande dokumentation under processen, dels för att bedöma om den dokumentation som forskare lämnar tillsammans med data är tillräckliga för sekundäranvändning. Att göra en sådan bedömning är inte helt enkelt och kan till exempel kräva viss ämneskunskap och kännedom om olika datatyper. Lite längre fram i pass 5 presenteras några grundprinciper som kan användas som en utgångspunkt för att bedöma om metadata är tillräckliga.

Nu kommer en övning där du, utifrån ett fiktivt forskningsprojekt, får möjlighet att reflektera kring dokumentation och metadata.

Fallbeskrivningen är följande:

> *Forskningsprojektet kommer att samla in data via enkäter och fysiska mätningar i samband med hälso-undersökning. Projektet har ännu inte startat insamlingen. Forskargruppen har påbörjat en data-hanteringsplan men har lite funderingar kring delen som avser dokumentation och metadata.*
>
> *De vill gärna ha lite råd om vad de bör tänka på i samband med dokumentation under insamlings- och analysfasen*

Pausa presentationen och försök att komma på minst tre råd som du kan ge till forskargruppen avseende dokumentation. När du är klar kan du starta presentationen igen.

Här kommer några exempel på råd som kan ges till forskarna:

- Dokumentera det som sker under datainsamlingen med hänseende till vad som gjorts, hur det genomförts, vilka

tidpunkter, och av vem. För data som samlas in via fysiska mätningar är det t.ex. relevant att beskriva vilka mätmetoder som använts och eventuella instrument.

- När data sedan är insamlade behöver forskarna beskriva om data har rensats och bearbetats på något sätt. Vad har gjorts, hur har man gått tillväga, när och av vem. Ju utförligare information desto bättre.
- Under analysfasen skapas ofta många olika versioner av datafiler och det är därför viktigt att dokumentera vilka datafiler som använts till vilka analyser.
- Klargör vilka rutiner som projektets medlemmar ska följa med hänseende till mappstruktur, filnamn och versionering.
- I en datahanteringsplan kan forskargruppen skriva ner rutiner kring dokumentation så att det blir tydligt för alla projektets medlemmar vad som gäller. De kan exempelvis skriva ner vilka dokument eller verktyg som används för att dokumentera olika saker och vilka principer som man kommit överens om avseende mappstruktur, versionering och filnamn.

Sammanfattning

Sammanfattningsvis vill jag lyfta upp några centrala delar från presentationen. Varje forskningsprojekt är unikt, och vilken dokumentation som är viktig att producera är beroende av projektets specifika förutsättningar. En bra utgångspunkt för att stödja forskare avseende dokumentation är att utgå från de olika steg som forskningsprocessen består av. Några viktiga råd till forskare är att börja dokumentera så tidigt som möjligt i processen och skapa rutiner som underlättar arbetet, samt att fundera på vilken information som andra behöver för att kunna förstå och återanvända datamaterialet.