Pass 5: Dokumentation av forskningsdata

Bedömning av metadata för sekundäranalys

Välkommen till en presentation om bedömning av metadata för sekundäranalys. Jag kommer att ta upp några principer som kan användas som stöd vid en sådan bedömning.

Under den första delen i BAS Online nämndes FAIR-principerna, vilka i korthet innebär att forskningsdata ska vara möjliga att hitta, att de ska vara tillgängliga och åtkomliga samt möjliga att återanvända. En central uppgift för en DAU är att hjälpa forskare att göra sina data så FAIR som möjligt. Det innebär bland annat att se till att det finns tillräckliga metadata vid publicering av data så att andra kan förstå och återanvända datamaterialet.

Under presentationen kommer du att få ta del av några grundläggande principer för bedömning av metadata. Dessa är inte helt beroende av ett arbetsflöde men utgångspunkten är att metadata som ska publiceras i SND:s nationella metadatakatalog i första hand inkommer via SND:s formulär för att beskriva och lämna in data. SND har i uppgift att arbeta fram fler arbetsflöden som mer detaljerat beskriver riktlinjer för DAU:s arbete med att granska metadata. Vissa lärosätesspecifika rutiner och arbetssätt kan emellertid också komma att påverka det lokala arbetet.

Metadata för sökbarhet

När metadata publiceras i den nationella metadatakatalogen blir informationen synlig och sökbar för andra. För en återanvändare är det viktigt att kunna hitta resursen och sedan, med hjälp av informationen, kunna bedöma om datamaterialet är intressant eller ej. De metadata som beskriver forskningsstudien eller projektet behöver vara relevanta och tillräckligt omfattande. Det behöver till exempel finnas beskrivet vad som har gjorts, hur det har genomförts, mellan vilka tidsperioder och av vem eller vilka. Det innefattar allt från uppgift om huvudman och ansvariga forskare till beskrivning av urvalsmetod, datainsamling och geografisk information. De metadata som forskaren angett i SND:s formulär för att beskriva data kan granskas med hänseende till datamaterialets sökbarhet. En bra utgångspunkt är att

undersöka om informationen i formuläret kan besvara frågorna vad, var, när, hur, och av vem. Förutom att kontrollera vilka fält som är ifyllda är det också relevant att se huruvida det finns uppenbara stavfel, om uppladdade dokument går att öppna och om deras innehåll är läsbart. Om det finns information i bifogade dokument som saknas i formuläret kan man helt enkelt komplettera. Många forskningsprojekt producerar ofta dokument som är relevanta vid återanvändning av datamaterialet, som t.ex. innehåller beskrivning av projektet, dess urvalsmetod, och annan information om data-insamling och datamaterialet. Alla sådana dokument eller rapporter är viktiga att bevara tillsammans med materialet och bör skickas med när formuläret fylls i. Eftersom metadata som läggs in i formuläret blir synliga och sökbara finns det en poäng med att fylla i så mycket information som möjligt. Utgångspunkten bör vara att forskaren är den som fyller i formuläret, men att en DAU kan identifiera vilka delar som behöver kompletteras eller korrigeras.

Metadata begripligt organiserade

En andra sak att ta reda på är hur datamaterialet är organiserat och om det finns dokumentation som beskriver datamaterialets struktur. Datamaterialet kan till exempel vara organiserat i en enskild fil eller i många filer som är mer eller mindre sammankopplade till varandra. Här är det relevant att bedöma om data är organiserade på ett sådant sätt att det är begripligt för en sekundäranvändare. Ett datamaterial som består av tabulära data som finns samlade i en datafil har inte samma behov av förklaring som ett datamaterial med en mer komplex struktur. Ett omfattande datamaterial kan bestå av hundratals filer med olika typer av data. Hur dessa förhåller sig till varandra behöver finnas beskrivet.

Metadata för återanvändning

För att datamaterialet skall kunna återanvändas räcker det inte med att resursen hittas i en metadatakatalog och att datamaterialet är organiserat på ett begripligt sätt. I det här steget behöver man ta reda på vilka metadata som finns på variabelnivå, dvs. beskrivning av datamaterialets innehåll. Här behöver man bedöma om de variabler, enheter eller objekt som data består

av är tillräckligt väl beskrivna eller om ytterligare metadata behöver kompletteras. Varje datamaterial är unikt på sitt sätt och det är därför svårt att säga vilka metadata som generellt behövs för sekundäranvändning. Alla datamaterial kräver inte lika mycket metadata, och vilka metadata som är relevanta behöver således bedömas för varje enskilt fall.

Vid det här laget känner du antagligen till att forskningsdata kan vara insamlade på olika sätt, bestå av olika typer av filer, och vara organiserade olika. Forskningsdata kan till exempel samlas in med hjälp av intervjuer, frågeformulär, inspelningar och observationer och bestå av texter, numeriska data, filmer, fysiska samlingar osv. Hur datamaterialet sedan är organiserat skiljer sig beroende på vilket typ av data det är.

Data som består av siffror, kategorier eller liknande och som kan kvantifieras brukar ofta struktureras i rader och kolumner. Såsom exemplet på bilden.

	A	B	C	D	E
1	F1_Kon	F2_Halsa	F2_Halsa_diko	P_Glukos	P_Glukos_3gr
2	1	3	1	5,1	1
3	1	2	1	4,7	1
4	2	2	1	4,2	1
5	1	1	1	5,6	1
6	2	4	2	3,1	1
7	2	998	998	4,9	1
8	2	3	1	6,2	2
9	999	3	1	5,5	1
10	2	2	1	6,7	2
11	2	1	1	4,1	1
12	1	1	1	7,4	3
13	998	5	2	12,7	3
14	2	3	1	4,9	1
15	1	3	1	5,1	1
16	1	3	1	3,8	1

Till en sådan datafil finns ofta dokumentation i separata dokument, t.ex. i en variabellista eller ett frågeformulär. Det behöver då framgå vilken variabel, enhet eller vilket objekt som informationen avser, så att data och dokumentation kan matchas mot varandra. Det här kan vara svårt att bedöma om man inte är kunnig i det aktuella ämnesområdet eller om det handlar om filformat som man aldrig varit i kontakt med innan. Det är därför bra att stämma av med forskaren själv som har kunskap om det specifika ämnet och datamaterialet. Det är dock inte säkert att forskaren har tänkt på alla aspekter avseende metadata för sekundäranvändaren och en DAU kan därför ställa relevanta frågor som säkerställer att metadata är tillräckliga.

Ytterligare metadata

Även om datamaterialet har tillräckliga metadata för en sekundäranvändare kan det finnas skäl att fundera kring om det är möjligt att berika formuläret med ytterligare metadata så att datamaterialet blir mer synligt, sökbart och går att jämföra med andra dataresurser. En åtgärd kan t.ex. vara att komplettera beskrivningen med fler nyckelord som är specifika för studien. Att dokumentera data strukturerat med hjälp av ett dokumentationsprogram är en annan åtgärd som t.ex. kan göra så att variabler i ett dataset eller frågor i ett frågeformulär också kan bli sökbara i SND:s nationella metadatakatalog. Det senare är givetvis arbete som är tänkt att utföras av forskaren.

Metadata om mjukvara och filformat

Forskaren bör redan från början ha valt ett filformat som är lämpligt för långtidsbevarande. I det första passet nämndes tre kriterier som indikerar om filformatet med stor sannolikhet kommer att fungera på sikt.

1. Formatet ska vara vanligt förekommande. Ett vanligt format löper nämligen mindre risk att avvecklas.
2. Formatet bör vara leverantörsoberoende, för då är man inte beroende av en viss programvara för att kunna öppna och läsa filen.
3. Slutligen så bör formatet ha en öppen teknisk specifikation, vilket innebär att det inte kontrolleras av en enskild person eller organisation.

Det är dock inte alltid möjligt att välja ett format som uppfyller dessa kriterier, då specifika instrument, analysredskap eller egentillverkad programvara kan påverka valet av dataformat. Det behöver då framgå vad som krävs för att datamaterialet ska kunna användas.

De principer som du nu har fått ta del av är förhållandevis generella, men kan ses som en utgångspunkt i arbetet med att bedöma om ett datamaterial har tillräckliga metadata för publicering i SND:s nationella metadatakatalog. Det finns inte någon strikt ordning på de nämnda principerna, och vad som görs i vilken ordning kan komma att variera. Som också nämnts har SND i uppgift

att arbeta fram rutiner som kan följas mer detaljerat än dessa övergripande principer.

Sammanfattning

En viktig del för att data ska vara FAIR är att det finns tillräckliga metadata för en sekundäranvändare. Förutom metadata som forskaren anger i SND:s formulär kan kompletterande dokument också laddas upp. Ofta produceras olika slags dokument med information som är relevant vid återanvändning. Forskaren har expertkunskap inom sitt område och om det specifika forskningsmaterialet och är således viktig att samarbeta med vid bedömning av metadata.