SND
Svensk nationell datatjänst

Dokumentation på variabelnivå

Pass 5: Dokumentation av forskningsdata

BAS Online 2021-01-20

Nu har vi lämnat övergripande principer kring dokumentation av forskningsmaterial och går över till passets andra del som handlar om dokumentation på variabelnivå.

Dokumentation på variabelnivå handlar om att beskriva de data som samlas in och/eller används för analys i forskningen. Variabler, objekt eller enheter är olika sätt att benämna de fenomen eller specifika delar som data består av och vilket begrepp som används kan skilja mellan olika ämnesområden och typer av data. En forskare som studerar föremål från en arkeologisk utgrävning brukar vanligtvis använda begreppet *objekt* medan forskare som studerar mätdata eller strukturerade data insamlade via exempelvis register eller enkäter använder begreppet *variabler*. En klimatforskare kan istället prata om *parametrar*, som används synonymt med variabler.

Under pass 4 fick du en introduktion till olika metadatastandarder som kan användas för att beskriva forskningsdata. En av dessa är metadatastandarden Data Documentation Initiative, DDI. Den används av SND och många andra datarepositorier, främst för att beskriva tabulära data som är insamlade via frågeformulär. Variabler är en central del av DDI, och elementet beskrivs enligt följande: “This is the applied expression of a data item within a data set.”

En variabel är alltså en del i en datafil och mäter något specifikt som efterfrågats i frågeformuläret, som t.ex. ”Antal koppar kaffe om dagen”. Till varje kolumn finns sedan respondentens svar från frågeformuläret angivet i form av ett värde (det kan vara kategorier, koder, nummer). Du kommer snart att få se ett exempel på hur det kan se ut.

Några begrepp

Innan vi pratar mer om dokumentation av data ska jag nämna några olika begrepp som kan vara bra att känna till. Data kan nämligen vara organiserade på olika sätt, vilket i sin tur kan ha betydelse för möjligheter att dokumentera och använda olika verktyg för dokumentation. I kontakt med forskare kommer du antagligen lägga märke till hur de använder olika begrepp för att beskriva data-materialets struktur och hur data är organiserade. Poängen är inte att försöka komma fram till något rätt och fel i sammanhanget, utan snarare att lyfta fram det faktum att vissa begrepp kan ha olika betydelse för olika forskare. Det finns dessutom inga tydliga definitioner och begreppen kan användas olika beroende på sin kontext. Här ser du tre olika sätt att omnämna datamaterialets organisering.

Vi börjar med att titta på begreppen kvantitativa och kvalitativa data. En del forskare använder dessa begrepp när de beskriver sin datamängd. Vad som är viktigt att ha i åtanke är att dessa inte handlar så mycket om själva organiseringen av data, utan snarare om hur de är insamlade eller analyserade. Data kan vara insamlade med en kvalitativ undersökningsmetod, t.ex. djup-intervjuer eller videoinspelningar, men kan sedan struktureras på ett sådant sätt att kvantitativa analyser kan göras. Data som är insamlade via kvalitativa metoder kan till exempel taggas eller beskrivas på ett strukturerat sätt i en datafil. Exempelvis kan man markera alla ställen i ett intervjumaterial där talaren använder logiska eller emotionella argument för att sedan kunna säga vilken typ av argument som är vanligast. Det är alltså inte filerna i sig som är kvantitativa eller kvalitativa utan innehållet och på vilket sätt det kan analyseras.

Andra begrepp som kan användas är strukturerade eller ostrukturerade data. Det kan finnas uppfattningar om att struktur-erade data är kvantitativa medan ostrukturerade data är kvalitativa,

men som tidigare nämndes handlar dessa begrepp snarare om metoder istället för hur data är organiserade. Strukturerat material avser sådant som dataregister, databaser och ärendehanterings-system, medan ostrukturerat material avser sådant som ljud, bild och löpande text. Generellt kan man säga att strukturerade data brukar vara organiserade på ett sådant sätt att de är läsbara för såväl människor som datorer. Det är emellertid inte säkert att forskaren själv som arbetar med bild, ljud, text eller liknande uppfattar sina data som ostrukturerade. För forskaren kan dessa ha en tydlig struktur, men ofta en struktur som en människa har mycket lättare att förstå sig på jämfört med en dator.

Vad menas då med tabulära eller icke tabulära data? Vi börjar med tabulära data, som synonymt kan kallas för rektangulära data. Vad som utmärker dessa är att de i allmänhet kan kvantifieras i siffror eller kategorier och brukar struktureras i rader och kolumner. Det kan till exempel vara mätvärden, befolkningsstatistik från register eller svar från enkätundersökningar med slutna svarsalternativ som finns inmatade i Excel eller statistikprogram som SPSS, Excel, SAS, R osv. Tabulära data behöver dock inte alltid vara numeriska.

Icke tabulära data är vanligtvis producerade under en kvalitativ undersökningsmetod och kan t.ex. bestå av texter, bilder, ljud eller filmer. Till skillnad från tabulära data kan dessa normalt sett inte struktureras i rader och kolumner, däremot kan de ändå betraktas av forskaren som strukturerade. Det kan handla om ett forsknings-material som består av tusentals bilder som är organiserade på ett mycket strukturerat sätt med tydlig koppling till metadata som finns inmatade i en Excelfil.

Under presentationen kommer vi att använda begreppen tabulära och icke tabulära data när vi visar exempel på data och hur de kan dokumenteras. Vi undviker begreppen kvantitativa och kvalitativa, då de snarare relaterar till metoden, och begreppen strukturerade

och ostrukturerade data, då de inte nödvändigtvis används av forskare för att beskriva hur data är organiserade i datafilen.

Dokumentation av data

Generellt kan man säga att det finns många likheter vid dokumentation av data, oavsett om de är tabulära eller icke tabulära. Vissa metadata kan oftast anges direkt i datafilen eller det analysprogram som används. En bra början är därför att ta reda på vilka metadata som är möjliga att ange i det program som forskaren använder för att organisera och analysera sina data. Utöver de metadata som är möjliga att ange i datafilen och/eller analysprogrammet kan också kompletterande dokument behövas, som innehåller information om forskningsprojektet och dess data. Det kan till exempel vara textfiler med beskrivningar av urvalsmetod och datainsamling, eller en detaljerad förteckning med beskrivningar av varje variabel eller objekt som datamaterialet består av.

Dokumentation av tabulära data

Nu ska vi titta lite närmare på dokumentation av tabulära data. I bilden överst på nästa sida ser du ett exempel på data som är strukturerade i rader och kolumner, och inmatade i Excel. I det här fallet motsvarar varje rad en individ som deltagit i studien, men kunde även ha varit ett specifikt objekt som studerades. Kolumnerna motsvarar undersökningens variabler och innefattar här data för respektive individ. I varje ruta där en rad och en kolumn möts ser man värdet för t.ex. hur en individ har svarat på en specifik fråga i enkäten. Namnet på varje variabel anges längst upp och av praktiska skäl brukar dessa ofta innefatta en förkortning av variabelns innehåll. Förkortningar som IDnr, Kon, Civilst osv. ger en viss indikation på dess innehåll, och även om forskaren själv vet exakt vad varje variabel avser så behövs mer information för att en sekundäranvändare ska vara helt säker på vad förkortningarna betyder. Något annat som behöver förklaras är vad siffrorna under varje variabel betyder. Är de mätvärden, kategorier eller kanske

koder? Under kolumnen C: ”F2_Civilst” finns exempelvis siffrorna 1, 2, 3, 4 och 999. Vad dessa siffror betyder framgår inte av den information som datafilen visar.

	A	B	C	D	E
1	Idnr	F1_Kon	F2_Civilst	F3_Fodar	F4_Utb
2	17675638	2	2	56	1
3	17675687	1	3	53	3
4	17675695	1	3	56	4
5	17675729	1	999	60	1
6	17675752	2	6	67	5
7	17675760	2	2	67	5
8	17675778	2	2	78	5
9	17675794	2	1	65	5
10	17675802	1	3	69	5
11	17675810	1	3	76	5
12	17675844	2	1	80	5
13	17675919	2	1	46	4
14	17675935	1	1	48	1
15	17675943	2	4	59	2
16	17675950	2	1	61	3
17	17675968	2	2	67	5
18	17675992	2	2	76	5
19	17676016	1	3	60	4
20	17676024	2	1	87	3
21	17676032	2	2	50	4

Nu går vi vidare till ett exempel på data som är inmatade i statistik-programmet SPSS. I SPSS finns två olika flikar: Data View och Variable View.

	IDnr	F1_Kon	F2_Civilst	F3_Fodar	F4_Utb
1	17675638	2	2	56	1
2	17675687	1	3	53	3
3	17675695	1	3	56	4
4	17675729	1	999	60	1
5	17675752	2	6	67	5
6	17675760	2	2	67	5
7	17675778	2	2	78	5
8	17675794	2	1	65	5
9	17675802	1	3	69	5
10	17675810	1	3	76	5
11	17675844	2	1	80	5

Data View | Variable View

Data View

	Name	Type	Wi...	D...	Label	Values	Missing
1	IDnr	Numeric	8	0	ID-nummer	None	None
2	F1_Kon	Numeric	3	0	Fråga 1 Kön	{1, Kvinna}...	999
3	F2_Civilst	Numeric	3	0	Fråga 2 Civilstånd	{1, Singel}...	997 - HI
4	F3_Fodar	Numeric	3	0	Fråga 3 Födelseår	{44, 1944}...	999
5	F4_Utb	Numeric	3	0	Fråga 4 Utbildningsnvå	{1, Grundsk...	997 - HI

Data View | Variable View

Variable view

Den första bilden, som visar Data View, innefattar samma slags information som vi såg i Excel och är även här strukturerad i rader och kolumner. På den andra fliken, Variable View, finns möjlighet att lägga till metadata om varje variabel som finns i datasetet. I kolumnen Name står namnet på variablerna och som redan nämnts brukar dessa anges med förkortningar. I många statistikprogram kan man t.ex. inte heller använda å, ä eller ö i variabelnamnet. Under kolumnen som heter Label finns däremot möjlighet att i fritext ange en förklaring till den specifika variabeln och under kolumnen Values kan man ange en förklaring till koder och kategorier som används. Svarsalternativ i enkäter är oftast kodade med olika siffror som läggs in i datasetet och under Values kan man ange vilken kod som avser vilket svarsalternativ.

Oavsett vilket program som används för datamängden är en variabellista ett väldigt bra komplement till data. Den är en strukturerad beskrivning med information om varje variabel i datasetet. På nästa sida ser du ett exempel.

Variabellista		
Variabelnamn	*Beskrivning*	*Kodning*
F1_Kon	Fråga1. Är du kvinna eller man?	1= Kvinna 2= Man 999=Uppgift saknas 998=Dubbelmarkering
F2_Halsa	Fråga 2. Allmänt hälsotillstånd	1= Utmärkt 2= Mycket gott 3= Gott 4= Någorlunda 5= Dåligt 999= Uppgift saknas 998= Dubbelmarkering
F2_Halsa_diko	Fråga 2. Allmänt hälsotillstånd dikotomiserat där Utmärkt/Mycket gott/Gott = Gott hälsotillstånd (1) Någorlunda/Dåligt = Dåligt hälsotillstånd (2)	1= Gott hälsotillstånd 2= Dåligt hälsotillstånd 999= Uppgift saknas 998= Dubbelmarkering
P_Glukos	Mätvärde: Blodprov: P-glukos (mmol/L). Kontinuerliga värden. Lägsta värdet i datamaterialet är 3,1 och högsta värdet är 12,7.	3,1 3,2 3,3 …osv… 12,7 999= Uppgift saknas
P_Glukos_3g	Mätvärde:Blodprov: P-Glukos (mmol/L). Indelad i tre grupper.	1= ≤6,0 mmol/L 2= 6,1-6,9 mmol/L 3= ≥7mmpl/L 999= Uppgift saknas

Variabellistan som visas innehåller namnet på de variabler som ingår i datasetet, en beskrivning av varje variabel, och information om de koder, kategorier eller siffror som används för respektive variabel. För en sekundäranvändare ger variabellistan en bra beskrivning av datamaterialet men förutom det kan relevanta metadata även finnas i dokument, såsom frågeformulär, projekt-beskrivning och teknisk rapport. I sådana dokument kan det finnas väsentlig information för en sekundäranvändare, såsom beskrivning av urvalsmetod, information om datainsamling, bortfallsanalys osv.

I tabellen på nästa sida framgår olika metadata som är kopplade till en specifik variabel i en datamängd.

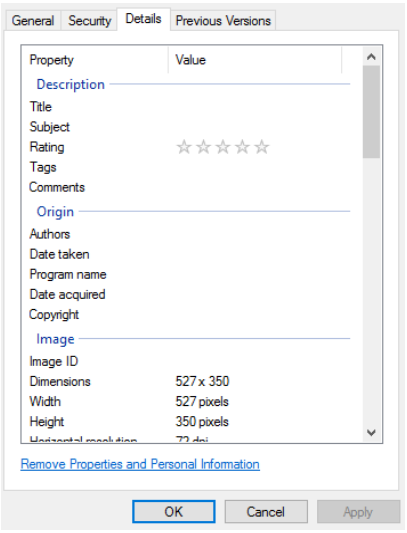
F2_Halso:	Fråga 2 Hälsotillstånd		
Frågeområde i enkät:	B. Hälsa		
Frågetext i enkät:	Hur upplever du ditt allmänna hälsotillstånd?		
Instruktion:	Utgå från det senaste året		
Kod	Förklaring	Individer (antal)	Procent (andel)
1	Bra	1097	38,0
2	Någorlunda	1393	48,2
3	Dåligt	398	13,8
999	Uppgift saknas	103	

Förutom variabelnamnet, som är F2_Halso, framgår en beskrivning av variabeln, vilket frågeområde i enkäten som frågan tillhör och exakt hur frågan har ställts i enkäten, den instruktion som respondenten fick för att besvara frågan, och vad de olika koderna betyder. Dessutom framgår frekvensfördelningen som visar hur många individer som svarat på respektive svarsalternativ samt den procentuella fördelningen. Den information som tabellen visar är alltså exempel på metadata som är kopplade till en specifik variabel i ett dataset. Vilka metadata som är relevanta att producera varierar såklart mellan olika forskningsämnen och typer av data. De metadata som bilden visar kan vara relevanta för data som är insamlade via frågeformulär medan arkeologiska data eller klimatdata kräver annan typ av metadata. Forskaren är den som känner datamaterialet bäst och som också kan producera de metadata som behövs.

Dokumentation av icke tabulära data

Nu har vi fokuserat en del på data som är tabulära, men hur kan icke tabulära data dokumenteras? Som tidigare nämnts är icke tabulära data vanligtvis producerade under en kvalitativ undersökningsmetod. Det kan till exempel vara text-, bild- eller ljudfiler. När det kommer till metadata finns ofta möjlighet att ange vissa metadata direkt i filerna. Ett annat sätt är givetvis att anteckna i ett separat

textdokument där det tydligt framgår vilken del av datamängden som beskrivs, så att datafiler och dess innehåll kan matchas mot dokumentationen. Generellt är det ovanligt med dokumentationsprogram som är avsedda för datatyper som inte är tabulära, men en del analysprogram ger möjlighet för viss dokumentation. Du kommer strax att få se några exempel på analysprogram.



I bilden till höger ser du ett exempel på hur information kan skrivas direkt i en bildfil. Under egenskaper går det t.ex. att ange en beskrivning och lägga in uppgift om plats, datum, nyckelord. På det här sättet blir metadata tätt knutna till själva datafilen. För ljudfiler finns ofta liknande möjligheter där metadata kan läggas in i filen. När datamaterial sedan ska tillgängliggöras är det bra att skicka med en readme-fil eller liknande där det framgår att vissa data finns i respektive bild- eller ljudfil samt om det finns ytterligare metadata i andra kompletterande dokument.

Ett annat sätt att dokumentera bildfiler är att samla alla metadata på ett strukturerat sätt i ett separat dokument. Information om respektive bild finns samlad i ett Exceldokument. Se exemplet nedan.

D	E	F	G	H	I	J
Grave plot	**Surname**	**First name/initials**	**Name as inscribed**	**Gender**	**Proba**	**Age**
V 5 45	Sjöblom	Johan	Johan Sjöblom	Male	1888	57
V 5 45	Sjöblom	Anna Julina	Anna Julina	Female	1888	64
V 6 50	Montell	Mortimer	Mortimer Montell	Male	1881	38
V 6 50	Montell	Valentin	Valentin Montell	Male	1925	86
II 4 106	Simonsen	M. C	M.C. Simonsen	Male	1919	29
III 2 33	Hellstrom	Karl Johan	Karl Johan Hellstrom	Male	1902	47
III 2 33	Hellstrom	Amalia	Amalia Hellstrom	Female	1929	73
III 2 34	Lindström	Johan	Johan Lindström	Male	1902	76
V 7 61	Sandberg	Ewa Justina	EWA JUSTINA SANDBEI	Female	1905	57
V 8 74	Sandberg	Matts	MATTS SANDBERG	Male	1907	62
V 9 83	Treffz	Herman	HERMAN TREFFZ	Male	1918	41
III 1 15	Lindblom	Teodor	TEODOR LINDHOLM	Male	1918	45

Varje rad i bilden motsvarar en gravplats som fotograferats, och kolumnerna innefattar olika slags information om respektive grav. En förutsättning för att dokumentationen ska vara användbar är att det finns en tydlig koppling mellan de objekt som beskrivs i dokumentet och de faktiska bilderna som dokumentationen avser. Att namnge bildfilerna på ett tydligt sätt är därför viktigt.

Verktyg för dokumentation av icke tabulära data

Här är några exempel på verktyg för datamaterial som inte är tabulära. Syftet med dessa är främst för att använda i samband med analys av data, men de ger också möjlighet att lägga in vissa metadata. Något som är viktigt att tänka på, för att säkerställa data-materialets möjligheter till återanvändning, är att spara filer i format som är lämpliga för långtidsbevarande. Vilka möjligheter som finns för export till olika format skiljer sig mellan olika program och är därför en viktig sak att ta reda på innan användning.

- Transana är ett verktyg för kvalitativ analys av text-, bild-, ljud- och video-filer. Det ger möjlighet att organisera video- och/eller ljudklipp. För ljudfiler finns t.ex. möjlighet att skriva transkript som sedan kan synkroniseras med ljudspåret. Detsamma gäller för video.
- ELAN är ett program som främst är tänkt att användas i syfte att analysera ljud, bild eller video. Det är möjligt att skapa komplexa anteckningar på video- och ljudresurser (t.ex. registrera rörelser i video, dokumentera intonation i röster).
- Nvivo används vanligtvis inom kvalitativ forskning och lämpar sig bland annat för ostrukturerade data, dvs. intervjuer, öppna enkätsvar, sociala medier, litterära texter m.m. Programmet kan hjälpa till att strukturera, organisera och analysera många olika typer av data, t.ex. video, bild, ljud och text.
- Voyant Tools är en webbaserad analysapplikation och är designad för att vara ett generellt verktyg för analys av text-baserade data.

- OpenRefine är ett verktyg för att rensa data, transformera från ett format till ett annat osv.

Sammanfattning

Metadata på variabelnivå handlar om att ge en detaljerad beskrivning av de specifika delar som datamängden består av. Ett sätt att skilja mellan olika former på data är att prata om tabulära och icke tabulära data. Data som är tabulära kan t.ex. bestå av siffror, koder och kategorier medan icke tabulära data t.ex. kan vara bilder, texter och filmer. Många program som forskare använder har stöd för att lägga till vissa metadata. En bra utgångspunkt är att använda de möjligheter för dokumentation som finns och komplettera metadata i ytterligare dokument. Det väsentliga är att all dokumentation som behövs för att förstå och använda datamaterialet finns tillgängligt för sekundäranvändaren.