SND
Svensk nationell datatjänst

Centrala områden inom datahantering

Pass 2: Datahantering och datahanteringsplaner

BAS Online 2021-01-20

I den här presentationen ska jag ta upp tre huvudsakliga områden inom datahantering och några centrala delar av dessa. Först och främst är det viktigt att komma ihåg att datahantering – och datahanteringsplaner – är något som man bör ha med sig genom hela forskningsprocessen, från den spirande projektidén till det färdiga projektet med dess publicerade artiklar och tillgängliggjorda data. Och oavsett vilken sorts projekt det rör sig om och vilka sorters data som ingår i det så är det tre huvudsakliga områden som man behöver fundera kring: organisation, dokumentation och teknik.

Organisation

Det första området som datahanteringsplaner omfattar handlar om att ha ordning på saker och ting. Ett exempel på en väsentlig form av *organisation* är versioneringen. Det är väldigt lätt att råka göra fel som kan leda till onödigt merarbete. Bra versionering ska göra det möjligt att se vilken som är den senaste versionen av ett dataset och hålla ordning på vad som har gjorts med olika versioner, så att man vet var i arbetsflödet en specifik fil befinner sig och så att eventuella fel kan spåras bakåt. Versioneringsregler bör vara så enkla som möjligt och så komplicerade som nödvändigt, med det är en god idé att åtminstone ha en lista med information om vad som skiljer varje ny version från den föregående. God versionering gör också att man kan koppla olika versioner av olika filer till varandra inom projektet, och versioner kan användas för att markera på vilket stadium i en process som filer befinner sig.

Faktum är att versionering inte endast kan användas för data. Enkla varianter av versionering gör även texthantering effektivare,

och inte minst datahanteringsplanen i ett projekt bör versioneras eftersom den ska vara ett levande dokument som uppdateras utifrån nya beslut och ändrade förutsättningar.

Hur många personer känner du som har perfekt ordning på sina filer på hårddisken? Under en löpande verksamhet så verkar det vara ofrånkomligt att man improviserar lite vad gäller strukturer, men för ett forskningsprojekt finns det allt att vinna på att man har data och annat material på ett logiskt och planerat sätt. Enklaste sättet att åstadkomma det är att i förväg fundera på vad en rimlig struktur skulle kunna vara för den information som kan förväntas dyka upp i projektet. Ju färre nivåer man har i sin mappstruktur, desto svårare är det att hitta rätt, men det går att gå till överdrift också. Och det kan vara en idé att tänka igenom strukturerna om de inte visar sig tillräckliga istället för att börja improvisera: konsekvens är ett nyckelord! Egentligen är det här inga konstigheter, men med tanke på hur ofta man träffar någon som verkar ha alla sina filer på Windows-skrivbordet så tål det att påpekas med jämna mellanrum.

Konsekvens och struktur är viktigt även när det gäller filnamn. Låt oss börja med ett exempel som visserligen är fingerat men ligger ganska nära sanningen. För samma projekt skulle filerna antingen kunna vara namngivna så här:

- talare1_ord_v0_orig.wav
- talare1_ord_v1_ren.wav
- talare1_ord_v2_ren.wav
- talare1_ord_v3_ren_slutlig.wav

eller så här:

- PeterS_ordlista_17jun.wav
- talare1_ord_slutlig.wav
- talare1_slutlig2.wav
- talarePS_ord_ren.wav

I det första exemplet är det lätt att förstå hur filerna hänger ihop. Filnamnet anger vilken talare det rör sig om – talare 1 i det här fallet. Det syns att det rör sig om materialet ”ord”, och vilken version av datasetet som varje fil innehåller. Den första filen är tydligt markerad som originalinspelning, de andra är rensade på något sätt, och version 3 är den slutgiltiga versionen.

Det andra exemplet är mer svårtytt. Rör det sig om samma talare i alla filerna? Är ordlista samma sak som ord? Hur förhåller sig ”ren” till slutlig – och slutlig till slutlig 2? Inkonsekvenserna gör det om inte omöjligt så svårt att avgöra vad filerna innehåller. En anledning till att vara noggrann med hur man namnger sina filer och inte bara förlita sig på en god mappstruktur är risken att filerna skulle kunna trilla ur sina mappar av misstag eller genom att någon kopierar bara enstaka filer till någon annan mapp eller till en annan lagringsenhet. SND fick in ett dataset med en väldigt bra mappstruktur som innehöll hundratals bildfiler. Man kunde tydligt se vilken plats de kom ifrån, vilket motiv på just den platsen som hade fotograferats och när bilderna tagits. Men i strukturens nedersta mappar låg bilderna, namngivna A, B, C, D, E, F och så vidare. Det här var inget problem så länge de stannade i mapparna, men om dessa skulle försvinna så skulle man bli sittande med femtioelva filer som heter A utan aning om vart varje fil hörde. Det finns därför anledning att ha ganska mycket information i filnamnen. För säkerhets skull.

Med andra ord, organisera data så att det är lätt att hitta rätt version och rätt filer och var framför allt **konsekvent** i tillämpningen av systemen.

Dokumentation

Det andra övergripande området rör *dokumentation*. Den första av våra tre centrala punkter är att man ska använda standarder. Förutom fördelarna som finns med att inte uppfinna hjulet på nytt så

blir det lättare att jämföra med andra material som andra har använt. Många forskare är ovana att använda metadatastandarder för att strukturera information om forskningsmaterial, så det viktiga att förmedla är att det helt enkelt handlar om att beskriva saker på samma sätt som andra gör. Loggar är vanliga inom vissa discipliners forskningsmetodik men ovanliga inom andra. Men oavsett vilka material man arbetar med är det en god idé att logga inte bara vad man gör med sitt forskningsmaterial utan även vilka beslut man fattar. Information om till exempel ”varför används de här filnamnskonventionerna?” eller ”varför ser mappstrukturen ut så här?” kan vara bra att logga både för egen och andras skull – det vill säga man uppdaterar sin datahanteringsplan med den.

Informationen man samlar om datahanteringen behöver vara lätt att hitta och lätt att följa. Är systemen för hur man dokumenterar svåra att förstå eller tar för mycket tid att hålla sig till är risken att man själv eller de man arbetar med inte följer dem. Då har man byggt in möjligheten att datahanteringen och dokumentering går åt skogen. Den dokumentation man har behöver dessutom vara **begriplig** både för en själv och för andra, oavsett om dessa andra är projektparter, tidskrifternas granskare eller sekundärforskare. Men det här gäller förstås inte bara forskare utan alla som på ett eller annat sätt jobbar tillsammans med andra människor.

Teknik

Det tredje området handlar om teknik. Vikten av säkerhetskopior har vi berört tidigare men den tål att upprepas. Det är två huvudsakliga problem som man behöver ha i åtanke när det gäller säkerhetskopior av data. Dels finns det en risk med att ha kopior av data på flera ställen, eftersom det kan leda till att man själv eller någon annan i projektgruppen råkar använda en gammal version istället för den senaste. Dels behöver man vara medveten om att alla lagringsmedia medför olika risker för dataförlust. Eftersom filer kan försvinna om hårddisken kraschar eller datorn blir stulen så

rekommenderas till exempel att originalfiler aldrig sparas på den egna datorns hårddisk. Forskningsdata bör inte heller lagras på USB-minnen, CD- eller DVD-skivor, externa hårddiskar eller liknande om det inte finns en säkerhetskopia på en säker lagringsyta, eftersom det finns risk att lagringsmediet går sönder, tappas bort eller stjäls. Skulle datamaterialet innehålla känsliga uppgifter så kan det krävas en högre nivå av säkerhetsklassning. Riktlinjen är att oavsett vilken typ av lagring som väljs så ska regelbundna backuper göras, gärna flera gånger varje dag. Är det flera personer som behöver komma åt och arbeta med samma forskningsdata är det ännu viktigare att data ligger säkert på en gemensam yta och att det inte skapas olika lokala versioner av samma filer. Finns det en osäkerhet om vilken lagringslösning forskaren eller forskargruppen ska välja bör lärosätets lokala IT-avdelning kontaktas.

Det finns mycket att säga om filformat. Generellt kan man säga att när man ska välja filformat bör de vara så vanliga och allmänna som möjligt. Ovanliga och specialiserade filformat riskerar att inte kunna öppnas i framtiden. Inte minst proprietära filformat – filformat som privata företag har upphovsrätt till – riskerar att vara svåra eller omöjliga att läsa i framtiden, eller kräver program som inte alla har tillgång till. Har programmet öppen källkod – Open Source – eller sparar filerna i ett standardformat så ökar chansen att informationen i filen är tillgänglig även i framtiden. En länk till rekommendationer om lämpliga filformat finns nedan. Där finns även en länk till ett antal dokument med *best practices* för olika typer av data.

När det gäller programvaror och verktyg ska man fundera på till exempel om inställningar, programvaruversioner och modeller som används för att samla in, analysera eller på annat sätt hantera data kan påverka utfallet och således kan behöva dokumenteras. Det beror väldigt mycket på inom vilket område man forskar: för vissa är svaret ett självklart nej och för andra är det ett lika självklart ja. En annan fråga att ställa sig är huruvida man har kod eller

program eller skript som man har specialdesignat för att analysera eller på annat sätt arbeta med sina data. Om så är fallet kan dessa också behöva dokumenteras och sparas tillsammans med data för att möjliggöra framtida användning.

En viktig fråga nu när databaser blir en allt vanligare form av resultat från forskningsprojekt är hur dessa databaser ska kunna fortsätta drivas vidare långt efter att projektet är avslutat, kanske för all framtid. Idag är det här ett vanligt problem, så om forskare söker medel för att skapa en databas som ska hållas tillgänglig efter projektslut är det värt att uppmana dem att fundera på hur databasen ska driftas, vem som ska drifta den och vem som ska betala för det. Det är inte roligt att hamna i samma situation som en forskargrupp vid ett stort universitet. De hade lovat finansiären att projektet skulle resultera i en databas som skulle drivas i tio år efter projektslut, men de hade inte begärt medel för driften. De plockade in en utländsk utvecklare som byggde databasen åt dem och som sedan åkte hem igen. Där satt de med ett åtagande till finansiären men utan någon som visste exakt hur databasen fungerade eller hur den skulle underhållas framöver. Det är med andra ord värt att tänka på det här på ett tidigt stadium. Det går att plocka ned en databas och spara den, men det är inte samma sak som att ha den uppe och tillgänglig via nätet för andra forskare. Det blir ännu mer komplicerat om det handlar om en databas som är tänkt att vara levande och utökas löpande över tid – då behöver den ännu mera vård och omsorg och finansiering.

Så med andra ord, **tänk framåt** så är en hel del förhoppningsvis vunnet.

Sammanfattning

Den här presentationen har gett en kort översikt över tre huvudområden inom datahantering: konsekvent organisation med fokus på versionering, lagringsstrukturer och namnkonventioner för

filer; begriplig dokumentation genom standarder, loggar och samlad information; och ett framåttänk vad gäller teknik, som säkerhetskopiering, filformat och programvaror. Nästa presentation kommer att ta upp hur SND:s checklista kan stötta utvecklingen av datahanteringsplaner.

Referenser

Svensk nationell datatjänst. (2017) *Best practices för olika datatyper*: https://snd.gu.se/sv/hantera-data/guider (Hämtad 2021-01-20)