SND
Svensk nationell datatjänst

Datahantering

Pass 2: Datahantering och datahanteringsplaner

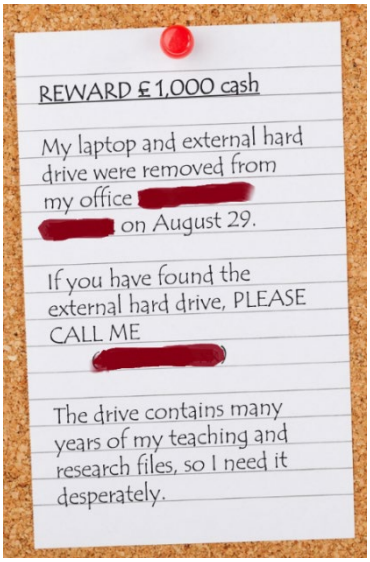
BAS Online 2021-01-20

I den här presentationen ska jag ge en kort inledning till vad datahantering är och vad det är bra för utifrån perspektivet att man i en DAU kan komma att presentera det för forskare. Låt oss börja med att konstatera att datahantering är något som forskare redan sysslar med. Beroende på vad de forskar på och inom vilken disciplin de forskar så finns det redan mer eller mindre utarbetade konventioner kring hur data ska hanteras under ett projekt. Vissa forskare tänker mer på en aspekt av hanteringen, andra tänker mer på någon annan aspekt. I den här presentationen utgår jag emellertid från något slags noll-läge: att forskaren man pratar med i princip inte vet överhuvudtaget vad det handlar om. I verkligheten krävs det att man har en viss fingertoppskänsla och ödmjukhet när man pratar datahantering med forskare, tills man fått förståelse för vilka delar av datahanteringen de redan är bekanta med.

Vad är vitsen med datahantering? I det här passet kommer tonvikten att läggas på forskarperspektivet och på hur man som DAU kan kommunicera med dem. Hur kan man få forskare att inse att de har något att vinna på god datahantering, det vill säga, vilka morötter kan man erbjuda?

Datahantering ska först och främst syfta till göra forskningen effektivare och lättare. Hanteringen ska inte vara en administrativ börda som läggs på forskaren, och när man utbildar forskare i datahantering måste man hela tiden trycka på att det finns vinster av olika slag för dem att göra här. Man måste alltså betona morötterna, och om man osäker på vilka morötterna är så är det en anledning till eftertanke.

Tanken är att man i mötet med forskare understryker att datahantering gör det lättare att komma åt projektets data när man behöver dem; att man ska kunna vara säker på att dessa data har rätt format; och att man ska vara säker på att det inte har hänt något med dem på vägen.



När man hör om datahantering som fallerar rör det sig oftast om tekniska problem, och det är ganska ofta problem som hade kunnat förutses eller förebyggas. Den här efterlysningen från London School of Economics med en hittelön på tusen pund vittnar om vilken desperation förlusten av den externa hårddisken gett upphov till. År av arbete som har försvunnit, uppenbarligen utan att det finns några säkerhetskopior. Även om det här fallet kan tyckas ovanligt hemskt så är det egentligen ganska typiskt. De flesta exempel man hittar på hur bristande datahantering har ställt till det handlar om förlust av data på grund av något slags tekniska problem. Dataförlust på grund av att säkerhetskopior saknas ligger högt på listan. Därmed inte sagt att man inte kan göra fel på andra sätt också, förstås.

Efter att ha inlett sin presentation för forskarna med en förklaring om hur datahantering gör livet, eller åtminstone forskningslivet, lättare så kan man sedan ta upp mer konkreta områden. Det är avsiktligt som listan över fördelar inleds med vad som skulle kunna uppfattas som det mer egennyttiga: att det handlar om organisation, och om lagring och teknisk kopiering. De här två anledningarna är lätta för de flesta att relatera till. Även om man

själv inte har alla sina filer i en enda röra på skrivbordet så känner man till dem som har det, och många håller antagligen mindre ordning än vad de vet med sig att de borde. Går det att ge förslag på praktiska lösningar på det här så har man kommit långt. På samma sätt är det förvånansvärt många som är väl medvetna om att de borde säkerhetskopiera men som inte gör det. När man tagit upp de bitarna kan man sedan gå vidare med mer altruistiska anledningar, som att dela data och bevara data för framtiden. Kör man enbart på att det är bra att bevara data för framtiden så kommer man inte att vinna sin publiks hjärtan. Det finns ett fåtal som tycker att det är viktigt, men de allra flesta är intresserade av vad just de kan vinna på det. Det bör även nämnas att det här med att dela data inte har hög prioritet inom många discipliner medan det inom andra discipliner är väldigt populärt. Det görs även på olika sätt. Inom bland annat medicin så sker det inte ofta via institutionella eller ämnesspecifika repositorier. Istället delar man med sig till kollegor som man känner, och samskriver eventuellt en artikel med dem. Samskrivande är alltså ett sätt att dela data som fungerar för forskare i medicin och också inom stora delar av naturvetenskaperna och samhällsvetenskaperna. Men inom många humanistiska områden betraktas samskrivna artiklar fortfarande med viss tveksamhet eller till och med stor misstro.

Det finns även andra sätt att uttrycka vinster med genomtänkt datahantering: den sparar tid och gör forskning effektivare, och den underlättar dokumentation och långtidsbevarande. Även här ligger tonvikten på nyttan för den egna forskningen. Samtidigt finns det fördelar med att poängtera värdet av god dokumentation, lite grann beroende på vem som lyssnar på presentationen: om man talar till en stor forskargrupp i naturvetenskap eller medicin bestående av en professor som leder labbet, kanske ett par seniora forskare, några postdoktorer, ett stort antal doktorander och några mastersstudenter, och alla dessa ska kunna hålla ordning på vad som händer med alla data som finns i projektet, ja,

då är dokumentation antagligen definitivt ett säljande argument. Rör det sig om en enskild humanist så är det mindre så, för då handlar det om en person som behöver hålla ordning på sitt material. Vitsen med datahantering beror alltså på vilka forskare man möter. Men man behöver vara medveten om att forsknings-effektivitet är ett starkare argument än långtidslagring för att locka de flesta forskare att bry sig om datahantering: nästan alla kan intressera sig för att bli effektivare men i dagsläget funderar merparten av forskarna inte ens på att långtidslagra sina data.

Sedan är det förstås så att många grupper inom teknik, natur-vetenskap, medicin med mera redan idag har med en eller flera personer som har ansvar för datahantering, men inte alltid. Det beror på vilka discipliner, och vilka lärosäten, och vilka forskar-grupper. Vissa forskargrupper har en datahanterare, eller data-samordnare eller data manager – det kan kallas för lite olika saker – som har som heltidssyssla att hålla ordning på det data-relaterade. Andra har det inte. En forskare som kontaktade SND sa ”ja, just det, det kanske är en bra idé att hålla koll på vad alla har för data i arbetsgruppen” och det är ju bara att hålla med om. Man ska med andra ord vara medveten om att det där skiljer sig åt och att det finns en risk att man trampar någon på tårna om man antyder att de behöver jobba med datahantering när de redan sysslar med det.

Att dokumentera handlar inte bara om att tala om för andra hur man utfört en studie utan även om att själv komma ihåg vad man har gjort i sin forskning. Det här är ett säljande argument, inte minst till doktorander. Det må vara så att man tänker sig att man aldrig kommer att glömma vad man gör under sitt avhandlings-arbete, men detaljer kan blekna snabbt, framför allt om andra projekt kommer emellan. Det är värt att uppmana forskare att fundera över några tänkbara scenarion, som till exempel: vad gör man som forskare om man skulle vilja utveckla en flera år gammal

studie? Kommer man ihåg hur filnamn konstruerades, hur data har rensats, vad olika förkortningar betyder och så vidare? Tänk om någon kommer och ber att få tillgång till data från studien för att de ska kunna bygga vidare på ens forskning? Vet man var data finns och är de i sådant skick att de går att lämna ut? Eller ännu värre, vad händer om man blir anklagad för forskningsfusk? Kan man bevisa att man inte har falsifierat data utan faktiskt har utfört arbetet så som man har beskrivit det, och i så fall hur? Kan man bevisa att data inte är tillrättalagda eller påhittade? Vad skulle man kunna ha gjort under projektets gång för att underlätta om man någonsin skulle hamna i en sådan situation?

Nedan finns det länkar till två kortfilmer som illustrerar en del av problemen som skulle kunna uppstå: en presentation av TROLLing, The Tromsø Repository of Language and Linguistics, och en animerad dramatisering av tre vanliga misstag inom tillgängliggörande och hantering av forskningsdata.

Slutligen vill jag understryka att det är viktigt att ta hänsyn till de stora skillnader som råder mellan olika forskare och olika projekt, och hur dessa skillnader påverkar vad som är rätt datahantering för någon. Vissa forskare arbetar ensamma eller i par medan andra arbetar i stora forskargrupper. Även om det finns likheter i hur man organiserar sina egna forskningsdata på den egna hårddisken, så ska man komma ihåg att om det rör sig om en forskargrupp är arbetet på den egna hårddisken mindre relevant. I stora grupper jobbar forskare förmodligen mot ett säkert system där data finns. På hårddiskar har de bara arbetsfiler, kopior eller utsnitt av dataseten. När man informerar om datahantering gäller det alltså att inse att forskarna i slutändan är experter på sina egna behov och på sina egna processer. Vi kan emellertid erbjuda möjliga redskap för dem att förbättra dessa.

Sammanfattning

I den här presentationen har jag tagit upp grundtankarna bakom datahantering och hur man kan förmedla dem till forskare. Exakt vad datahantering omfattar beror på faktorer som disciplin, datatyper och analysmetod. Grundtanken är att man själv och andra ska kunna hitta, komma åt och återanvända data samtidigt som man är säker på att inga fel smugit sig in – det vill säga, de ska kunna följa FAIR-principerna som nämns i pass 1. Vissa forskare är vana vid datahantering i den här betydelsen, andra är det inte. Det svenska forskarsamhället kan ses som ett lapptäcke där det förekommer områden med enormt god datahantering och områden där man inte är det minsta medveten om behovet av datahantering, och allt däremellan. Vissa delar av datahanteringen är väl inarbetade i forskningsprocesserna medan andra delar är mer eller mindre okända. För att kunna ta hänsyn till den här bredden av olika behov ska vi i resten av det här passet titta på ett redskap som erbjuder en möjlighet att skapa ett stöd för datahantering inom ett visst projekt, nämligen datahanteringsplanen.

Referenser

TROLLing:
https://www.youtube.com/watch?v=uEf0c0NT9_A&feature=youtu.be

Data Sharing and Management SNAFU in 3 Short Acts:
https://www.youtube.com/watch?v=N2zK3sAtr-4