Open data flagship pilot 2022

Slutrapport

Urban Andersson
Lina Andrén
Jeremy Azzopardi
Olof Olsson

https://dx.doi.org/10.17196/snd.flagship-open-data.2022
Publicerad: 2023-02-15

1. Uppdrag

Uppdraget har bestått i en ”Open Data Flagship-pilot" där man som best-practice-exempel i samverkan bygger upp enkla och för forskaren effektiva arbetsflöden på respektive lärosäte. Dessa använder lokal information om forskare och forskningsprojekt och lokala verktyg för datahanteringsplaner för att stödja publicering genom SND. Det handlar både om att finna, skapa och använda modulära tekniska lösningar och tjänster (inklusive API:er) och om datalogistik, ansvarsfördelning och policyer, dvs vem som gör och ansvarar för vad i olika delar av processen.

I uppdraget ingår t ex att;

- Återanvända ifylld information (metadata) från en datahanteringsplan vid senare publicering av data
- ta fram en gemensam målbild
- kompetensspridning
- bygga samarbete och utveckling tillsammans

Uppdraget skall vara utmaningsdrivet, dvs. utgå från målgruppernas behov.

Enligt uppdragsbeskrivning i ansökan skall projektet, med fokus på (maskinläsbara) datahanteringsplaner och deras centrala och sammanhållande roll i forskningsdataprocessen, ge en fördjupad förståelse av hur nyttjande av öppna API:er och maskinläsbara metadata kan underlätta för forskare och övriga aktörer (Data Office/DAU, SND, lärosäte, finansiärer), samt bidra till bättre, aktuell och korrekt information i berörda system. Projektet skall resultera i (minst) en testad och implementeringsfärdig teknisk lösning för att kunna åstadkomma detta i praktiken.

SND flaggskepp

De lärosäten som ingår i SND-konsortiet har åtagit sig att medfinansiera SND:s verksamhet in kind med en heltidsekvivalent per lärosäte, vilken motsvarar cirka 1,5 miljoner kronor per år.
I ansökan till Vetenskapsrådet beskrivs **flaggskepp** som ett sätt att organisera medfinansieringen.

För att flaggskeppen skall kunna göra nytta nationellt behöver resultaten bli tillgängliga FAIR och öppet. Detta är en viktig del av flaggskeppsarbetet. Vid flaggskeppets slut avges normalt en kort skriftlig rapport. Detta behövs dock inte om resultat producerats som står för sig själva och utgör leverabel, tex utbildningsmaterial. Flaggskeppen förväntas sprida sina resultat inom SND-nätverket genom presentationer på SND:s nätverksträffar och rapportering via Basecamp. Beroende på flaggskeppets natur kan ytterligare former för spridning vara lämpliga, till exempel workshops, webbinarier, hackathons o. dyl. När resultaten har relevans utanför SND-nätverket sprids de exempelvis genom presentation på internationella konferenser och/eller publikation i tidskrift.

2. Genomförande

Arbetet har vägletts av en styrgrupp bestående av representanter för de tre ingående organisationerna - Maria Kinger (Chalmers), Rosa Lönneborg (KTH) och Gustav Nilsonne (SND). Projektgruppen har bestått av Urban Andersson (Chalmers, projektledare), Lina Andrén (KTH), Jeremy Azzopardi (Chalmers) och Olof Olsson (SND).

Projektgrupp och styrgrupp har haft återkommande digitala möten, medan projektgruppen har haft möten både digitalt och vid heldagsmöten på plats i april och augusti.
Vid SND:s nätverksträff i Göteborg den 27 april presenterades och diskuterades projektet med andra nätverksmedlemmar. I samband med detta inhämtades också synpunkter från användargruppen i en anslutande workshop. Vid ett arbetsmöte i augusti deltog Joakim Philipsson (SU) för att diskutera hur Stockholms universitet har arbetat med maskinläsbarhet i datahanteringsplaner. Projektet har också presenterats vid SND:s IT-forum samt för SND:s personal respektive styrgrupp och vid en webbinar arrangerad av RDA Nordic den 29 juni.
Projektet och resultatet kommer att redovisas vid en webbinar den 13/3 2023.

3. Bakgrund och problembeskrivning

En datahanteringsplan analyserar och dokumenterar hanteringen av data i ett forskningsprojekt. Den är i optimala fall ett levande dokument som omfattar samtliga faser i projektets genomförande – från planering till publicering av output och data, samt arkivering av data för långtidsbevarande. Den har ofta ett särskilt fokus på hantering av data som kan vara problematisk och/eller kräva extra stöd och resurser, såsom mycket stora eller särskilt skyddsvärda och känsliga data. Den utgör därmed en viktig källa för olika stödfunktioner och intressenter, exempelvis för att kunna planera för datalagring eller att identifiera forskningsprojekt som hanterar sådana data.

Allt fler finansiärer kräver också att det skall finnas en datahanteringsplan för beviljade projekt. Det finns dock goda skäl till att ha en sådan även i de fall där finansiären inte explicit kräver detta. Sist, men inte minst, är datahanteringsplanen ett viktigt verktyg för forskarna själva, för att redan på ett tidigt stadium planera för och organisera data, samt att löpande kunna dokumentera hanteringen av dessa. Dels för att underlätta publicering och arkivering i slutet av projektet, men också för att kunna få bättre stöd med datahanteringen under genomförandefasen.

Datahanteringsplanen kan i det enklaste fallet bestå av ett fysiskt, mer eller mindre statiskt dokument, som i normalfallet diarieföras vid lärosätet. Detta dokument ersätts då helt enkelt av nya versioner vid förändringar. Idag erbjuds ofta ett särskilt systemstöd för detta, vilket underlättar för forskaren, både när en ny plan skall upprättas och när den skall uppdateras under projektets genomförande. I dessa system kan forskaren också få stöd och ledning genom anpassade formulär med frågor och hjälptexter. De två system som har undersökts i detta projekt är DMP Online och Data Stewardship Wizard (se mer nedan).

Även om dessa system underlättar för forskaren, när det handlar om att upprätta och underhålla datahanteringsplaner, så kan det ändå vara svårt för intressenter att ta del av innehållet på ett effektivt sätt om det som systemet genererar endast är ett statiskt dokument. Ofta är dessutom stora delar av datahanteringsplanens innehåll uttryckt i fritext, i form av fråga-svar, vilket kräver genomläsning och analys av varje enskild plan för att kunna identifiera eventuella problem och behov. Det kan också vara svårt att återvinna information från datahanteringsplanen i andra delar av processen, exempelvis när data skall publiceras.

För att skapa bättre förutsättningar talar man idag om *maskinläsbara datahanteringsplaner* (”machine actionable dmps”, fortsättningsvis benämnda *maDMP*). Dessa skapas och underhålls som vanligt i befintliga system. Men utvalda delar av innehållet kan levereras digitalt till andra system och tjänster. Antingen på forskarens eget initiativ (push) som till exempel ett dokument i JSON-format, eller genom ett API, där auktoriserade tjänster, vid behov, kan söka fram och läsa de maskinläsbara delarna av innehållet automatiskt. Exempel på det sistnämnda skulle kunna vara tjänster som bevakar och fångar upp behov av datalagring eller särskild hantering av känsliga data. Det skulle också kunna vara tjänster för publicering eller arkivering av data, som kan läsa validerad och kvalitetskontrollerade metadata från datahanteringsplanen, och återanvända denna som underlag för att skapa en publikation eller ett arkivobjekt.

Om maskinläsbarhet skall kunna nyttjas optimalt krävs att informationen i datahanteringsplanen är i relativt hög grad standardiserad. Detta innebär att de generella, breda frågor som renderar beskrivande svar och som är mycket vanliga i befintliga systems frågeformulär, kan behöva ersättas med mer detaljerade och specifika frågor, i bästa fall, om så är möjligt, med valbara alternativ. Det krävs också att metadatamodeller i de system som används är standardiserade, samt givetvis att det finns tekniska möjligheter att kommunicera och leverera den maskinläsbara metadatan på ett effektivt och säkert sätt.
De system som vi har tittat på uppfyller i stort dessa krav, men implementeringen skiljer sig åt mellan systemen, vilket skapar vissa utmaningar. Dessa beskrivs närmare nedan.

Det är viktigt att påpeka att maskinläsbara datahanteringsplaner inte i sig utgör ett alternativ till, eller en ersättning för traditionella planer. Tanken är inte heller att hela planen nödvändigtvis skall eller kommer att vara fullt tillgänglig i denna form. Nyttan med maskinläsbarhet skall snarare ses mot bakgrund av befintliga, praktiska behov från intressenterna, inklusive forskarna själva. Med det primära syftet att underlätta och effektivisera delar av hanteringen och överföringen av dokumentation kan en maskinläsbar datahanteringsplan bidra till exempelvis undvika att samma administrativa information måste matas in manuellt upprepade gånger vid olika tillfällen. Till exempel kan identifikatorer för projektdeltagare och organisationer överföras automatiskt från datahanteringsplanen till en forskningsdatakatalog vid publicering. Ett annat exempel är att om systemet tydligt kan indikera förekomst av personuppgifter eller andra känsliga data kan detta användas för att tillse att forskaren får kontakt med och tillgång till stödprocesser redan tidigt i planeringsfasen.
Datahanteringsplanen kan också vara ett underlag och ett dokumentationsverktyg för processer som behöver göras under, innan och efter datahanteringen sker i praktiken, såsom riskbedömningar, dokumentation av personuppgiftsansvar, reglering av åtkomst till data med mera.
Detta sätt att se på maDMP uttrycks för övrigt också i valet av termen ”actionable”, snarare än ”readable”.

4. Kontext

4.1 RDA Common standard

Som nämnts ovan så syftar maskinläsbara datahanteringsplaner till att möjliggöra ett bättre och mer effektivt nyttjande, validering och återanvändning av värdefull information i dessa planer. Detta kan åstadkommas genom integrationer med andra system och kommunikationskanaler som berör intressenter i forskningsprocessen, såsom lärosätet och olika aktörer inom detta (DAU-funktionen, stödfunktioner för it-säkerhet, dataskydd, datalagring med mera), finansiärer, publiceringstjänster och, inte minst, forskarna själva. Att hantera datahanteringsplaner i ren digital form skapar också bättre möjligheter att uppdatera informationen vid behov och därigenom åstadkomma de levande dokument som dessa planer bör vara.

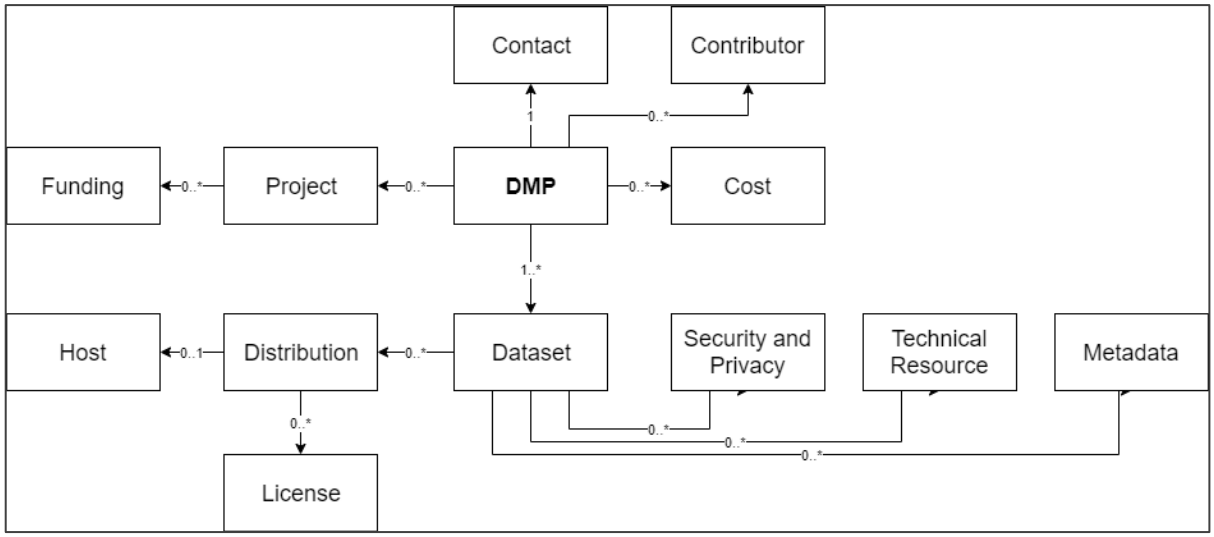
Den metadatastandard som idag finns och som har utvecklats till en de-facto standard för maDMP är RDA Common. Den är skapad och utvecklas av en arbetsgrupp inom Research Data Alliance (RDA), RDA WG DMP Common Standards, som bildades 2017. Den första versionen av denna standard publicerades 2019 och den senaste (1.1) 2021. Standarden är idag, mer eller mindre, implementerad i samtliga vedertagna system för hantering av datahanteringsplaner.

RDA Common bygger på tio grundläggande principer. Dessa finns publicerade i en artikel från 2019 (https://doi.org/10.1371/journal.pcbi.1006750).

I korthet så är målsättningen med den maskinläsbara datahanteringsplanen inte att den helt och direkt skall ersätta den traditionella datahanteringsplanen, utan snarare att möjliggöra utbyte av information mellan datahanteringen och relaterade intressenter och tjänster. För att på så sätt kunna skapa arbetsflöden som både underlättar och förbättrar för alla parter, samtidigt som god metadata, inklusive olika kontrollerade och persistenta identifierare kan medfölja metadatat under hela processen.
Det är en generell modell som är avsedd att kunna täcka en stor mängd behov. Den är inte anpassad till specifika system eller finansiärskrav och skall omfatta hela datahanteringens livscykel.

Det är möjligt att lägga till och använda egna, specifikt anpassade metadatafält. Det kan dock givetvis inte garanteras interoperabilitet med befintliga integrationer och tjänster.
Ett antal case studies och prototyper (s.k. mockups) för olika ändamål och intressenter finns. Dock saknas i dagsläget fortfarande praktiska tillämpningar.

Ett initiativ som dock bör nämnas är DAMAP (https://damap.org/). Ett nytt verktyg för datahanteringsplaner, skapat vid TU Wien, fullt kompatibelt med RDA Commons och utvecklat med särskilt fokus på integration med CRIS-system och andra relaterade tjänster. Hur väl detta system kan fungera i olika lokala miljöer, med olika förutsättningar och behov, återstår att se. Men det är hur som helst ett mycket intressant initiativ.



RDA Common 1.1, metadatamodell. https://rda-dmp-common.github.io/RDA-DMP-Common-Standard/

4.2 Chalmers

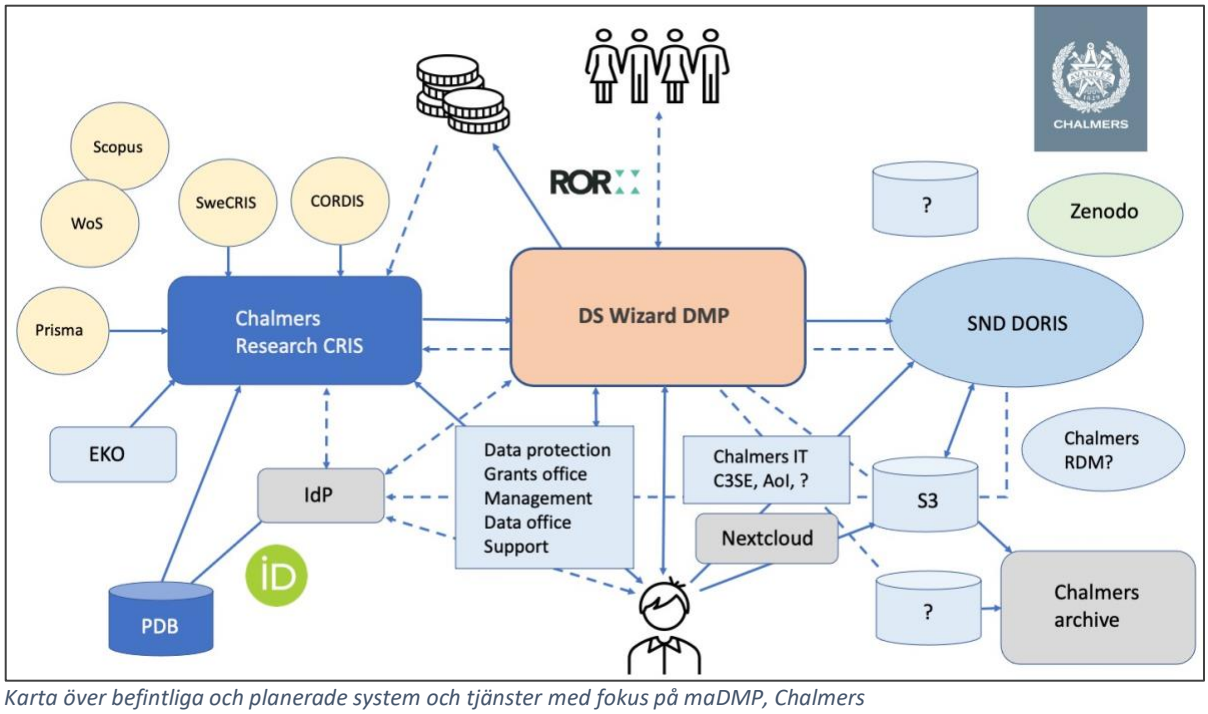
Chalmers data office (CDO) har funnits sedan 2020 och utgör DAU-funktionen på Chalmers. Det är en del av Chalmers eCommons, en infrastruktur för samordning av all datahantering på lärosätet (https://www.ecommons.chalmers.se).

CDO erbjuder sedan 2021 forskare tjänsten Chalmers DSW (https://dsw.chalmers.se) för upprättande och underhåll av datahanteringsplaner. Denna använder systemet Data Stewardship Wizard (se "System" nedan) som en lokalt installerad och i viss mån anpassad tjänst. I kommande riktlinjer uppmanas lärosätets forskare att upprätta datahanteringsplaner för samtliga forskningsprojekt, även sådana där finansiären inte explicit kräver detta. Nyttjandet av Chalmers DSW är då en rekommendation, om än inte ett absolut krav i detta sammanhang.

Sedan 2018 används ett lokalt och egenutvecklat CRIS (https://research.chalmers.se) för registrering och synliggörande av Chalmers forskningsoutput och relaterade aktiviteter, främst publikationer och information om forskningsprojekt. Här finns möjlighet att koppla publicerade dataset med publikationer och projektinformation, samt numer även möjlighet att länka projekt med datahanteringsplaner på forskarnas egna sidor.

För att underlätta upprättandet av nya datahanteringsplaner tillämpas fr o m hösten 2022 en rutin, där beviljade projekt från utvalda finansiärer (primärt VR och Formas) autogenererar nya planer, med kontrollerade metadata (inklusive identifikatorer) från Swecris, Prisma och Chalmers CRIS, samtidigt som forskningsprojektet skapas upp automatiskt i CRIS. Det sistnämnda blir då också en platshållare och ingång till datahanteringsplanen för forskaren, för framtida uppdateringar.

I framtiden planeras en än mer automatiserad process, där öppna api:er och maskinläsbara data nyttjas i högre grad för att underlätta för forskare och övriga intressenter. Som ett första steg kommer man att möjliggöra för forskare att själv initiera och påbörja en ny datahanteringsplan, med förifyllda data, från forskningsprojektets sida i CRIS.



Karta över befintliga och planerade system och tjänster med fokus på maDMP, Chalmers

Maskinläsbarhet nyttjas av Chalmers i dagsläget primärt för att, på ett tidigt stadium:

- fånga upp och identifiera särskilda behov av datalagring under och efter projekttiden, samt att få en uppdaterad bild av sådana behov generellt
- identifiera forskningsprojekt som kommer att hantera persondata eller andra skyddsvärda data. Sådan information skickas automatiskt som en alert till Chalmers dataskyddsombud och förväntas helt kunna ersätta befintligt fysiskt formulär.

4.3 Kungl. Tekniska högskolan (KTH)

KTH:s forskningsdataverksamhet bedrivs i form av ett forskningsdatateam med deltagare från olika delar av det gemensamma verksamhetsstödet på KTH. Forskningsdatateamet koordineras från KTH Biblioteket. KTH organiserar IT tjänster för forskning inom en delportfölj för forskning. Vad gäller IT-tjänster relevanta för datahanteringsplaner finns tjänsten DMP Online för att skriva datahanteringsplaner. DMP Online är tillgängligt för alla forskare som skaffar ett konto. Att KTH har ett eget konto innebär att de användare som kopplar sina konton till KTH får ett anpassat grafiskt gränssnitt och tillgång till KTH-specifika mallar. Personal på KTH med administratörsrättigheter kan utveckla mallar och se de datahanteringsplaner som skapas av KTH-anknutna användare.

Till skillnad från Chalmers har KTH inte samlat data om forskning från olika källor i ett CRIS-system, utan det är utspritt över ett flertal interna och externa system. Det blir därför inte möjligt att utveckla en funktion som gör att forskare kan "skicka" data från CRIS-systemet till DMP Online. Då KTH inte kan utveckla egna funktioner i DMP Online går det inte heller att enkelt införa en funktion i DMP Online för att hämta data från till exempel Prisma. Därför är det i nuläget svårt att göra en automatisk överföring av data mellan olika system som forskare kan välja att initiera. För administrativ personal är det däremot möjligt att göra sådana överföringar, antingen aktivt eller genom ett script eller program som gör det regelbundet.

4.4 Svensk nationell datatjänst (SND)

För att koppla ihop datahanteringsplaner med publicering av dataset så skall DORIS lista forskarens samtliga datahanteringsplaner, oavsett vilket system de skapades i. För att möjliggöra detta så har vi i gruppen undersökt möjligheten att använda RDA Common för att på bygga ett centralt index med datahanteringsplaner som kan listas för användaren.

Då datahanteringsplanerna och datapubliceringarna innehåller en stor del gemensamma, administrativa metadata så kan dessa användas för att automatiskt populera vissa fält för dataset som publiceras i DORIS. Detta kan ske utifrån lisning av datahanteringsplanerna där användaren klickar på en knapp för att skapa ett dataset utifrån en datahanteringsplan som listas i gränssnittet.

Det är viktigt att identifierare för personerna följer med i metadatan för datahanteringsplanen för att matchningen i DORIS ska gå att göra urifrån den inloggade användaren. DORIS har möjlighet att matcha identitet utifrån ORCID (om detta är kopplat när användaren loggar in) eller via e-postadress.

De fält som finns i RDA-dmp commons som kan användas för att populera fält vid publicering av ett dataset är:

- Contact
- Contributor

- Title
- Description
- Language
- Funding

Fältet *ethical_issues_exist* skulle även kunna användas som indikator på att fälten i DORIS som hanterar klassificering av personuppgifter/känsliga data behöver specificeras även om det blir svårt att göra en rak mappning då fältet inte matchar ett specifikt fält.

Viss modifiering och anpassning av den information som förs över från en maDMP till ett nytt dataset kommer förmodligen behöva editeras och kompletteras som t.ex. titel och beskrivning, men skulle fungera som en bra grund för att påbörja publicering av ett dataset.

Referensen från datahanteringsplanen är även den relevant att spara då vi har diskuterat att en framtida återrapporteringsfunktion skulle kunna byggas, där en berikad datahanteringsplan med en populerad lista över publicerade dataset skulle kunna genereras enligt RDA Common.

5. System

5.1 DMP Online

DMP Online är en tjänst som levereras av Digital Curation Center (DCC) och som bygger på mjukvaran DMPRoadmap (https://github.com/DMPRoadmap). Detta är en open sourcelösning som utvecklas och underhålls av DCC, i samarbete med University of Californa Digital Library, som också levererar en liknande tjänst med namnet DMPTool. I Sverige använder Lunds universitet sedan några år tillbaka en lokalt installerad version av DMPRoadmap. Ett flertal andra svenska lärosäten använder institutionella versioner av DMP online genom ett avtal mellan DCC och Sunet.

KTH har haft tillgång till en sådan version av DMP Online sedan 2020, först via ett eget avtal men sedan 2021 genom Sunets:s avtal (https://dmp.kth.se). Flera svenska och internationella lärosäten använder genom detta avtal samma system, alla med sina egna grafiska profiler och mallar. En fördel med att finnas i samma system som många andra är att det underlättar vid samarbeten med forskare från flera lärosäten. Genom att logga in med sitt lokala (eller ej lärosätesbundna) konto kan alla inblandade vid behov få tillgång till datahanteringsplanen. Systemet blir också bekant och därmed lättare att använda för fler.

KTH har via ett API åtkomst till metadata som KTH:s användare matar in i systemet, men kan inte kontrollera hur systemet fungerar och ser ut, utöver grafiska element och vissa hjälptexter. Detta, i kombination med avsaknaden av ett lokalt system för forskningsinformation, leder till att det inte finns ett naturligt ställe i KTH:s system för forskaren att initiera ett informationsutbyte mellan system. Om ett sådant utbyte skall ske så måste det istället initieras i ett system som enbart har denna funktion. Under dessa förutsättningar verkar det osannolikt att forskare på egen hand kommer att initiera ett sådant informationsutbyte – det måste i stället ske antingen automatiskt eller på initiativ av administrativ personal.

DMP Online som system lanserades 2010 och är alltså betydligt äldre än arbetet med att göra datahanteringsplaner maskinläsbara. Grundformatet, och version 0 av API-funktionaliteten, bygger på strukturen fråga-svar. Frågor kan konfigureras så att ett eller flera alternativ kan väljas, eller så kan de vara fritextfrågor. Det finns också ett antal maskinläsbara fält som kan användas i version 1 av API-funktionen. Dessa är inte kopplade till någon mall, och är av mer administrativ karaktär.

Mjukvaran som ligger till grund för DMP Online har också utvecklat en möjlighet att dokumentera dataset, något som är ett föräldraelement till många av de mer specifika egenskaperna i RDA Common standard (se bild ovan). DCC har dock ännu inte infört denna funktionalitet för dataset (”research outputs”) i den version av DMP Online som de levererar.

5.2 Data Stewardship Wizard

Den första versionen av Data Stewardship Wizard (https://ds-wizard.org) släpptes hösten 2016 och var resultatet av ett samarbete mellan Dutch Techcentre for Lifesciences (DTL) och Czech Technical University (CTU), inom ramen för ELIXIR - europeisk infrastruktur för livsvetenskaper. Den första instansen togs i drift våren 2018. Programvaran är open source och underhålls av ett utvecklingsteam vid CTU. Det finns även en advisory board och en växande användargrupp.

Systemet har en datamodell som är väl anpassad för maskinläsbarhet, men kan i nuvarande version (3.20) enbart leverera maskinläsbara metadata till andra tjänster på direkt initiativ av forskaren (genom s.k. ”push”). En mall för att transformera datahanteringsplaner till RDA Common 1.1 medföljer systemet och kan användas i standardutförandet med medföljande knowledge models.

Systemet har ett API, men detta är främst avsett att användas av interna systemfunktioner och relaterade systemnära tjänster. För att kunna använda det för automatisk leverans av standardiserade data – exempelvis för integration med andra tjänster - krävs i nuläget ett (lokalt) mellanlager som kan transformera det interna formatet till RDA Common. Vid Chalmers har för detta ändamål utvecklats en tjänst, där metadata hämtas i en automatiserad rutin från DSW API, konverteras till RDA Common och lagras i ett sökbart index där externa tjänster (som exempelvis DORIS) kan söka och läsa standardiserade datahanteringsplaner i ett kontrollerat API.
För att metadata skall göras tillgängliga i denna tjänst krävs ett aktivt godkännande av forskaren. En sådan “disclaimer” finns i de mallar som används i Chalmers DSW (se ”7. Utmaningar” nedan).

5.3 DORIS

SND:s dataorganiserings- och informationssystem (DORIS, https://doris.snd.gu.se) är ett nationellt system för att göra forskningsdata från svenska organisationer sökbara och tillgängliga, och som är utvecklas och underhålls av SND. En viktig del i detta system är de databeskrivningar som skapas (primärt) av forskaren när nya data publiceras. Metadata från dessa beskrivningar följer därefter med publiceringen i den DOI som skapas och är mycket viktiga både för synlighet och sökning av publicerade data. Att kunna kontrollera och återanvända sådana data när nya databeskrivningar skapas både underlättar för forskaren och säkerställer att viktiga uppgifter, som exempelvis identifierare, både existerar och dessutom håller hög kvalitet.

6. Resultat

6.1 Beskrivning av integrationen i DORIS

För flödet så har vi tittat på olika sätt för att kunna göra en integrering i DORIS. Modellen vi tagit fram är ett centralt dmp-index som populeras av de olika lokala systemen hos lärosätena. På så vis behöver DORIS endast skicka i väg en fråga för att se en komplett lista över DMP som en viss forskare är deltagare i. Forskare identifieras via ORCID eller e-postadress. Båda dessa identifierare bör finnas tillgängliga för DORIS i samband med att forskaren loggar in. I dagsläget

kan dock ORCID saknas som attribut i det som lokala autenticeringstjänster levererar (se ”7. Utmaningar” nedan).

I det centrala indexet lagras endast den informationen som går uttrycka enligt RDA DMP Common standard.

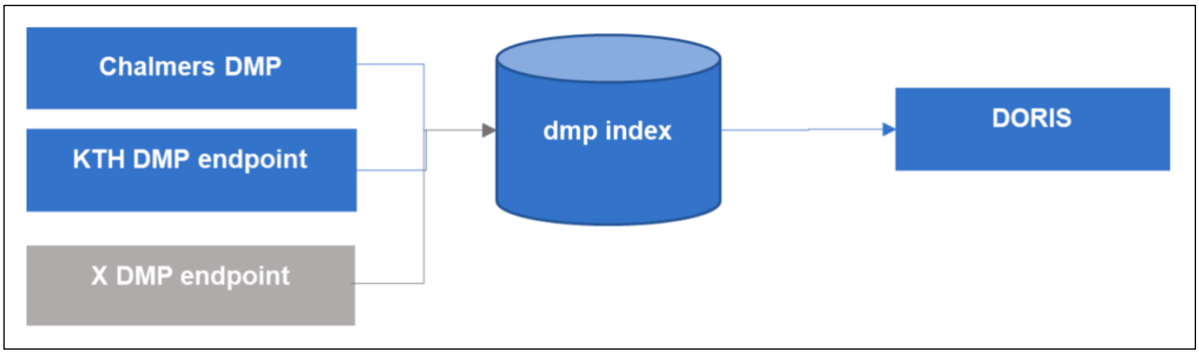


Bild över det tänkta flödet vid integration med DORIS (https://doris.snd.gu.se)

- Processen för att få in datahanteringsplanerna till det centrala DMP-indexet är tänkt att gå via daglig körning för att föra över nya publicerade datahanteringsplaner. Nackdelen med denna metod är att det blir en viss fördröjning från att användaren publicerar datahanteringsplanen tills att den blir synlig i DORIS. En alternativ push-metod skulle kunna införas som ett separat endpoint men bör ses som en frivillig implementation för varje lokal lösning.

- Det centrala indexet har i sin tur en endpoint som söker igenom samtliga skördade datahanteringsplaner utifrån orcid med e-postadress som alternativ matchning. Detta index skulle även kunna exponeras till andra intressenter på sikt, men DORIS blir den primära användaren.

- Vid listan i DORIS som visar användarens datahanteringsplaner presenteras en åtgärd för att skapa ett nytt dataset som tar med den informationen som finns i datahanteringsplanen, samt sparar identifieraren till datahanteringsplanen för att kunna presentera länkningen. Informationen som kopieras över kommer dock endast att kopieras över när datasetet skapas. Förändringar av informationen i datahanteringsplanen förs ej över till redan kopplade dataset då det kan ge oönskade effekter, då den informationen ska gå att modifiera fritt av användaren för att ge en så bra beskrivning som möjligt av datasetet.

7. Utmaningar

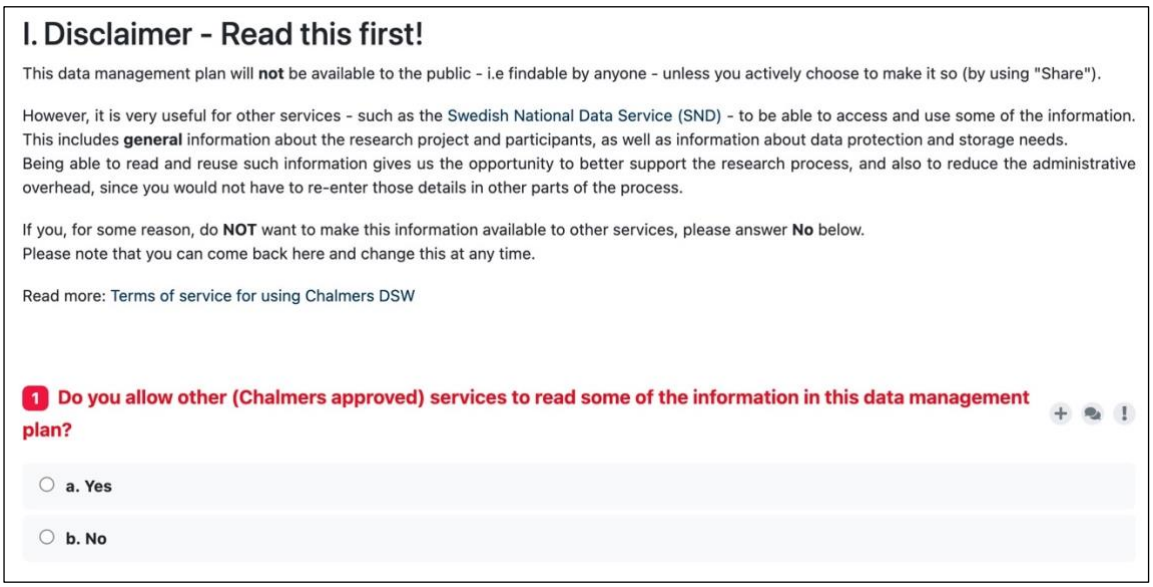
Vi har under projektarbetet identifierat svårigheter av lite olika karaktär, främst juridiska, organisatoriska och tekniska/systemspecifika.

7.2 Juridiska svårigheter

En datahanteringsplan är i grunden forskarens egendom och ansvar. Den innehåller normalt personuppgifter och skulle potentiellt även kunna innehålla andra känsliga och/eller skyddsvärda uppgifter
Detta kan kraftigt begränsa möjligheten att återanvända ingående metadata på ett effektivt (automatiserat) sätt, liksom möjligheten att lagra denna information på andra plattformar,

exempelvis i en temporär cache eller externt sökverktyg (för api-åtkomst). Möjligheten att skapa de sömlösa integrationer som vi önskar begränsas ju då också avsevärt.

Alla etablerade system inom området har möjligheten att låta forskaren ge andra intressenter behörighet att läsa och/eller modifiera informationen i datahanteringsplanen. På samma sätt skulle man också kunna ge exempelvis en DAU behörighet att återanvända informationen maskinellt i olika tjänster. Detta kräver dock en extra, aktiv handling av forskaren.
En bättre lösning är förmodligen att låta forskaren ge (eller neka) denna behörighet genom ett enkelt avtal ("disclaimer") som ingås när planen skapas, och som kan uppdateras vid behov. I Chalmers DSW har denna möjlighet lagts till som en första fråga när en ny plan upprättas.



Disclaimer från Chalmers Data Stewardship Wizard (https://dsw.chalmers.se)

DMP Online har en uppsättning användarvilkor som gäller för alla användare av DMP Online, och som tillåter viss återanvändning av de data som läggs in i systemet. Användarvillkoren är dock, som sådana texter ofta är, en text av mer juridisk karaktär som kan verka svårgenomtränglig. För att göra detta tydligt har KTH bett DCC att uppdatera första sidan för KTH:s DMP Online, dvs den sida som användaren ser innan ett konto skapas och innan inloggning. När detta har implementerats kommer texten att lyda:

This instance of DMP Online is provided by KTH Royal Institute of Technology to help you write data management plans. Create an account to access guidance, advice, and templates from KTH Royal Institute of Technology. Once you have an account you can link this to your university account and log in with your KTH credentials. KTH DMP Online will store and use the information you enter according to the terms of use.

Do not enter information about high security projects into KTH DMP Online. If you need a DMP for such a project, feel free to log in, create a test plan and download the questions from the appropriate template. Use the preferred template and reply to the questions in order to write a DMP using your preferred writing tool and storing it in a secure location.

The information you enter into KTH DMP Online will not become available to the public unless you choose to make it so. However, KTH Research Data Services can access the information (general information such as project description, participating researchers, as well as more specific information about needs for support and storage). This information can be used for support purposes, but also for reuse in administrative systems to simplify and provide better services.

If you do not agree with this, please log in and create an empty plan to download the questions from your preferred template. You can then use the template and write down your answers to the questions outside the system.

If you have questions about research data management at KTH, please contact researchdata@kth.se

7.3 Organisatoriska svårigheter

I första hand på KTH finns det metadata/data som behövs för att automatiskt populera en maskinläsbar datahanteringsplan inte samlat, utan är utspritt på flera olika system med olika nivåer av interoperabilitet och varierande kvalitet på data.
På Chalmers är detta data i större utsträckning samlat i det lokala CRIS-systemet, vilket underlättar skapandet av integrationer.

7.4 Teknik och systemstöd

Som nämnts ovan så är det primära syftet med maDMP inte att helt ersätta den mer traditionella datahanteringsplanen, i bemärkelsen omfatta samtliga frågor eller all kringinformation om datahanteringen som en sådan kan innehålla. Syftet är snarare att kunna utnyttja värdefull information i datahanteringsplanen för att automatisera och effektivisera delar av processen. Exempel på sådana delar skulle kunna vara att automatiskt bevaka och fånga upp särskilda behov, exempelvis när ett forskningsprojekt hanterar känsliga, mycket stora eller särskilt skyddsvärda data. Viss information skulle också kunna återanvändas i andra system, för att undvika dubbelarbete.

Standarden ger även möjlighet att ange information om dataset som skall delas. Detta är dessutom definierat som ett obligatoriskt fält, vilket kan vara ett problem då man ju inte alltid har, eller borde behöva ha denna information i en datahanteringsplan. Det leder kanske också in i en större fråga om vad en datahanteringsplan är (och bör vara). Frågan om huruvida det är relevant och rimligt att en plan i förlängningen skall innehålla detaljerad information om enskilda dataset kan diskuteras. Detta speglar möjligen en övertro på forskarnas vilja och behov, liksom systemstödet, alternativt kanske man tänker sig en framtid där sådan information skulle kunna tillföras och extraheras från datahanteringsplanen på ett mer automatiserat och intelligent sätt än idag. I nuläget är dock lösningen normalt att skapa ett tomt element för dataset, för kompabilitet med standarden.

Av information som *saknas* i nuvarande version kan främst nämnas uppgifter om behoven av datalagring under (och efter) projekttiden. Fältet ”cost” finns, men detta är främst avsett för att dokumentera – eller automatiskt beräkna - kostnaderna för lagringen, inte direkt storlek och egenskaper hos lagringsytorna. Det senare ser vi dock som en av de viktigaste delarna i datahanteringsplanen, åtminstone initialt. Det är dessutom en information som borde vara både enkel att ange och att hantera maskinellt. I Chalmers DSW har denna fråga lagts till i både lokala standardmallar och som ett lokalt definierat fält i maDMP.

Att koppla ihop olika system med datahanteringsplaner förutsätter mallar med frågor som i hög grad är strukturellt anpassade för mappning mot fälten i standarden. Detta innebär i korthet färre frågor som efterfrågar svar i beskrivande fritext och fler i formen av ja/nej och val mellan olika alternativ. I synnerhet gäller detta de delar som i standarden förväntas kunna trigga olika mekanismer i andra system. Exempelvis frågor om datasäkerhet.

I såväl DMP online som DS Wizard kan frågemallar anpassas relativt fritt.
En särskilt anpassad mall för DMP online har arbetats fram av Joakim Philipsson, SU. Denna kan användas som underlag för att skapa egna mallar som är optimerade för maskinläsbarhet. På KTH kommer vi under 2023 att titta på hur vi kan anpassa den för KTH.

För att kunna nyttja fördelarna med maskinläsbara datahanteringsplaner är det också mycket viktigt att detta finns med som ett behov även vid utveckling av lokala system och tjänster för datahantering. Det är önskvärt att alla sådana kan leverera maskinläsbara data enligt befintlig

standard. För att kunna automatisera och effektivisera maximalt bör dessa kunna tillhandahållas genom ett **sökbart API**, där aktuella datahanteringsplaner kan identifieras och sökas fram med hjälp av exempelvis:

- Forskares personliga id (ORCID, lokalt användar-id, e-post) som korresponderar med de identifierare som returneras av lärosätets IdP vid inloggning i DORIS eller andra tjänster
- Finansiär, standardiserat namn/akronym och identifierare (t ex ROR), i kombination med
- Projekttid

En sista teknisk (och möjligen organisatorisk) utmaning som vi vill beröra är relaterad till den federerade inloggningen i DORIS (SWAMID). För att optimalt kunna matcha inloggade användare (forskare) och befintliga datahanteringsplaner bör såväl e-post som ORCID finnas med bland de attribut som returneras av de lokala lärosätenas autenticeringstjänster (IdP). I nuläget returneras dock inte alltid ORCID av alla lärosäten, men det är önskvärt att så kan komma att ske i framtiden. Detta kräver både att detta ID finns registrerat och tillgängligt i respektive lärosätes system och att det skickas tillbaka till den anropande tjänsten vid inloggning.

8. Slutsatser och vägen framåt

Användandet av maskinläsbara datahanteringsplaner ger stora möjligheter till automatisering och därigenom effektivisering av många centrala processer i datahanteringen. Det kan också bidra till en högre metadatakvalitet, då möjligheter finns att integrera med och hämta kontrollerade metadata från andra system och tjänster, snarare än att helt förlita sig på manuell inmatning i olika delar av processen.

Om den maskinläsbara datahaneringsplanen används för att publicera dataset så kan en utökad maskinläsbar datahanteringsplan genereras där maskinläsbar metadata om dataseten ingår. Denna maskinläsbara metadata skulle kunna användas för att få en helhetsvy över forskningsprocessen både för forskaren, men även för det lokala forskningsstödet på lärosätet och därigenom också underlätta utvärdering.

Man bör vara medveten om de begränsningar som finns och framför allt att den maskinläsbara datahanteringsplanen inte, åtminstone inte i nuläget, avser att vara heltäckande, dvs en ersättning för den tradionella datahanteringsplanen. Snarare skall den ses som en möjlighet till effektivisering och förbättring av vissa centrala processer. Begränsade lokala förutsättningar och rutiner, såväl tekniska som organisatoriska, kan givetvis också begränsa det fulla nyttjandet av dessa effektiviseringar och förbättringar.

För att nyttja maskinläsbarhet optimalt så måste frågemallar vara anpassade till standarden. Detta innebär i de flesta fall färre frågor av beskrivande karaktär (fritext) och fler alternativfrågor som lättare kan tolkas och hanteras maskinellt. Man måste också beakta behörighetsfrågan när det gäller att kunna läsa och återanvända metadata. En datahanteringsplan är i normalfallet att betrakta som forskarens ansvar och egendom. Den skulle dessutom kunna innehålla känsliga uppgifter. I vanliga fall kan man åtminstone utgå ifrån att den innehåller personuppgifter. Detta måste hanteras primärt av de tjänster som tillhandahåller dessa uppgifter, exempelvis (som i Chalmers fall) med någon form av godkännande från forskaren.

Om maskinläsbara datahanteringsplaner skall kunna fylla sin funktion så krävs naturligtvis också det att det finns och skapas tjänster och rutiner som erbjuder denna möjlighet. I dagsläget är sådana tjänster mycket få. Detta beror givetvis främst på att det är en ny standard, som

fortfarande är under utveckling, men också på att befintliga rutiner och arbetsflöden fortfarande är outvecklade och, i den mån de finns, fortfarande manuella. Ibland dessutom av nödvändighet.

Likväl är det mycket viktigt att beakta maskinläsbarhet vid införande och/eller utveckling av nya systemlösningar, inklusive lokalt utvecklade tjänster. Alla sådana bör kunna leverera och ta emot metadata i den utsträckning som behövs för att minimera onödiga administrativa steg för forskare och annan personal, samtidigt som det upprättas tillräcklig dokumentation för att både kunna användas i lärosätenas administrativa processer och för att stödja forskarna i en transparent och högkvalitativ forskningsprocess.

I detta sammanhang är det viktigt att också ställa den större frågan om vad en datahanteringsplan är och vad den syftar till, utöver eventuella finansiärskrav och lokala policies. Man bör kunna identifiera de största behoven samt hur och om maskinläsbarhet kan utnyttjas för att underlätta och förbättra för involverade parter. I Chalmers fall nyttjas exempelvis maskinläsbarhet i dagsläget (endast) för två syften: att fånga upp särskilda behov av datalagring samt att på ett tidigt stadium identifiera forskningsprojekt som hanterar persondata och andra skyddsvärda data. Det är därför viktigt att denna information kan anges i frågemallarna och även kan levereras enligt standarden för samtliga datahanteringsplaner. Ett agilt förfarande vore kanske att utgå från sådana konkreta behov, för att över tid utöka efter nya och förändrade behov.

Implementationen i DORIS är en praktisk tillämpning av maskinläsbarhet och RDA Commons. Den kan tjäna som ett gott exempel på hur detta kan användas för att förenkla och förbättra ett befintligt arbetsflöde, i det här fallet vid publicering av forskningsdata.
Ett nästa steg skulle kunna vara en lösning för att initiera en ny datahanteringsplan, förpopulerad med grunddata från externa system, som exempelvis Swecris och lokala system, med hjälp av maDMP.

Litteratur och resurser

Cardoso, J., Castro, L.J., Ekaputra, F.J. *et al.* DCSO: towards an ontology for machine-actionable data management plans. *J Biomed Semant* 13, 21 (2022). https://doi.org/10.1186/s13326-022-00274-4

DAMAP : a tool for machine actionable DMPs
https://damap.org/

Data Stewardship Wizard
https://ds-wizard.org/

DMP Online [Digital Curation Center (DCC)]
https://dmponline.dcc.ac.uk/

DMP Roadmap [software]
https://github.com/DMPRoadmap

Miksa, T, Walk, P, Neish, P, Oblasser, S, Murray, H, Renner, T, Jones, S. (2021) Application Profile for Machine-Actionable Data Management Plans. *Data Science Journal*, 20(1), 32.
http://doi.org/10.5334/dsj-2021-032

Miksa, T, Oblasser, S and Rauber A (2021) Automating Research Data Management Using Machine-Actionable Data Management Plans. *ACM Trans. Manage. Inform. Syst.* 13, 2, Article 18 (December 2021), 22 pages. https://doi.org/10.1145/3490396

Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. *PLoS Comput Biol* 15(3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750

Oblasser, S, Miksa, T and Kitamoto, A (2022) Finding a Repository with the Help of Machine-Actionable DMPs: Opportunities and Challenges. *International Journal of Digital Curation* 15(1). https://doi.org/10.2218/ijdc.v15i1.704

Miksa, T, Walk, P and Neish, P. RDA DMP Common Standard for Machine-actionable Data Management Plans. http://doi.org/10.15497/rda00039

Philipson, J (2022) "SU-EOSC Nordic 5.3.2 maDMP project", https://doi.org/10.7910/DVN/MGZBAL, Harvard Dataverse, V1

RDA DMP Common Standard for machine-actionable Data Management Plans https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard

RDA DMP Common Standards Workgroup https://www.rd-alliance.org/groups/dmp-common-standards-wg

Simms S, Jones S, Mietchen D, Miksa T (2017) Machine-actionable data management plans (maDMPs). Research Ideas and Outcomes 3: e13086. https://doi.org/10.3897/rio.3.e13086