SND flaggskeppsprojekt Spårbara forskningsdata – metadata för data-användning

Rosa Lönneborg (KTH), ORCID: 0000-0002-7679-1369

Projektdeltagare: Urban Andersson (Chalmers), Therese Tikkanen (Chalmers) Stina Johansson (GU), Mattias Vesterlund (KTH) , Agne Larsson (KTH) + stöd från André Jernung & Olof Olsson (SND)

Re-cap: Tidigare flaggskeppsprojekt

- Tidigare projekt på **temat spårbara forskningsdata**
  - 2022-23 Maskinläsbara DMP:er – hur kan information flöda genom ett DMP verktyg? – Återanvändning av öppna administrativa metadata (ex projektinformation från SweCRIS, ytterligare information från DMP för mer spårbara lärosätesinterna processer)
  - 2024-25 Var finns lärosätenas publicerade forskningsdata? Hur kan de spåras?

- Rapporter och digitala outputs 2022-2025:

Mjukvara maDMP: https://github.com/snd-sweden/dmp-scripts
Mjukvara (under utveckling) bakom ROAGG skördning: https://github.com/snd-sweden/roagg

Presentation av projekt 2025: https://doi.org/10.5281/zenodo.19094219

Rapport 2025 – rekommendationer för ökad spårbarhet: 10.5281/zenodo.17365329
Rapport 2025 - gapanalys: 10.5281/zenodo.18846122

Rapport 2023 - Open data flagship pilot 2022 : slutrapport

Årets projekt – bygger vidare på tidigare

Går det att spåra om forskningsdata som ligger till grund för publicerade resultat tillgängliggörs med öppen tillgång? Tillgängliggörs den enligt FAIR-principerna?

Utdrag ur tidigare rapport om gapanalys:

"I vilken utsträckning forskningsdata faktiskt publiceras i Sverige är dock svårt att svara på – bland annat då **det saknas effektiva metoder** för att spåra och följa upp sådana publiceringar.

Uppföljningen försvåras bland annat av **brister i angivna metadata för organisationsaffilieringar och klassificering av material**. Den kompliceras också av **skillnader i rutiner för hur forskning redovisas lokalt och nationellt.**

För att uppnå bättre spårbarhet hos forskningsdatapubliceringar krävs förbättringar av såväl nationell samordning och teknisk infrastruktur som incitamentsstrukturer."

Projektets syfte och delfaser 2026

- Som tidigare flaggskepp - samarbete KTH, Chalmers och GU.
- **Syfte:** Att utarbeta processer för leveranser av metadata till en nationell datapubliceringsplattform med stöd av kartläggning av nuvarande metadataflöden i det digitala landskapet för publicering.
- **Avgränsning:** Projektets fokus är att spåra inomvetenskaplig användning av data /digitala research outputs*, vi analyserar inte bredare användning i samhället.
- Arbetet planerat i 3 olika delfaser

De olika delfaserna

- Del 1: Hur citeras/anges forskningsdata som ligger till grund för publicerade resultat från Sveriges lärosäten idag? Är den forskningsdatan tillgängliggjord enligt FAIR?
- Del 2: Metadata & PID:ar –övergripande kartläggning av ansvar för processer och flöden mellan olika forskningsaktörer för att uppnå högre spårbarhet i linje med nationella rekommendationer.
- Del 3: Vilka skäl finns till att existerande större samlingar forskningsdata inte i högre utsträckning är tillgängliggjorda enligt FAIR?

Forskarens perspektiv: användare av data i forskningen

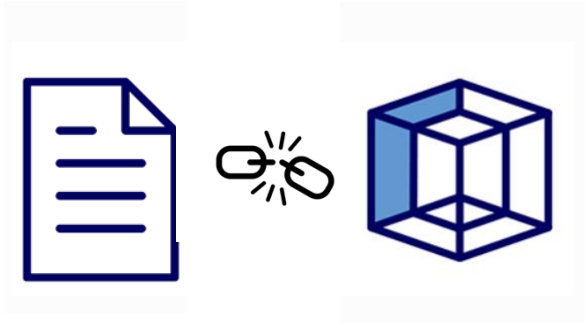
Finns data? Var?
Är den i användbart skick?
Är det tillförlitlig data?
Vem står bakom källan?

Hur får jag använda datan? Var och hur kan jag lagra & analysera den? Referera till den?

Hur tillgängliggör jag mina digitala outputs? Hur f-n ska jag hinna det och varför?

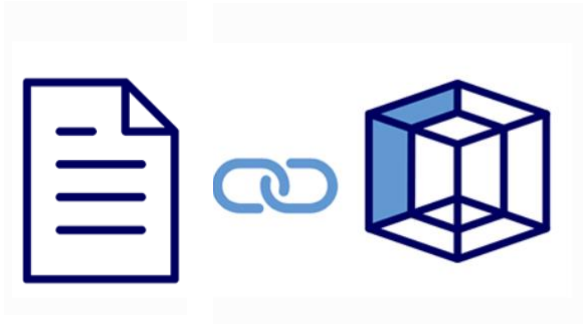
Vetenskaplig artikel – **hur ange data-användning ?**

Hur anges forskningsdata i publicerade artiklar?



Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.



Repository: Manually annotated miRNA-disease and miRNA-gene interaction corpora.
https://doi.org/10.5256/repository.4591.d34639.

This project contains the following underlying data:
- Data file 1. (Description of data.)
- Data file 2. (Description of data.)

Data are available under the terms of the Creative Commons
waiver (CC0 1.0 Public domain dedication).

Figure 1: Data citation example.

...highly site specific, potentially limiting their wider value. However, applying the approach as conducted in this paper to data such as that presented by Barnett et al (2013) to (1) ve relative values for different organisms should provide a more generic set of 'reference data'. In taking the REML approach forward it will be beneficial to target...

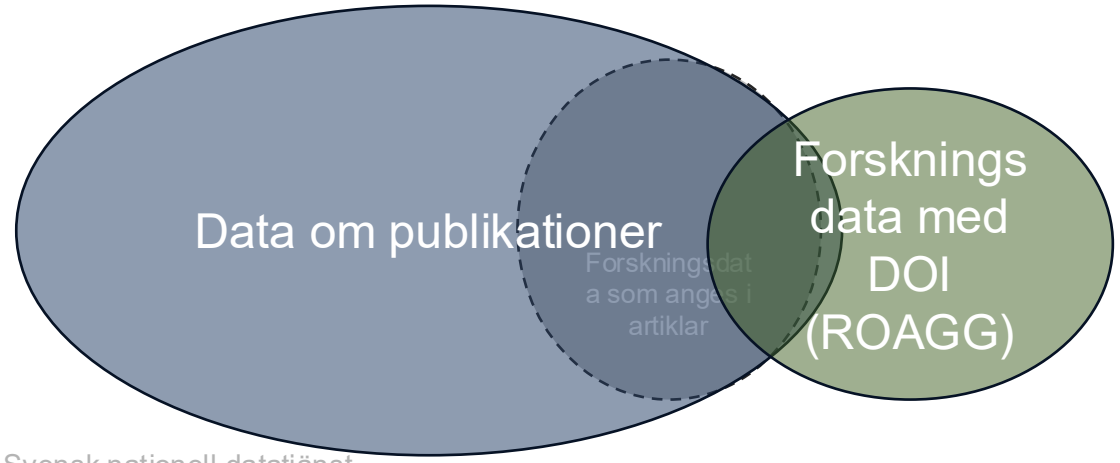
References

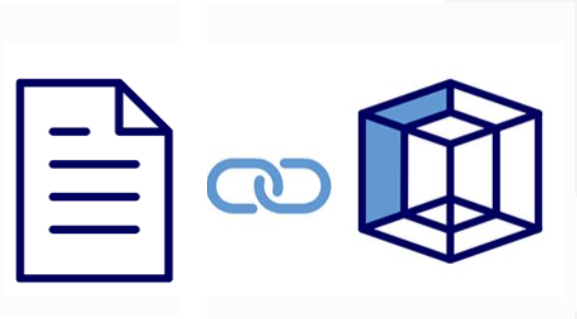
Barnett et al., 2013 C. (2) arnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone Element and radionuclide concentrations in representative species of the ICRP's reference animals and plants and associated soils from a forest in North-west England. NERC - Environmental Information Data Centre (3) http://doi.org/10.5285/e40b53d4-6699-4557-bd55-10d196ece9ea

Figure 1. Data citation example derived from: Cousijn, H., Kenall, A., Ganley, E. *et al.* A data citation roadmap for scientific publishers. *Sci Data* **5**, 180259 (2018). https://doi.org/10.1038/sdata.2018.259

Status –arbete pågår med delfas 1:

- Hur anges forskningsdata i publicerade artiklar?
  - Test och jämförelse av olika metoder för text-mining för att spåra användning av data via omnämnande av dataset/digitala research outputs i publicerade vetenskapliga artiklar.
  - Metoderna testade på mindre test-data mängder
    - Nästa steg: Tvätta, normera och validera större data-set med fulltext-artiklar för bättre representation av olika forskningsfält.
    - Jämför med ROAGG-dataset – vilken forskningsdata citeras?
  - Rättsliga hänsynstaganden och kvalitetssäkring av fulltext-dataset – liten insikt i sig.





Rättsliga hänsynstaganden och kvalitetssäkring av fulltext-dataset

- Diskussion om licensierad tillgång till fulltext-artiklar. Två vägar:
  - A) enbart använda OA-fulltexter
  - B) OA samt icke OA fulltexter för mer fullständigt dataset
  - A eller B ? Det beror på avvägning mellan riskaptit och kvalitetskrav...
  - A innebär att vi är juridiskt safe. Men kan vi enbart använda OA-artiklar med tydligt angiven CC-BY licens och få tillräckligt bra dataset för representation av olika forskningsdomäner? Kanske
  - B innebär något mindre "juridiska säkerhetsmarginaler" Det är *troligtvis* ok att använda ALLA fulltexter, så länge vi inte tillgängliggör data-dumpen publikt och processar den utan att data läcker.

  Om vi väljer alternativ A :
  Fördel: icke-konfidentiellt material, dataset kan tillgängliggöras, ingen juridisk risk. **Nackdel:** Risk att humaniora och konstnärlig forskning blir under-representerat i det normerade & validerade data-setet.

  Hinder: CC-BY licensen behöver vara tillgänglig i maskinläsbart format för återanvändning - men DiVA anger inte detta. Då behöver vi "kors-referera" med OpenAlex/CrossRef som har metadata-fält om OA, men kvaliteten?

- **Insikt:** om publikationsdata i Sverige funnits tillgängligt via API med både öppen tillgång och maskinläsbar CC-BY licens med API-åtkomst hade projektet förenklats avsevärt.
- OM vi dessutom haft tillgång till gemensam nationell tjänst för virtualiserad behandlingsmiljö för att testa metoderna också på konfidentiellt material hade det varit ännu bättre.

Planering delfas 2 & 3

- Del 2: Metadata & PID:ar –ansvar och flöden mellan lärosäte & forskningsinfrastrukturer & den nationella forskningsdataportalen.
  - Vilken metadata och processtöd behövs för att:
    - Tydliggöra proveniens och spårbarhet? Ansvarsfördelning för sådana processer?
    - Säkerställa data-kvalitet? Ansvarsfördelning för sådana processer?
    - Möjliggöra rättsäkert tillgängliggörande av olika typer av digitala forskningsoutputs?
- Del 3: Vilka skäl finns till att större mängder insamlad forskningsdata inte i högre utsträckning är tillgängliggjorda på ett sätt att de kan återanvändas?
  - Intervjuer med ansvariga för större samlingar forskningsdata om de hinder de upplever för högre grad av tillgängliggörande.
  - Identifiera kontaktytor mellan forskare, lokalt stöd, infrastrukturellt och nationella aktörer och EOSC
  - Identifiera avsaknad av incitament eller infrastruktur för tillgängliggörande

Fråga till publik

- Finns det dataset eller större samlingar av data på ditt lärosäte som skulle kunna vidarenyttjas i forskning men där det arbete som krävs för att bearbeta och tillgängliggöra det för forskningsändamål inte har gjorts?
  - Har ni uppfattning om vad som är hindrande för att genomföra arbetet?