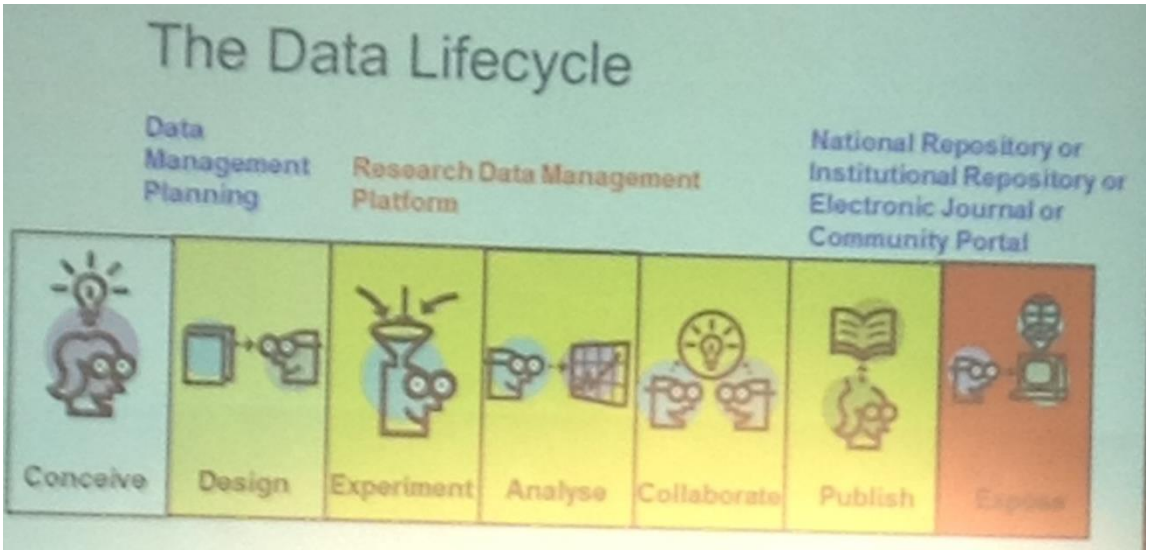
«Dealing with data – what's the role for the library?»

Rapport fra deltagelse på Joint OpenAIRE/LIBER workshop i Gent, Belgia 28. mai 2013.

Programmet var helhetlig oppbygd med tanke på å gi en komplett oversikt over utfordringer ulike aktører møter i arbeidet med forskningsdata og gikk fra forsker til publiseringsinstansene, videre til institusjonene og til sist innom ulike verktøy til bruk i arbeidet med forskningsdata. Det hele ble avsluttet med en paneldebatt som jeg dessverre ikke fikk med meg fordi jeg måtte rekke siste flyet tilbake til Oslo den kvelden.



Enda en variant av «the data lifecycle»

Keynote Erik Mannens begynte dagen med foredraget «What is open science, a researcher´s perspective", dette var minst spennende da det var et noe hoppende og overfladisk sammendrag av litt om alt og ingenting innen open science, open access og open data. Mannens bakgrunn er arbeid med linked open data (LOD) og foredraget gav derfor et interessant innblikk i hvordan noen som er engasjert i LOD ser et potensiale i deling av forskningsdata. Han nevnte også noe sitt arbeid med R&Wbase: Git for tripples

Research Data Overview by Sarah Callaghan 'A step by step guide through the research data lifecycle, data set creation, big data vs long-tail, metadata, data centres/data repositories' ga et godt innblikk i hvordan arbeid med data kan være. Callaghan jobber for The British Atmospheric Data Centre og delte av sine erfaringer herifra og fra tidligere arbeid med forskningsdata. Hennes fokus var på at "analysering og analyse prosessen må frem, det er dokumentasjonen her som gjør resultatene reproduserbare". Meteorologiske data er ikke reproduserbare fordi været er i konstant endring, det er også umulig å spore tilbake konteksten en måling er gjort i om denne ikke er registrert → lengede og breddegrader →definisjoner og relasjoner var viktige stikkord i hennes arbeid med metadata. Fokus på at et repository må ha en workflow, ikke bare fungere som en «data dumper» der professorer som nærmer seg pensjonsalder levere fra seg en pakke og går. Som eksempel på at lagring ikke gir mening om kunnskapen til å lese og forstå innholdet ikke lenger er tilstede brukte hun Phaistios Disk dette er i følge Callaghan kun av interesse for «data arkeologer».

Callaghan´s deffinisjon på et datasett:

- Result of a defined process
- Scientifically meaningful
- Well-defined (clear what is inn and what isn't)

Videre mente hun at et kriterie for lagring er at et datasett og metadataene må være komplett, først da gir det mening å bruke DOIs og sitere datasettene I videre bruk.
«Conclusions are only as good at the data they are based on. It you can´t dig into the data you can`t find out if the data is rubbish"

"Meet the scientists" A Researcher's perspective var en panelsesjon med biolog Aaike de Wever (biofresh), historiker og språkforsker Joris Van Zundert og professor i markedsføring (analytical consumer relationship managment) Dirk Van den Poel. Gjennom å beskrive sin bruk og lagring og arbeid med forskningsdata fikk de frem ulike problemstillinger knyttet til åpen lagring.

Aaike de Wever:

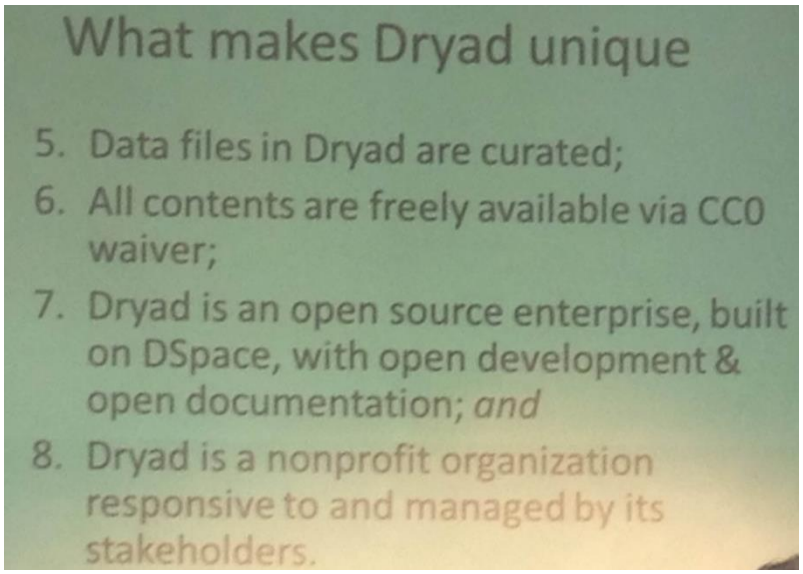
Biofresh hor harn jobbet samlet og lagret data om vann kvalitet, resultat av prøver tatt i ferskvann for å kartlegge forurensning. Han var veldig opptatt av rent vann. For dataene han jobbet med er «hva, hvor, når og av hvem» ekstremt viktig informasjon, både for å gi en forståelse av hvor prøvene kommer fra, men også for å gi de troverdighet, av hvem er derfor like viktig for hvor og når
Han var også inne på «distrust of own data» som en viktig faktor for forskerne, de må være kritiske i alle ledd for å oppdage mulige feil eller faktorer som kan gi misvisende resultat.

Dirk Van den Poel:
Jobber mye med data på våre forbruksvaner, disse er det vanskelig å gjøre tilgjengelig i dag pga. personvernhensyn, siden det ofte også etter anonymisering vil være mulig å spore personene bak dersom dataene kombineres på rett/feil måte med andre data. Beskriver tilgangene han har som veldig lukket, samtidig som det stadig kommer krav om deling av data og programvare fra tidsskrifter. Fokuserer på at han ikke bare beskriver dataene, men bruker de som grunnlag for modeller som forsøker å forutsi mønster.
«arbeidsflyten/prosesseringen/data innsamlingen må dokumenteres med fokus på sporbarhet og kode i prosessen.»

Joris Van Zundert:
Lagrer sine data på egne servere, kunne brukt DANS men gjør pr. i dag ikke det.
Til tider kan det også være uklart hva som er dataene, fordi disse kan være komplekse og av svært varierende art. Understreket behovet for en formalisering av forskningsprosessen, innenfor humaniora der tolkning og analyse ikke nødvendigvis er formaliserte prosesser.
Ting som bør være på plass er rutiner for kreditering av «data skaperne(creators)» og API tilgang.

The data publishing process: presentasjon av Dryad som er en non-profit medlemsorgansiasjon som tilbyr lagring/publisering av og Geoscience data journal fra Wiley

Dryad:
Dryad tilbyr lagring av data og programvare som er forbundet med publikasjoner – i utgangspunktet innen området evolusjonsbiologi og økologi, men i dag også brukt av andre fagområder innen naturvitenskap og medisin, men med fokus på biovitenskap. Dryad fokuserer på å gjøre datasett gjenfinnbare, gjenbrukbare og siterbare. Fordelene de gir er: synlighet, siterbarhet, arbeidsbesparende, bevaring. I best practise guidelines finnes eksempel på datasitering og måter å gjøre dette på. De ser hvor i artikler referansene til datasettene ligger, det er tydelig at det ikke er en ens praksis, mange har et dedikert «data kapittel» andre ganger er det nevnt i forord, etterord, referanser, eller mitt i teksten, dette viser at det er behov for å enes om en standard praksis på datasitering. Dryad anbefaler at det legges referanse både til publikasjonen der dataene først er brukt og til datasettet. Arkivet er Dspace basert.



Spesielt for Dryad satt det "curators" og gikk gjennom alt som ble lagt inn. Noe som må være en kjempejobb og ikke minst dyrt. Dette minner oss om kostnadasfaktoren og det faktum at det kan kreve en god del penger å gjøre arkiv/deling til en god forskningsressurs, det er ikke bare å sette opp en løsning som fylles opp, men en kurator rolle er også nødvendig.

Geoscience data journal:
Argumenter for beskrivelse av forskningsprosessen og siterte Susan Riley med krav om «Data accessebility statement» «opportunities for data exchange»

Data Management training/Data håndterings kurs

Robin Rice presenterte MANTRA og arbeidet med å samle data management resurser online ved Edinburgh University, Ellen Verbakel presenterte 3tu.datacentrum and data inteligence for librarians – kurset i Delft og Andrew Cox fra University of Sheffield presenterte RDMRose. Alle tre tilbyr resurser for arbeid med forskningsdata på sine nettsider, og det er også mulig å følge kursene ved universitetene, alle tre kursene er på engelsk. Jeg har tidligere vært i kontakt med Delft med tanke på å kurset de tilbyr der, men ser at også de andre tilbudene kan være interessante.

Services and tools – Zenodo og cKan

Zenodo tilbyr lagring og tilgjengeliggjøring i et åpent arkiv som er et samarbeid mellom OpenAIRE og CERN. Zenodo er finansiert med prosjektmidler, der er derfor usikre på fremtiden, men lover å sørge for at data som leveres bevares vedlikeholdes og tilgjengeliggjøres for ettertiden. Foreløpig er det kun mulig å lisensiere med cc0. Også Zenodo går til en viss grad igjennom dataene som leveres inn, men ikke alt, det de sa var at «dataene dine blir ikke presentert på førstesiden uten at vi har gått igjennom de».

cKan er en programvare som er mye brukt i arkivering av forskningsdata

Andre lenker jeg noterte meg i løpet av dagen er:

SIM4RDM

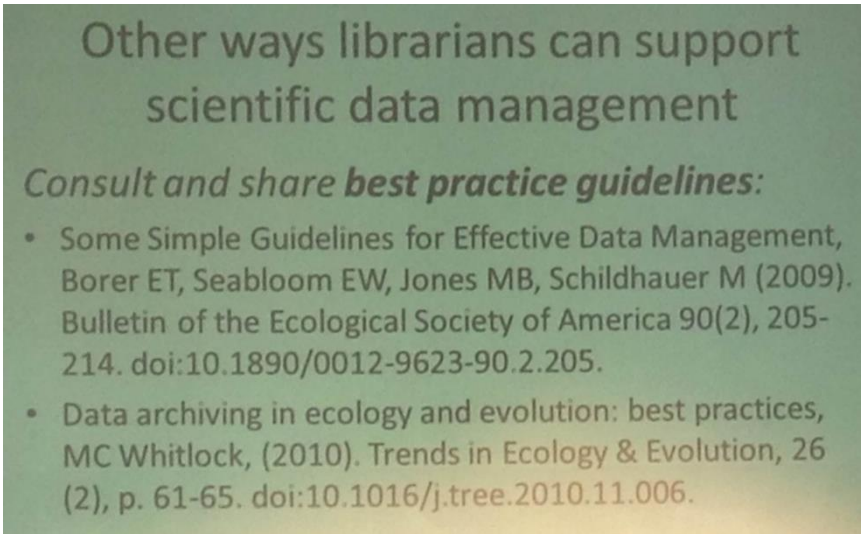
PREPARDE

Data citation index

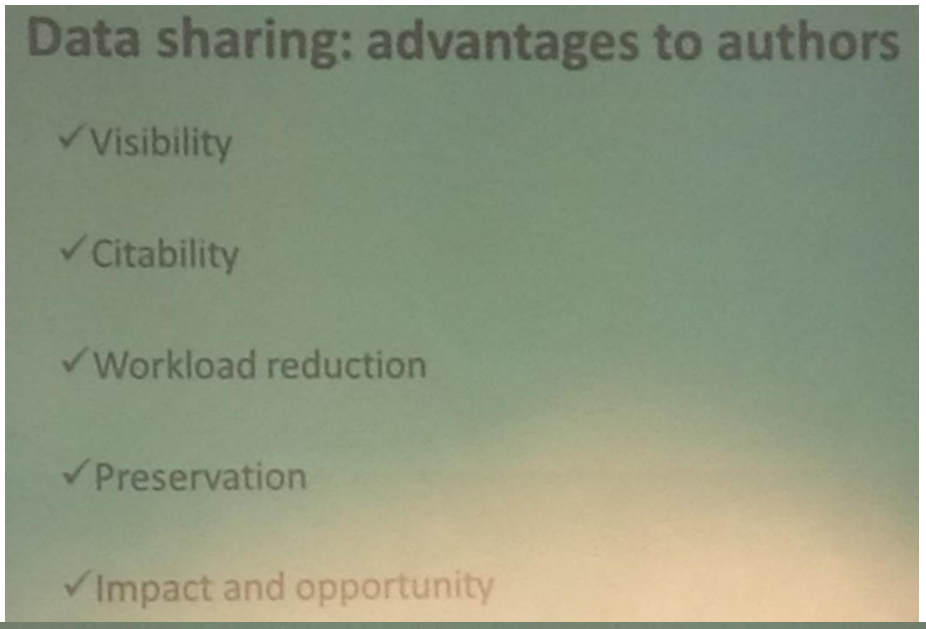
JoRD – Journal Research Data policy bank

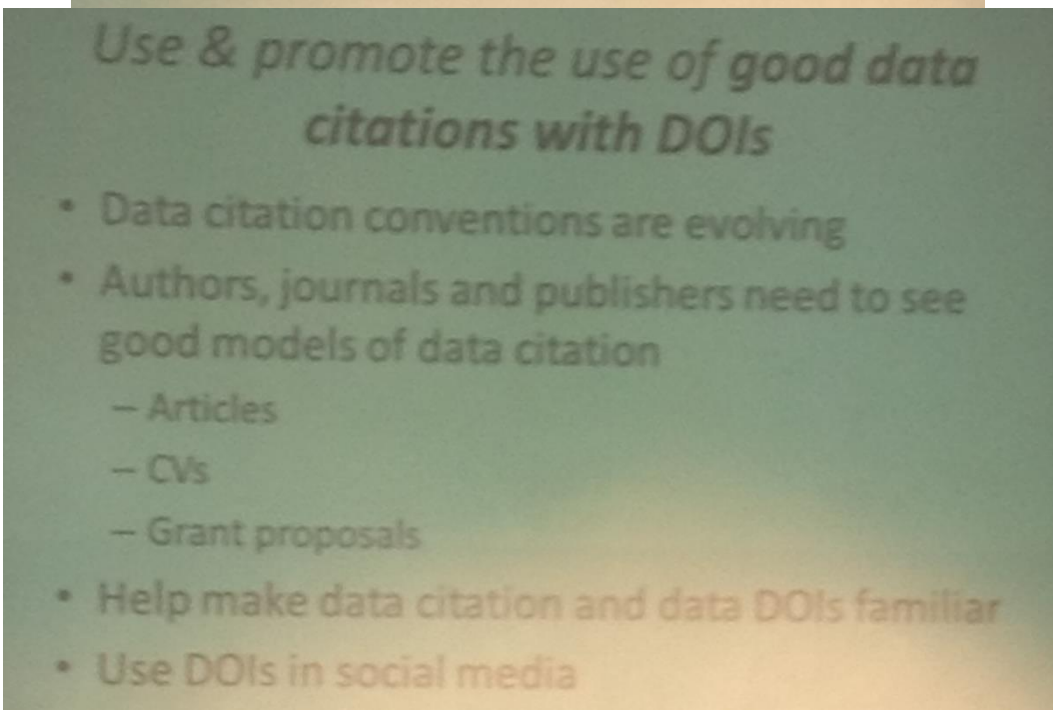
DANS

RDM Best practice for researchers



Andre slides:





Checking citations to the data in the data-sharing article

For 338 articles associated with Dryad data:

-- 253 did include a DOI for the data (75%)

-- 85 did not (25%)

- where the DOIs were located:

-- dedicated section (Data accessibility) n= 148

-- in or near article header n= 43

-- in-text (Methods, Acknowledgments): n= 71

-- in References: n= 28 (but: 17 are not actual full citations in the style of the other citations).

Our Research Data Management Roadmap (to implement the policy)