

Flaggskeppsprojektet "Certifiering av Bolincentrets Databas"

Merlijn de Smit

Stockholms universitetsbibliotek



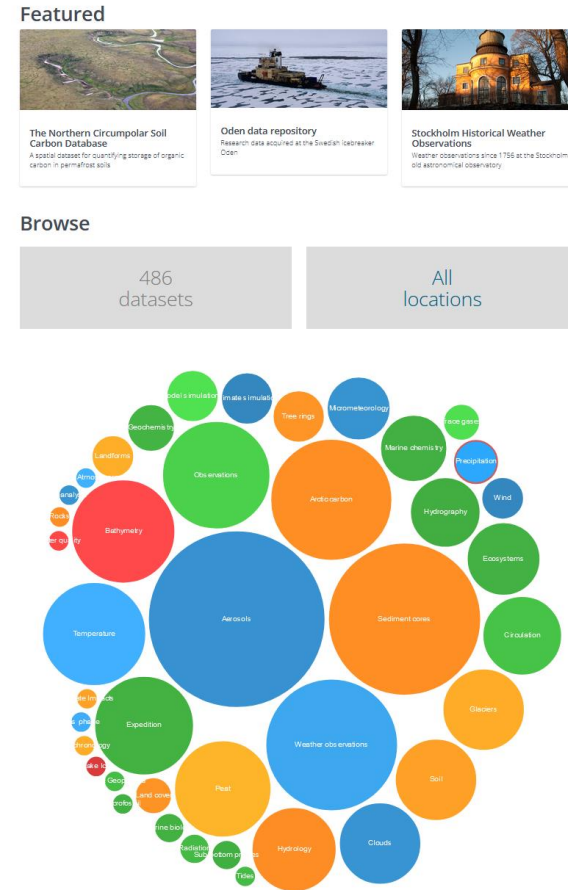
Stockholms
universitet

Bolincentrums databas är...

- **Ett stort repositorium** med nästan 500 publicerade dataset, däribland serier som Stockholm Historical Weather Observations samt forskningsdata från isbrytaren Oden.
- Som har drivits hittills med **små resurser** – två personer på deltid och en lokal server i Geovetenskapens Hus på Stockholms universitet.

Vad händer med databasen om till exempel:

- Anders går med pension eller Rezze byter jobb?
- Bolincentrets finansiella situation ändras negativt?
- Servern brinner ner?



Flaggskeppsprojektet

Övergripande mål: att förbereda och skicka in en ansökan om CoreTrustSeal-certifiering av Bolincentrets databas.

Men att uppfylla krav för CoreTrustSeal-certifiering innebär att:

- Vi kan redogöra för repositoriets långsiktiga hållbarhet, när det gäller finansiering men också personaltillgång och kompetensdelning.
- Vi kan redogöra för hållbarheten och säkerheten av databasens tekniska infrastruktur och lagring, t.ex. integritet men också redundans/katastrofsäkring, åtkomstbegränsning till fysiska servrar och digital infrastruktur, osv.
- Databasen och forskaren har samma bild om ömsesidiga rättigheter och förpliktelser, grundrad i ett skriftligt avtal/överenskommelse

Och allt detta ska vara dokumenterat!

Projektgruppen

Från Bolincentrets databas: Anders Moberg
(koordinator), Rezwan Mohammad (teknisk förvaltning).

Från Forskningsdatagruppen: Merlijn de Smit, Maria Almbro och Jorrig Vogels (huvudsakligen), andra vid behov.

Styrgruppen: Wilhelm Widmark (överbibliotekarie), Gustaf Hugelius (Bolincentret), Mia Wahlberg (sektionschef SUB), Iris Alfredsson (SND).

Regelbunden kontakt också med: representanter Sunet/Safespring (lagring), representanter SU-IT (teknisk infrastruktur), jurister SU.



Stockholms
universitet

Teknisk infrastruktur

CTS-krav:

- Dokumentera teknisk infrastruktur och att den är adekvat för repositoret och dess "designated community" (R15)
- Dokumentera säkerhet, t.ex. åtkomst till teknisk infrastruktur men också fysisk säkerhet (R16)
- Dokumentera hur data och metadata skyddas från oönskade förändringar (avsiktiga och oavsiktliga) (R14)

Vad vi gör

- Flytta infrastrukturen till en mer central tjänst (tillsvidare: virtuell maskin hos SU-IT).
- Infrastrukturen är ett nytt system som bygger på NodeJS och Kubernetes. Den kommer vara bättre säkerhetsmässigt (med särskilda inloggningsfunktioner för forskare, granskare, osv.).
- Dokumentera detta!

Kurering

CTS-krav:

- Dokumentera kriterier för mottagandet av dataset, samt versionering och spårbarhet ändringar (R07-R08).
- Dokumentera kriterier för kvalitetskontroll (R10).
- Dokumentera arbetsflöde och rutiner i kureringsprocessen.

Vad vi gör:

- Versioneringssystem av dataset innan publicering redan på plats – vi funderar på en system som tillåter snabbt publicering och sen ”efterhandskurering”.
- I nuläget sparas korrespondens med forskaren i vårt arbetssystem (Asana), detta kommer ändras.
- En intern dokument med kriterier och guidance för olika metadata (t.ex. namn, abstrakt, databeskrivning) finns och är under utveckling.

Kompetensdelning

CTS-krav:

- Garantera kontinuitet i repositoriets drift och tillgänglighet i en föränderlig värld (R03).
- Dokumentera att repositoryet har långsiktig tillgång till finansiering och personal (R05).
- Se till att repositoryets personal har relevanta kunskaper, och kontakt med vetenskapliga experter om det behövs (R06).

Vad vi gör:

- Introducera medarbetare av forskningsdatagruppen i arbetet med repositoryet, framförallt (i nuläget) kurering av nyinkomna dataset.
- Jobba på dokumentation när det gäller kurering samt teknisk infrastruktur.

Avtal och juridik

CTS-krav:

- Skapa tydlighet när det gäller rättigheter: både repositoriets rättighet att lagra, publicera, marknadsföra, flytta data och metadata och forskarens rättighet att sätta licens (R02).
- Se till att publicering av data/metadata kan ske i enlighet med lagen och forskningsetiska principer (R04).

Vad vi gör:

- Ta fram en "Terms and Conditions" som forskaren bockar fram i samband med datapublicering, med bl.a. garanti om inga personuppgifter/känsliga personuppgifter samt om samtycke av medförfattarna.
- Ta fram texter om behandling av personuppgifter, cookies, osv.

Quality Assurance (R10)

R10. The repository addresses technical quality and standards compliance, and ensures that sufficient information is available for end users to make quality-related evaluations.

The response statement and evidence should include references to the following items:

- The approach to data and metadata quality taken by the repository including variations for different curation-levels.
- The standards that data, metadata and documentation must comply with to be acceptable for preservation and access. Whether these are general external standards, internally developed standards or specific to a community of practice.
- The quality control checks in place ensure the completeness and understandability of data and metadata.
- The approach to resolving issues e.g. whether the digital objects are returned to the depositor for rectification, fixed by the repository, noted by quality flags, and/or included in the accompanying metadata.
- The approach to managing changes to expected standards (e.g. new or updated data formats of metadata schemas) in response to changes in the technical environment or to changes in the needs of the Designated Community.
- Any links provided to other digital objects' data and metadata e.g. related digital objects, publications, or the use of controlled vocabularies and ontologies.



Vad är syftet med kureringen?

- att data är lätt att leta fram (för både människor och maskiner)
- tillgänglighet, även över lång tid
- konsistens med andra dataset
- översiktlighet
- transparens
- reproducerbarhet

-> Uppfylla FAIR-principerna



(Meta)data

- Hur ser metadata (och data) ut i olika repositorer (Bolin, SND, Zenodo, figshare, Dryad, etc.)
- Vad är viktigt i metadata?
- På vilken nivå kureras metadata (CTS-nivåerna A-D)?
- Hur bedöms om databeskrivningen är tillräckligt begriplig? Finns expertkunskaper?
- Kureras datafiler också (nivå D)?

Geochemistry data from Gamlebyviken, Swedish east coast, for the last 3000 years

Elinor Andrén, Thomas Andrén, Simon Belle

This dataset consists of geochemistry data (stable carbon and nitrogen isotopes; organic matter and carbon accumulation rates), measured on a 494 cm long splice sediment record (core EL18-SH01-03) from the Gamlebyviken estuary, Swedish east coast (Baltic Sea).

update record

Praktikexempel kurering av dataset i BCD

1. Ge namn och preliminär DOI
2. Är titeln och sammanfattningen både begripliga (för en masterstudent i geovetenskap) och informativa ("vad, när, hur"?). Vilka ändringar skulle ni föreslå?
 - Är kravet "begriplig för en masterstudent" rimligt?
3. Finns det annat i första versionen (referens, nyckelord, osv.) som ska ändras?

Bolin Centre > Data

Geochemistry data from Gamlebyviken, Swedish east coast, for the last 3000 years

Elinor Andrén, Thomas Andrén, Simon Bellé

This dataset consists of geochemistry data (stable carbon and nitrogen isotopes; organic matter and carbon accumulation rates), measured on a 494 cm long splice sediment record (core EL18-SH01-03) from the Gamlebyviken estuary, Swedish east coast (Baltic Sea).

update record

Praktikexempel kurering av dataset i BCD

4. Ladda ner första versionen av datafiler och kolla på dem: finns det uppenbara saker som ni skulle föreslå ändringar på?
 - Filformat och struktur
 - Variabelnamn, förkortningar, osv.
 - Konstiga värden?
 - Tomma celler/rader
5. Med datafilen i bakhuvud, kolla på databeskrivningen. Vad ska en bra databeskrivning innehålla?
 - Information om filer och filformat.
 - Information om variabler och beskrivning – men hur mycket?



Frågor...

- Vi funderar på ett system för "efterhandskurering": dataset publiceras först, sedan kureras den. Pangaea har ett liknande system. Vilka för- och nackdelar finns det?
- Hur mycket tid är rimligt att lägga ner på kurering?
- Går det att kurera ett dataset bra utan expertkunskaper inom området? Var går gränsen mellan kurering och "data wrangling" som egentligen tillhör forskningsprocessen? Jämför kurering och peer-review av artikelpublikationer.