

Open Repositories 2021

*Panel: Dataverse Community & CoreTrustSeal:
Certifying generalist data repositories
June 10, 2021*

CoreTrustSeal certification of DataverseNO

Philipp Conzett

UiT The Arctic University of Norway

ORCID:  <https://orcid.org/0000-0002-6754-7911>

Twitter: @PhilippConzett @DataverseNO @dataverseorg #openrepos2021



Thanks for organizing this session and
for inviting me to talk here today!

Outline of presentation

- ❑ Key facts about DataverseNO
- ❑ Experiences from our certification work
- ❑ Where to get help?

Key facts about DataVerseNO

Key facts about DataverseNO

DataverseNO ...

- ❑ is a **national, generic** repository for **open** research data;
- ❑ is **curated**, aligned with the **FAIR principles** (cf. Conzett 2020), and **CoreTrustSeal**-certified;
- ❑ runs on the **Dataverse software**;
- ❑ is operated at **UiT The Arctic University of Norway**, and thereby
- ❑ the **northernmost** Dataverse repository in the world.

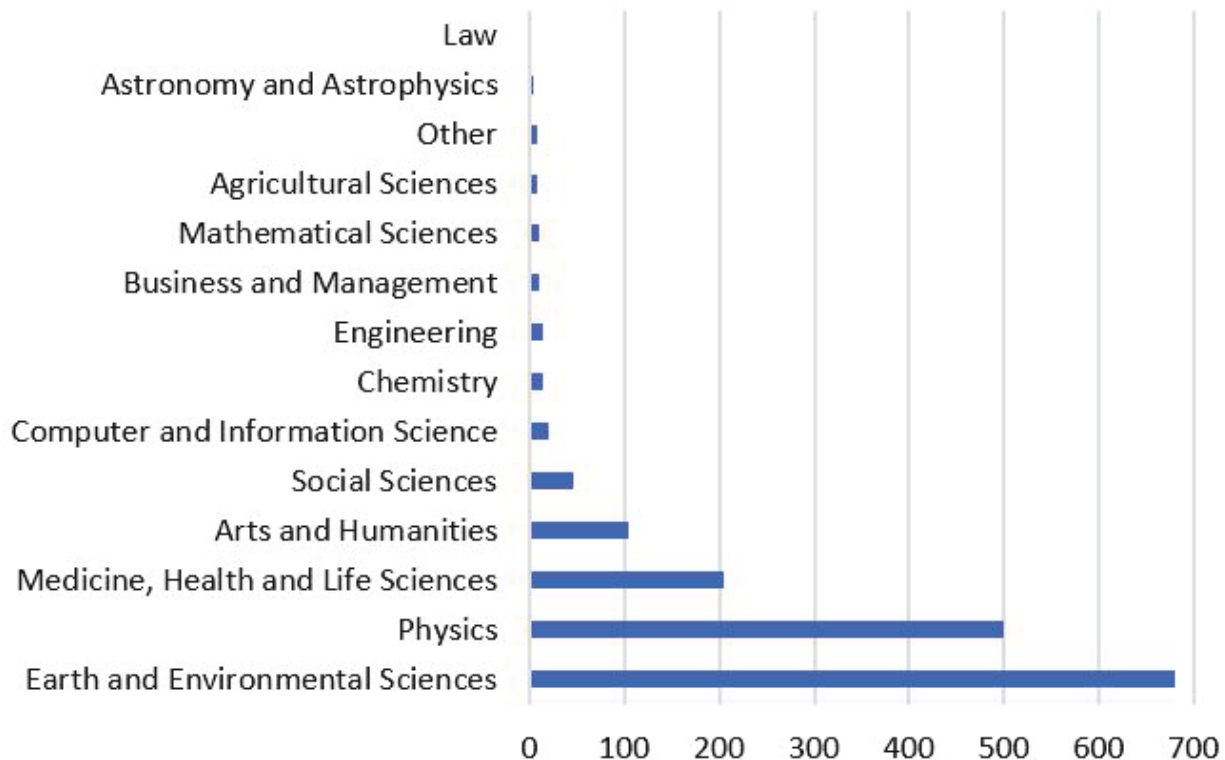
... a national repository

- ❑ **Institutional Focused** (cf. Schlatter & Ji, 2021)
- ❑ Currently **9 partner institutions** (+ a new one coming later this month...)
- ❑ Universities and university colleges
- ❑ But also open for (individual) researchers from **other Norwegian research organizations**
- ❑ Contains currently data from researchers affiliated with approx. **40 Norwegian organizations**

DataverseNO institutions



... a generic repository



- ❑ Accepting **all types** of research data from **all domains** of science
- ❑ Graph shows distribution of currently published data across domains
- ❑ High numbers within Physics and Earth Sciences are due to large **time series**.
- ❑ Apart from time series: Mostly **background data** for publications.

Numbers as of May 15, 2021

Note: Many datasets are classified as belonging to more than one domain.

Certification of DataVerseNO

CoreTrustSeal certification

To demonstrate its **commitment to FAIR data stewardship** and **trustworthy and sustainable repository management**, DataverseNO has documented its approaches and workflows to obtain **CoreTrustSeal certification**.



CoreTrustSeal requirements

The **CoreTrustSeal** evaluates the trustworthiness and sustainability of data repositories based on a **self-assessment of requirements (R)** grouped into 16 main themes:

Organisational Infrastructure:

- R01. Mission/Scope
- R02. Licenses
- R03. Continuity of Access
- R04. Confidentiality/Ethics
- R05. Organizational Infrastructure
- R06. Expert Guidance

Digital Object Management:

- R07. Data Integrity and Authenticity
- R08. Appraisal
- R09. Documented Storage Procedures
- R10. Preservation Plan
- R11. Data Quality
- R12. Workflows
- R13. Data Discovery and Identification
- R14. Data Reuse

Technology:

- R15. Technical Infrastructure
- R16. Security

Our application process

- ❑ **Started** working on the application **early in 2018**.
- ❑ **Three people** from the library (repository managers) with help from IT dpt.
- ❑ None of us had done this kind of self-assessment before.
- ❑ Divided CoreTrustSeal requirements between us, followed by common discussion.
- ❑ Submitted (**first version** of) application at the end of **June 2018**.
- ❑ Submitted **two more versions** based on valuable feedback from consultants.
- ❑ **Obtained the CoreTrustSeal** at the end of **March 2020**.



Main challenges

- ❑ We wanted to certify the **entire repository**. The **distributed organisation**, including multiple **institutional collections**, caused some challenges, mainly related to
 - ❑ data and metadata quality, and our
 - ❑ organizational infrastructure.
- ❑ To establish a fully-fledged **preservation plan** was another challenge.

Data and metadata quality (e.g. R08, R11)

Challenge: How to ensure data and metadata quality across collections?

Approach:

- ❑ Define **one set of common policies and guidelines** to be applied to all data. This includes:
 - ❑ **DataverseNO Policy Framework** (covering access and use, accession, deposit, preservation), fleshed out in the
 - ❑ **DataverseNO Guidelines** (aimed at depositors, curators, administrators)
- ❑ All datasets are **curated** by research data support staff before publication to ensure compliance with deposit guidelines.

Organizational infrastructure (R05)

Responsibility for collection management and data curation is **distributed** among partner institutions.

Challenge: How to ensure that **sufficient resources and qualified staff** are allocated for maintaining each collection?

Approach:

- ❑ DataverseNO partner agreement obliges partner institutions to manage their collections in compliance with common policies and guidelines.

But: This approach is not sufficient for compliance level 4. CoreTrustSeal consultants ask for **more specific documentation of resources and qualifications**. We'll have to revise some of our documentation, and probably point to a common **skills framework**.

Preservation Plan (R10)

Challenge: How to define a preservation plan containing **specific preservation actions**? All certified repositories have **high-level** preservation **policies**, but we could not find detailed plans for any of the certified repositories.

Approach:

- ❑ Create preservation plan based on Becker et al. (2009): Systematic planning for Digital Preservation: evaluating potential strategies and building preservation plans, and other resources
- ❑ Challenging work, because there were no good existing examples for research data **repositories**.

Where to get help?

Dataverse Software Guide for CoreTrustSeal Certification

The Dataverse Project community has written a **guide to help Dataverse repositories apply for the CoreTrustSeal certification**.

The guide describes how the **core functionality and design** principles of all 4.0+ versions of the **Dataverse software**, as well as the **Dataverse community** itself, **can help** complete most sections in the most recent version of the CoreTrustSeal application.

<https://dataverse.org/cts-guide>

But remember: Much of the CoreTrustSeal requirements is about **policies** and **good routines**.

Help from other projects and communities?

How could EOSC, CODATA, FAIRsFAIR, GO FAIR, RDA, repository communities, etc. help repositories to fulfill certification requirements?

For example by establishing **common frameworks** for

- ❑ preferred file formats for long-term preservation
- ❑ preservation plans
- ❑ ... or more generally for research data repository policies

References

About DataverseNO. <https://site.uit.no/dataverseno/about/>.

About The Dataverse Project. <https://dataverse.org/about>.

Becker, C., Kulovits, H., Gутtenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for Digital Preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, 10(4), 133–157. <https://doi.org/10.1007/s00799-009-0057-1>.

Conzett, Philipp. 2020. «DataverseNO: A National, Generic Repository and Its Contribution to the Increased FAIRness of Data from the Long Tail of Research». *Ravnetrykk*, 39, 74–113. <https://doi.org/10.7557/15.5514>.

Schlatter, Tania & Jonathan Ji. 2021. Personas for software? How and why we created archetypes for installation of an open source product. Poster presented at The information architecture conference (IAC21). Available at <https://drive.google.com/file/d/1SA2W7MKMRXTAzFrZmjVYM-E6o9tT1OQm/view?usp=sharing>.

Thank you for listening!

DataverseNO repository:



dataverse.no



[@DataverseNO](https://twitter.com/DataverseNO)



info.dataverse.no

Dataverse software:

Dataverse 

dataverse.org



[@dataverseorg](https://twitter.com/dataverseorg)