DataverseNO: Building a national research data management support service based on the Dataverse software

CSUC/CERCA webinar | May 18-19, 2021

Philipp Conzett
UiT The Arctic University of Norway
ORCID: 0000-0002-6754-7911
Twitter: @PhilippConzett

DataverseNO

Thank you for inviting me to this webinar!
Happy to share **share** our **experiences** with DataverseNO.
Welcome to the global Dataverse **community**!

Image: http://www.ministryfundraisingnetwork.org/blog/global-collaboration/

Course objective:

- ❑ Familiarize participants with the **Dataverse repository software**
- ❑ Give an introduction to how **DataverseNO** has established and provides a **national** research data management (RDM) **support service** based on the Dataverse software

Outline of the webinar:

❑ DAY 1 (May 18): INTRODUCTION AND ORGANIZATIONAL MATTERS

❑ DAY 2 (May 19): DEPOSIT, PUBLICATION, AND CURATION SUPPORT

Outline of DAY 1:

- ❑ PRESENTATION:
  Introduction to the Dataverse software and the DataverseNO repository
  - Main features and brief history of the Dataverse software
  - Main features and brief history of the DataverseNO
  - Organization of DataverseNO
  - Configuration of the Dataverse software at DataverseNO
  - Certification of DataverseNO

- ❑ SHORT BREAK (5 MIN.)

- ❑ DISCUSSION & ACTIVITIES:
  - Questions and comments
  - Strengths and weaknesses of Dataverse repositories
  - Pros and cons of different organizational models
  - Future challenges

Please write your **questions in the chat or Q&A**. We’ll address them in the **discussion session**.

Main features and brief history of the Dataverse software

The Dataverse® Project

Key facts about the Dataverse software

- Open source **web application** to share, preserve, cite, explore, and analyze **research data**, including support for
  - FAIR Data Principles
  - Persistent Identifiers
  - Versioning
  - Single Sign-on (SSO) Log-in
  - Integration with other tools
  - …
- Being developed at **Harvard**'s Institute for Quantitative Social Science (**IQSS**), along with **many collaborators and contributors worldwide**
- **69** known **installations** worldwide (as of May 15, 2021)
- An active and growing **user community** worldwide.

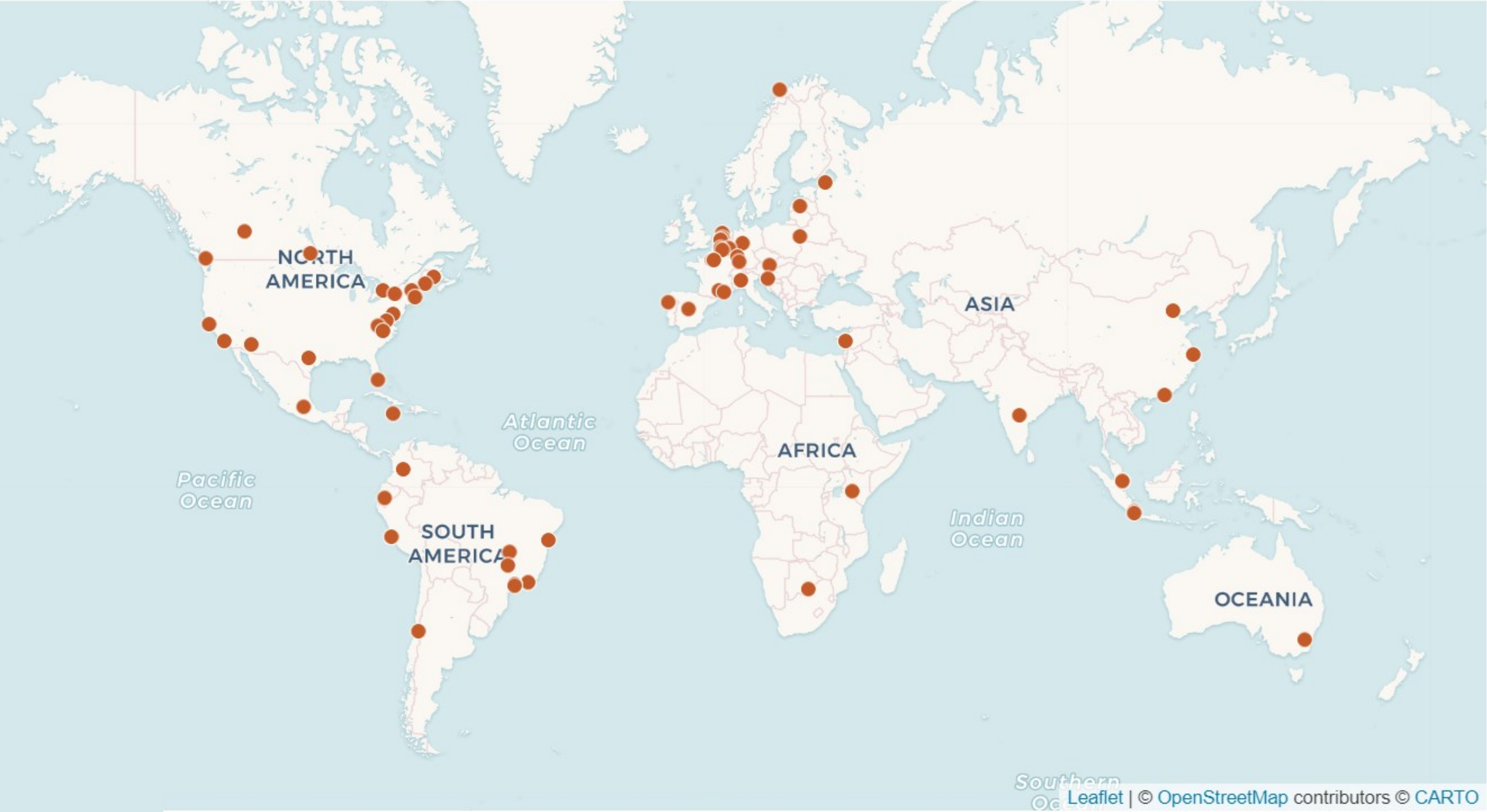
The Dataverse® Project

Brief timeline of Dataverse software

PRE-DATAVERSE		DATAVERSE		
1987	1997-2006	2006	2015	2020
Pre-web software to automatically **transfer cataloging information** by FTP to other sites **across campus**	**Virtual Data Center (VDC)**: collaboration between the Harvard-MIT Data Center (now part of IQSS) and the Harvard University Library	**Coding of the Dataverse** (previously: Dataverse Network) **software began** under the leadership of Mercè Crosas and Gary King.	Release of **version 4.0**. **Completely rewritten**. Improving usability, disciplines support, API, permissions model.	Release of **version 5.0** (= current main version)

Sources: https://dataverse.org/about and Wikipedia

69 Installations

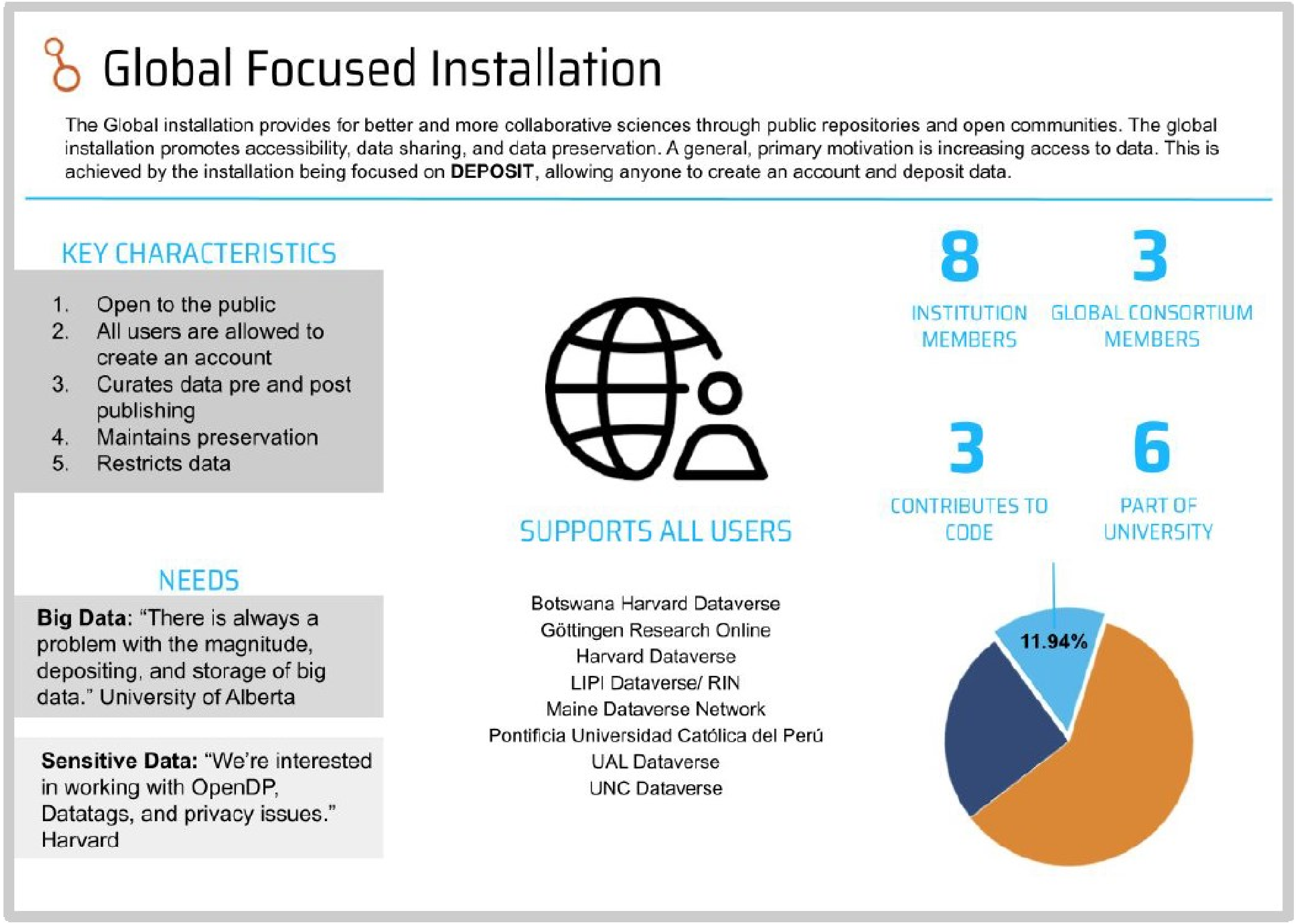


3 archetypes of Dataverse installations

Tania Schlatter & Jonathan Ji have carried out a survey among Dataverse installations to map the **typical characteristics** of the different installations. Based on the results, they distinguish between **three archetypes of Dataverse installations**:

- ❑ **Global** Focused Installation
- ❑ **Domain Specific** Installation
- ❑ **Institutional** Focused Installation

(Schlatter & Ji, 2021)



Global Focused Installation

The Global installation provides for better and more collaborative sciences through public repositories and open communities. The global installation promotes accessibility, data sharing, and data preservation. A general, primary motivation is increasing access to data. This is achieved by the installation being focused on **DEPOSIT**, allowing anyone to create an account and deposit data.

KEY CHARACTERISTICS

1. Open to the public
2. All users are allowed to create an account
3. Curates data pre and post publishing
4. Maintains preservation
5. Restricts data

NEEDS

Big Data: "There is always a problem with the magnitude, depositing, and storage of big data." University of Alberta

Sensitive Data: "We're interested in working with OpenDP, Datatags, and privacy issues." Harvard

SUPPORTS ALL USERS

Botswana Harvard Dataverse
Göttingen Research Online
Harvard Dataverse
LIPI Dataverse/ RIN
Maine Dataverse Network
Pontificia Universidad Católica del Perú
UAL Dataverse
UNC Dataverse

8 INSTITUTION MEMBERS

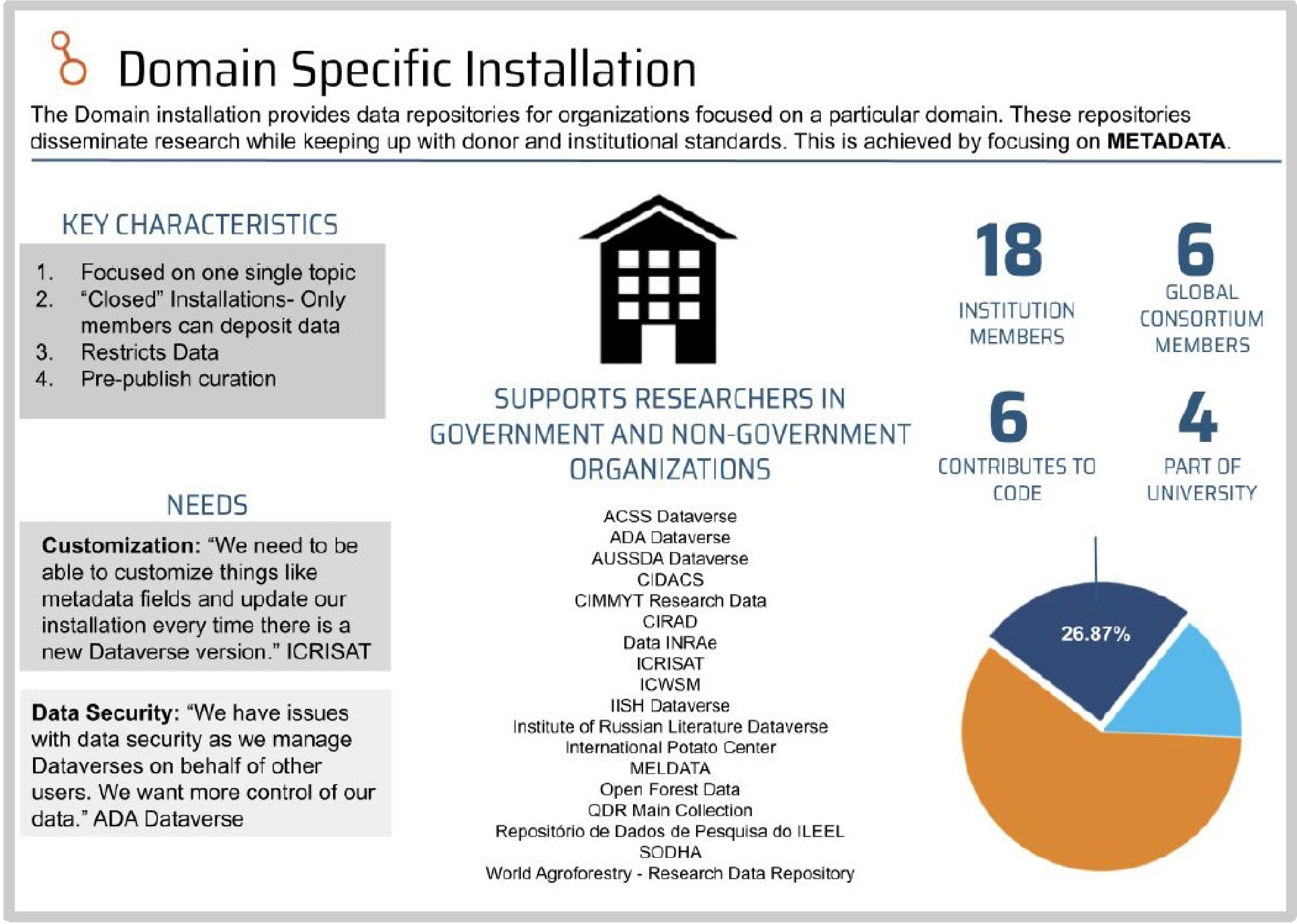
3 GLOBAL CONSORTIUM MEMBERS

3 CONTRIBUTES TO CODE

6 PART OF UNIVERSITY

From Schlatter & Ji (2021)

Typical example: Harvard Dataverse



Domain Specific Installation

The Domain installation provides data repositories for organizations focused on a particular domain. These repositories disseminate research while keeping up with donor and institutional standards. This is achieved by focusing on **METADATA**.

KEY CHARACTERISTICS

1. Focused on one single topic
2. "Closed" Installations- Only members can deposit data
3. Restricts Data
4. Pre-publish curation

NEEDS

Customization: "We need to be able to customize things like metadata fields and update our installation every time there is a new Dataverse version." ICRISAT

Data Security: "We have issues with data security as we manage Dataverses on behalf of other users. We want more control of our data." ADA Dataverse

SUPPORTS RESEARCHERS IN GOVERNMENT AND NON-GOVERNMENT ORGANIZATIONS

ACSS Dataverse
ADA Dataverse
AUSSDA Dataverse
CIDACS
CIMMYT Research Data
CIRAD
Data INRAe
ICRISAT
ICWSM
IISH Dataverse
Institute of Russian Literature Dataverse
International Potato Center
MELDATA
Open Forest Data
QDR Main Collection
Repositório de Dados de Pesquisa do ILEEL
SODHA
World Agroforestry - Research Data Repository

18 INSTITUTION MEMBERS

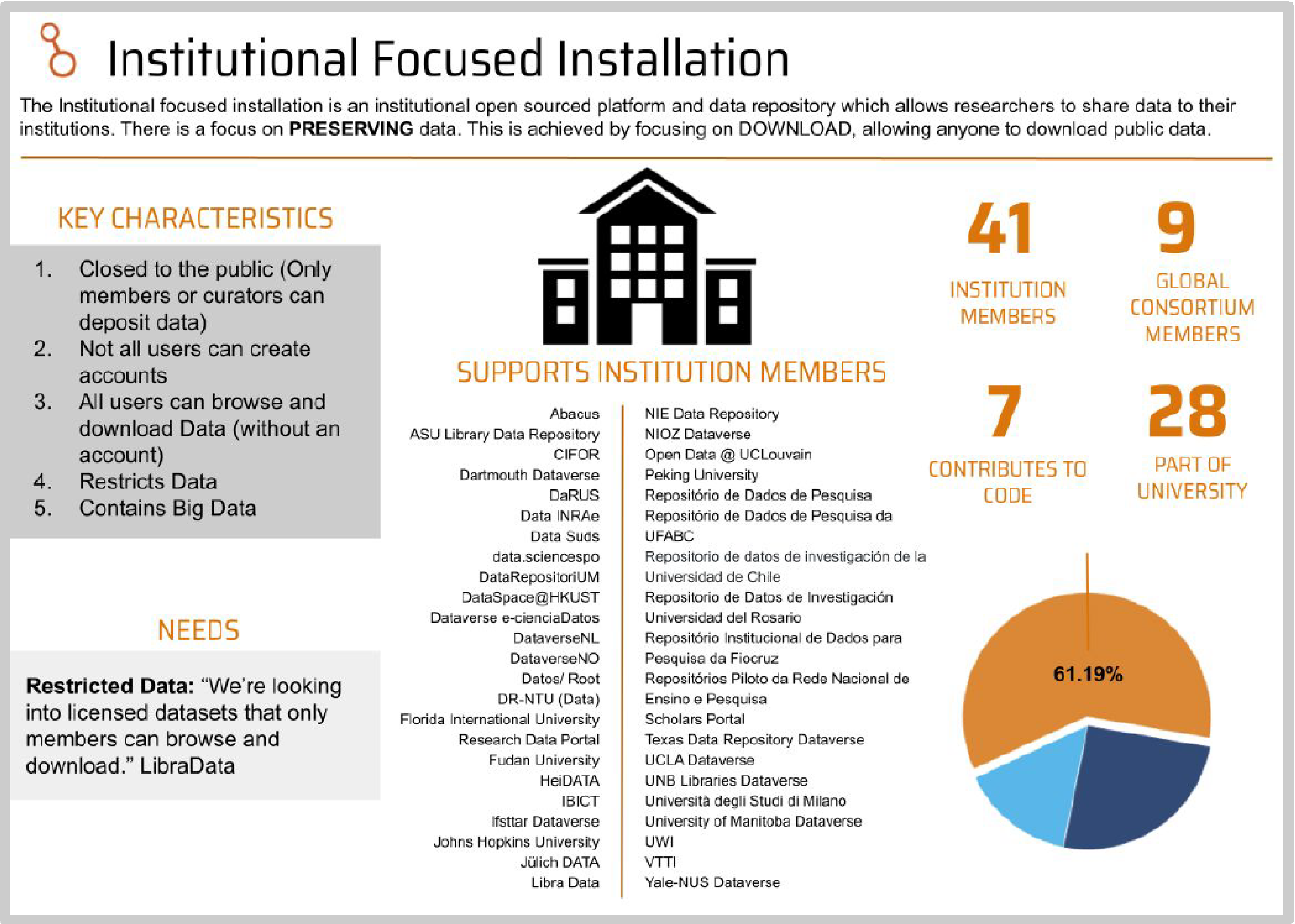
6 GLOBAL CONSORTIUM MEMBERS

6 CONTRIBUTES TO CODE

4 PART OF UNIVERSITY

From Schlatter & Ji (2021)

Typical example: AUSSDA Dataverse (The Austrian Social Science Data Archive)



Institutional Focused Installation

The Institutional focused installation is an institutional open sourced platform and data repository which allows researchers to share data to their institutions. There is a focus on **PRESERVING** data. This is achieved by focusing on DOWNLOAD, allowing anyone to download public data.

KEY CHARACTERISTICS

1. Closed to the public (Only members or curators can deposit data)
2. Not all users can create accounts
3. All users can browse and download Data (without an account)
4. Restricts Data
5. Contains Big Data

NEEDS

Restricted Data: “We're looking into licensed datasets that only members can browse and download.” LibraData

SUPPORTS INSTITUTION MEMBERS

Abacus
ASU Library Data Repository
CIFOR
Dartmouth Dataverse
DaRUS
Data INRAe
Data Suds
data.sciencespo
DataRepositoriUM
DataSpace@HKUST
Dataverse e-cienciaDatos
DataverseNL
DataverseNO
Datos/ Root
DR-NTU (Data)
Florida International University
Research Data Portal
Fudan University
HeiDATA
IBICT
Ifsttar Dataverse
Johns Hopkins University
Jülich DATA
Libra Data

NIE Data Repository
NIOZ Dataverse
Open Data @ UCLouvain
Peking University
Repositório de Dados de Pesquisa
Repositório de Dados de Pesquisa da
UFABC
Repositorio de datos de investigación de la
Universidad de Chile
Repositorio de Datos de Investigación
Universidad del Rosario
Repositório Institucional de Dados para
Pesquisa da Fiocruz
Repositórios Piloto da Rede Nacional de
Ensino e Pesquisa
Scholars Portal
Texas Data Repository Dataverse
UCLA Dataverse
UNB Libraries Dataverse
Università degli Studi di Milano
University of Manitoba Dataverse
UWI
VTTI
Yale-NUS Dataverse

41
INSTITUTION MEMBERS

9
GLOBAL CONSORTIUM MEMBERS

7
CONTRIBUTES TO CODE

28
PART OF UNIVERSITY

From Schlatter & Ji (2021)

Typical example:
Peking University Dataverse

Also DataverseNO and ?DataverseCAT belong to this category.

Main features and brief history of DataverseNO

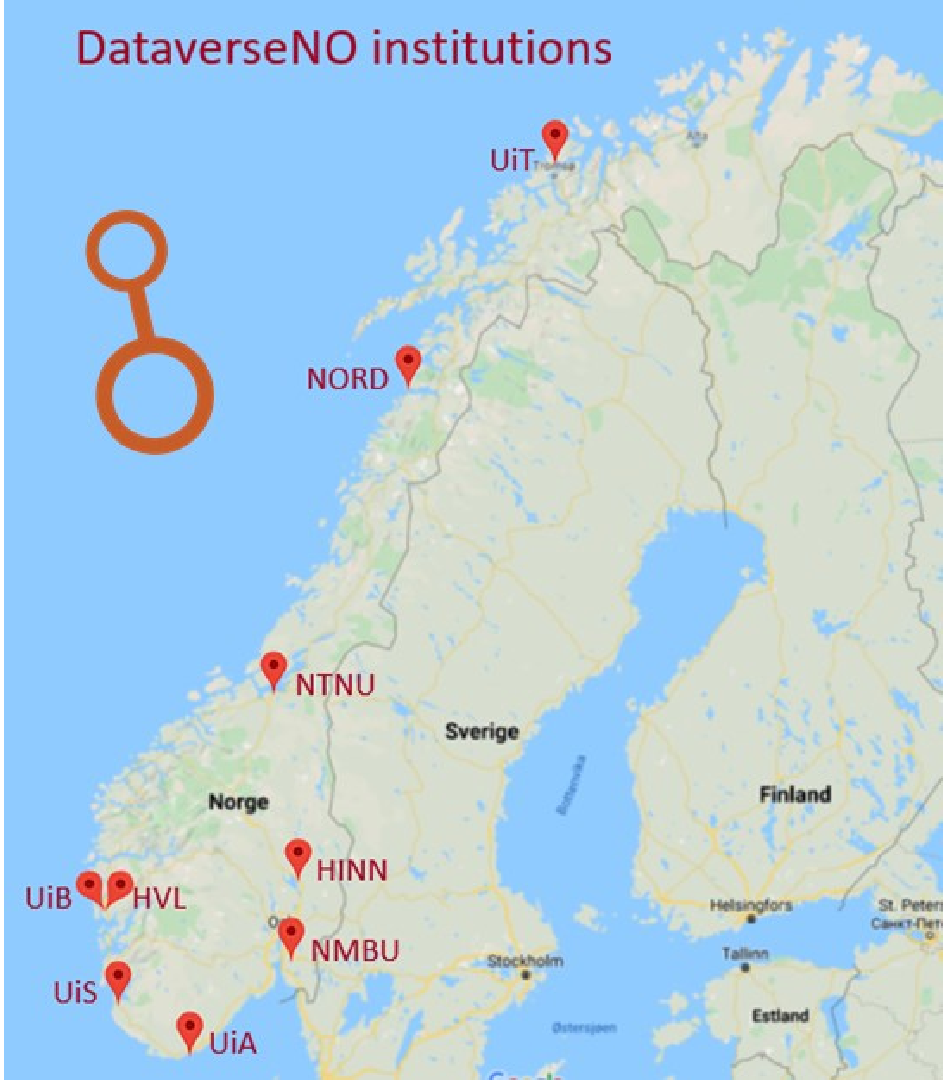
Key facts about DataverseNO

DataverseNO …

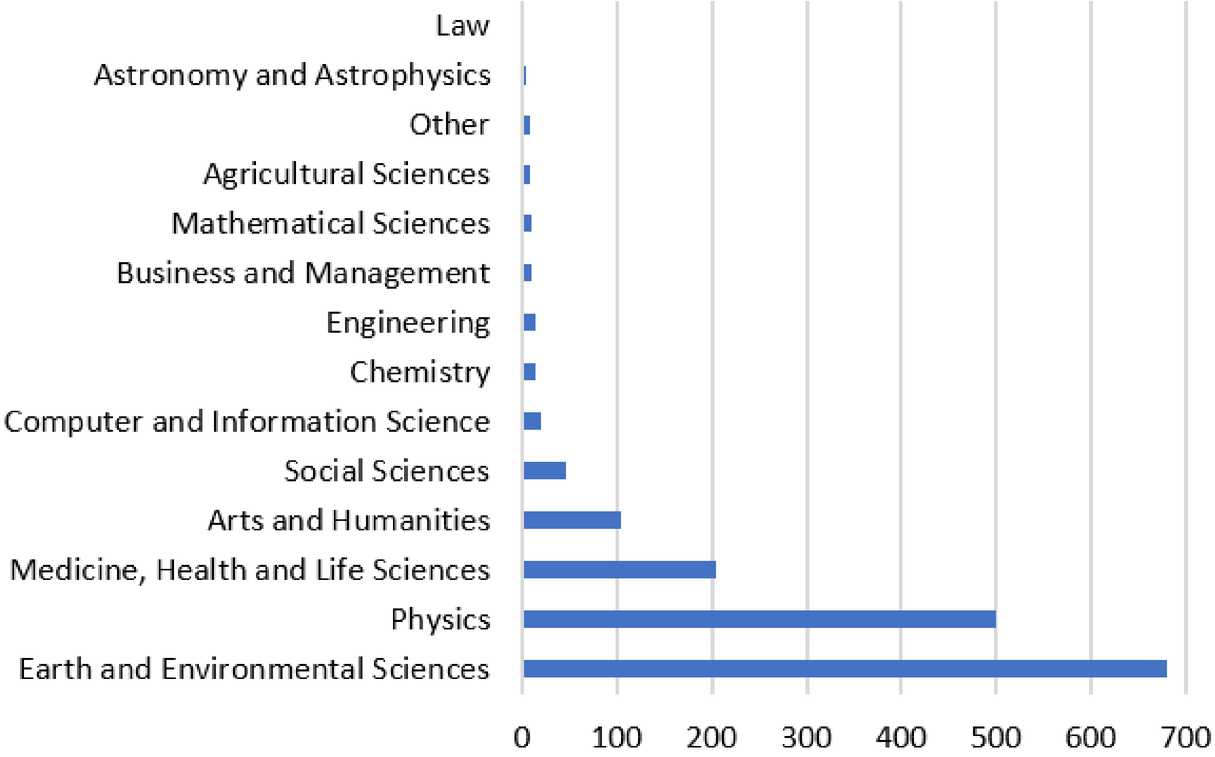
- is a **national**, **generic** repository for **open** research data;
- is **curated**, aligned with the **FAIR principles** and **CoreTrustSeal**-certified;
- runs on the **Dataverse software**;
- is operated at **UiT The Arctic University of Norway**, and thereby
- the **northernmost** Dataverse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

... a national **repository**

- **Institutional Focused** Installation (cf. Schlatter & Ji, 2021)
- Currently **9 partner institutions** (+ new one coming in June!)
- Universities and university colleges (all but 3 of the universities)
- But also open for (individual) researchers from **other Norwegian research organizations**
- Contains currently data from researchers affiliated with **37 Norwegian organizations**


DataverseNO institutions

… a generic **repository**



- ❑ Data from **all domains** of science
- ❑ Graph shows distribution across domains
- ❑ High numbers within Physics and Earth Sciences are due to large time series.

Numbers as of May 15, 2021

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

DataverseNO is one among many **repositories …**

Based on the OpenAIRE Guides for Researchers, UiT gives the following advice to its researchers on how to select a data repository:

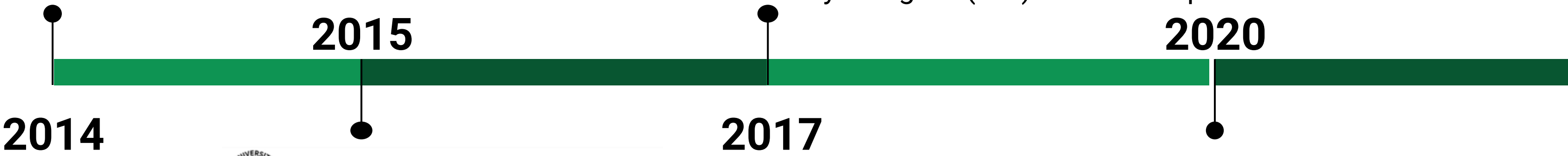
1. Funder or journal may **require** to use a **specific** repository.
2. Repository already established for your research **domain**. May use the re3data registry to find a suitable repository.
3. UiT's **institutional** collection within **DataverseNO**.
4. For data containing **person-identifying information**, we advise you to use **NSD**'s repository.

Brief history of DataverseNO --- or: Why DataverseNO?

TROLLing
The Tromsø Repository of Language and Linguistics

The Tromsø Repository of Language and Linguistics (TROLLing):

- Linguists at UiT needed a repository to share their data worldwide
- Result: **TROLLing** launched in 2014
- Based on Dataverse
- Part of CLARIN-ERIC



UiT Open Research Data

- UiT needed an institutional repository for research data
- **UiT Open Research Data** established in 2015 based on experiences and solutions used in TROLLing

DataverseNO

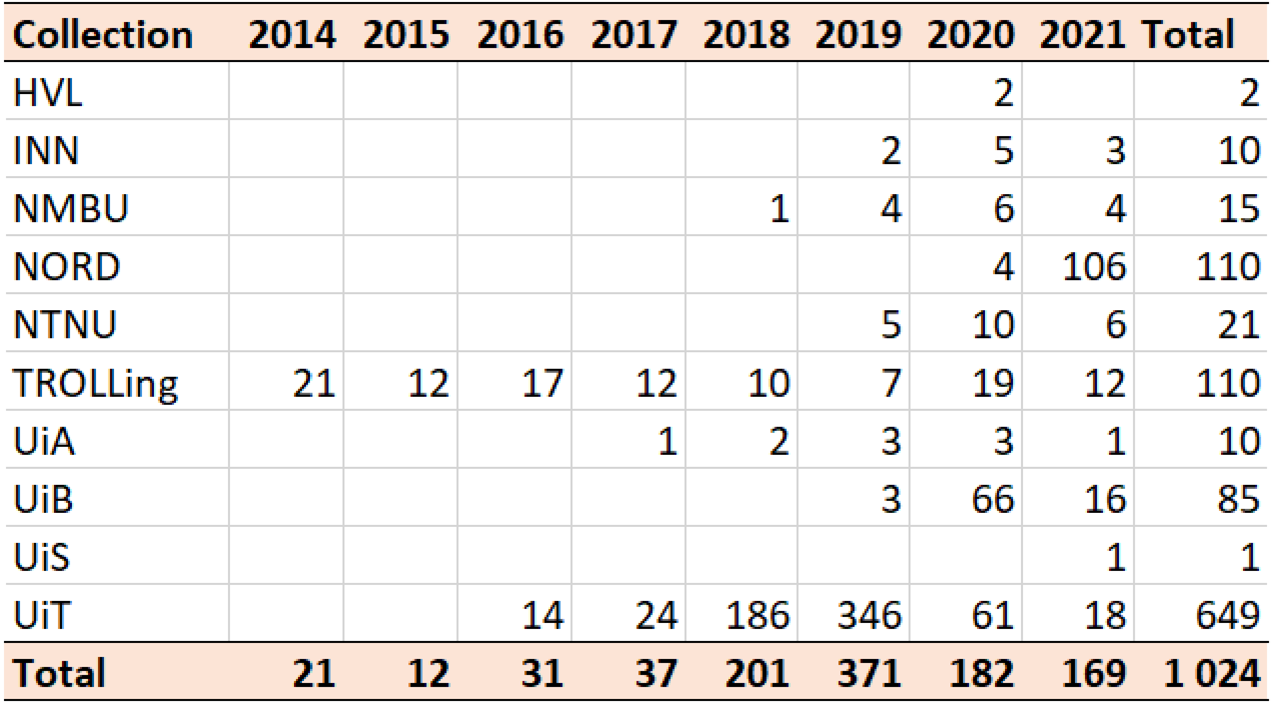
- Other universities in Norway became interested in UiT Open Research Data
- UiT decided to expand service to become a national repository
- **DataverseNO** launched in 2017 (inspired by DataverseNL) with UiT and the University of Agder (UiA) as the first partners

DataverseNO is CoreTrustSeal certified

- More details below

CORE TRUST SEAL

Number of published datasets (as of May 15, 2021)



Collection	2014	2015	2016	2017	2018	2019	2020	2021	Total
HVL							2		2
INN						2	5	3	10
NMBU					1	4	6	4	15
NORD							4	106	110
NTNU						5	10	6	21
TROLLing	21	12	17	12	10	7	19	12	110
UiA				1	2	3	3	1	10
UiB						3	66	16	85
UiS								1	1
UiT			14	24	186	346	61	18	649
Total	**21**	**12**	**31**	**37**	**201**	**371**	**182**	**169**	**1 024**

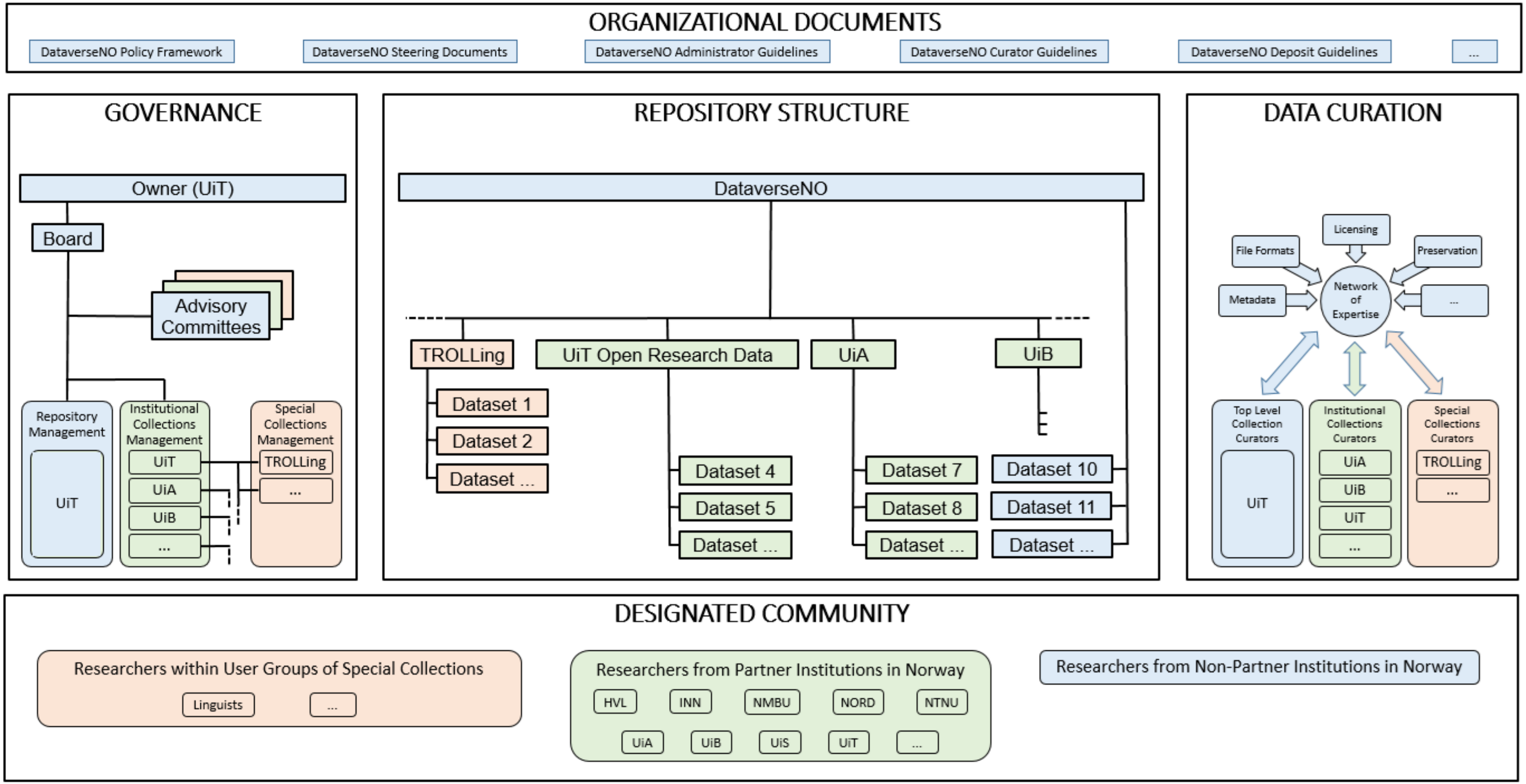
What kind of data?

- Most typically: **Background data for** article/book **publications** (other terms: supporting data, replication data, ...)
- Some larger sub-collections with **time series**:
  - UiT: Tromsø Geophysical Observatory: One dataset per month; so far: **402** datasets
  - Nord: Spawning behavior of Arctic charr: so far **127** datasets
  - UiT: NMDC Node UiT: so far **98** datasets
  - UiB: UiB Global Navigation Satellite System Data: so far **65** datasets

Organization of DataverseNO

… is outlined in the DataverseNO Organization Chart:

Organization of DataverseNO



ORGANIZATIONAL DOCUMENTS

DataverseNO Policy Framework

DataverseNO Steering Documents

DataverseNO Administrator Guidelines

DataverseNO Curator Guidelines

DataverseNO Deposit Guidelines

...

GOVERNANCE

Owner (UiT)

Board

Advisory Committees

Repository Management: UiT

Institutional Collections Management: UiT, UiA, UiB, ...

Special Collections Management: TROLLing, ...

REPOSITORY STRUCTURE

DataverseNO

TROLLing: Dataset 1, Dataset 2, Dataset ...

UiT Open Research Data: Dataset 4, Dataset 5, Dataset ...

UiA: Dataset 7, Dataset 8, Dataset ...

UiB

Dataset 10, Dataset 11, Dataset ...

DATA CURATION

File Formats

Licensing

Preservation

Metadata

...

Network of Expertise

Top Level Collection Curators: UiT

Institutional Collections Curators: UiA, UiB, UiT, ...

Special Collections Curators: TROLLing, ...

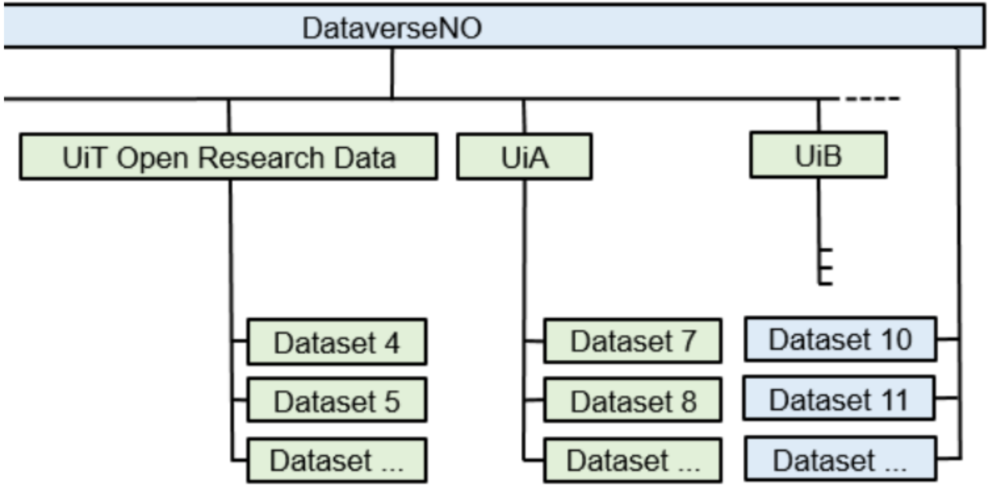
DESIGNATED COMMUNITY

Researchers within User Groups of Special Collections: Linguists, ...

Researchers from Partner Institutions in Norway: HVL, INN, NMBU, NORD, NTNU, UiA, UiB, UiS, UiT, ...

Researchers from Non-Partner Institutions in Norway

Repository structure

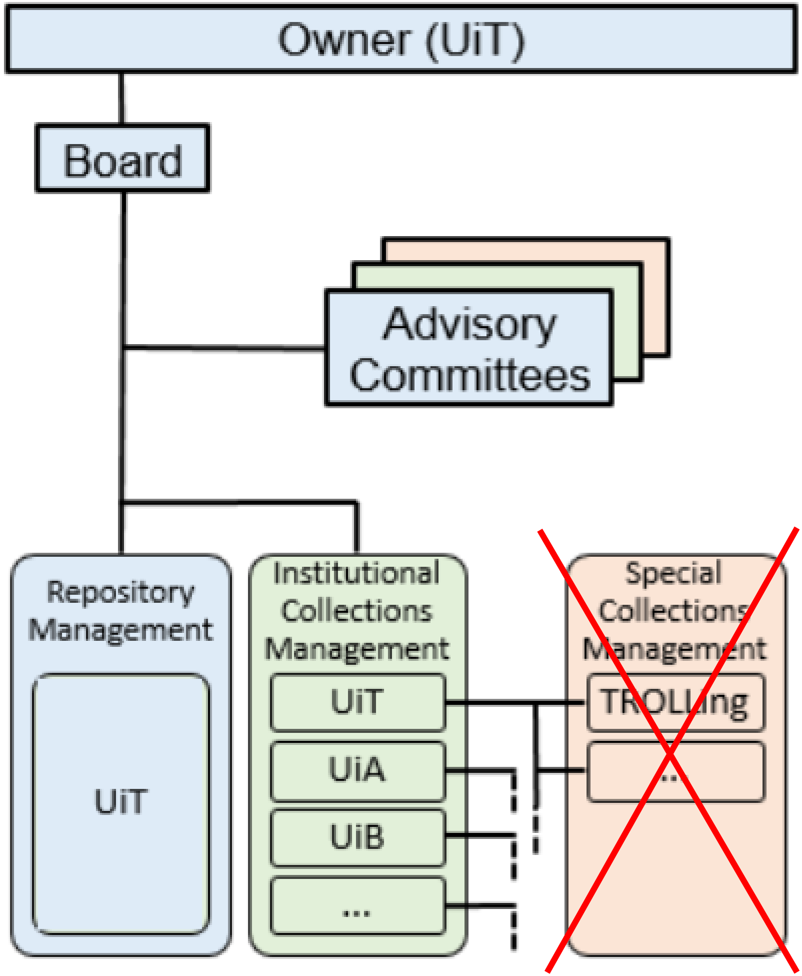


❑ **Each partner** institution has its **own collection** (sub-dataverse)

❑ Researchers from **non-partner** institutions publish their data in the **top-collection**

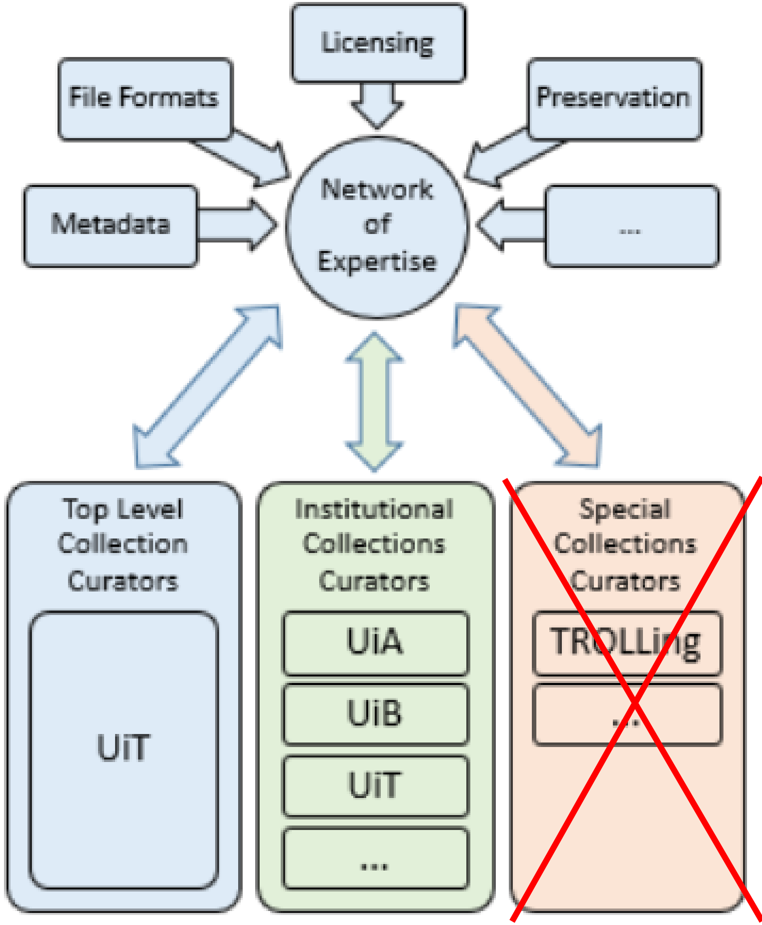
Governance

- **UiT** is responsible for management and development of **technical and functional core** of the **repository**, and training and support for collections managers.
- Each **partner** is responsible for management of **institutional collection**.
- The **Board** of DataverseNO is the highest management body of the repository.
- An **advisory committee** discusses RDM and collaboration issues and gives advice to the board.



Data curation

- ❑ In order to make published data as FAIR as possible, **each dataset** is **curated** by research **support staff** at the **partner institutions** before publication.
- ❑ Only curators can publish datasets.
- ❑ Curators share knowledge and experience in **curator network** across partner institutions.
- ❑ More details tomorrow!


Licensing
File Formats
Preservation
Metadata
Network of Expertise
...
Top Level Collection Curators
UiT
Institutional Collections Curators
UiA
UiB
UiT
...
Special Collections Curators
TROLLing
...

Organizational documents

The organization of DataverseNO is based on a set of **well-defined documents** describing the **responsibilities** of partners and the FAIR-aligned **stewardship of data** handled in the repository:

DataverseNO Policy Framework:

- ❑ Access and Use Policy
- ❑ Accession Policy
- ❑ Deposit Agreement
- ❑ Preservation Policy

DataverseNO Guidelines:

- ❑ Guidelines for Repository Managers
- ❑ Guidelines for Collection Managers
- ❑ Curatoration Guidelines
- ❑ Deposit Guidelines

DataverseNO Steering documents:

- ❑ Establishment of a Board for DataverseNO
- ❑ Mandate Board for DataverseNO
- ❑ Steering Document for DataverseNO

DataverseNO Partner Agreement (including data processor agreement):

- ❑ Partners commit to manage their collections **according to DataverseNO policies and guidelines**

DataverseNO Policy Framework

- ❏ Access and Use Policy
- ❏ Accession Policy
- ❏ Deposit Agreement
- ❏ Preservation Policy

DataverseNO Access and Use Policy

… outlines DataverseNO’s commitment to facilitating **maximum access and use** of research data.

Dissemination of content:

- Facilitating **indexing of metadata** by search engines
- Providing URLs for **harvesting** (OAI-PMH)
- Assigning Digital Object Identifiers (**DOIs**)

Access to content:

- **Discoverable and openly available**
- For **at least 10 years** after assigned DOI
- Intent is to ensure access in a **long-term perspective**

Licensing of content:

- Only accepting licenses providing **access** to deposited data **in one form or another**
- **Default** license: **CC0**

DataverseNO Accession Policy

… explains what DataverseNO can **accept for publication**.

Criteria for depositing (selection):

- ❑ Research data that are **publicly distributable**
- ❑ At least one author associated with a **Norwegian Research Institution**
- ❑ Provide **metadata** and **documentation**
- ❑ Preferred or accepted **file format**

Quality Control:

- ❑ Must **comply** with deposit **guidelines**
- ❑ DataverseNO is **not responsible** for **content** of dataset

Copyright and **IPR**:

- ❑ **If applicable**, depositor **retains copyright** to published data
- ❑ Depositor **grants to DataverseNO** the **non-exclusive right** to reproduce, translate, and distribute the dataset

DataverseNO Deposit Agreement

… defines **rights and obligations** of depositor and repository.

Depositor confirms to have read and **accept** the **terms** of the **agreement** and all related DataverseNO **policies**, including

- ❑ **transfer of custody** of datasets;
- ❑ that DataverseNO may **convert** the deposited data and/or metadata **files** to any medium or format and make multiple copies of the deposited dataset for the purposes of **security, back-up, and preservation**.

By submitting data, **depositor confirms** that

- ❑ s/he **has the right to grant the rights** contained in the Deposit Agreement;
- ❑ **nothing** in the dataset **infringes** on anyone’s copyright or other intellectual property rights;
- ❑ the dataset is in agreement with general **guidelines for research ethics**;
- ❑ … a number of other things “are OK” with the dataset …

DataverseNO Preservation Policy

… describes DataverseNO’s commitments and approaches to **responsible and sustainable stewardship** of published datasets in the **long term**.

Includes definition of:

- Preservation **Objectives**
- **Roles** and **Responsibilities**
- Preservation **Strategies**
    - Normalization
    - Format Migration
    - Bit Stream Copying
    - Fixity Checking
- **Levels** of Preservation
- Planning and **Monitoring**

Preservation Policy is fleshed out in **Preservation Plan**, including

- Asset Groups
- Preservation Action Plan

Configuration of Dataverse software in DataverseNO

Installation / deployment

- ❑ Using **main distribution** of Dataverse software
  - ❑ no forks
  - ❑ only minor adaptions (html/css and database trigger; see below)
- ❑ Runs on Linux **CentOs** distribution on **Virtual Machine** on **local server** at UiT.
- ❑ Planning to **migrate** installation **to cloud** service during 2021.

Collections

- One **institutional collection** per partner institution
- **Sub-collections** within institutional collections are created by repository manager **upon request**.

User authentication

- Researchers from **Norwegian research organizations** use **Single Sign-on** (SSO) through national authentication service (**Feide**).
- **Other users** (e.g. international collaborators) use **local authentication**:
  - Sign up in Google form for user account.
  - Repository manager creates user account.

Access rights management

Researchers from **Norwegian research organizations** are **automatically granted access** to right collection:

- ❑ Handled through trigger solution* in database based on email address:
  - ❑ Researchers from **DataverseNO partner institutions** get access to their institutional collection, e.g. ...@uit.no → UiT Open Research Data; ...@ntnu.no → NTNU Open Research Data
  - ❑ Researchers from **other Norwegian research institutions** get access to top-level collection.

Other researchers need to be granted access **manually** by repository or collection manager.

(* Thanks to DANS for help with implementing this solution.)

User groups and permissions

Three default user groups in all collections:

- Admin: admin rights
- Curator: curator rights
- Dataset Creator: **create**, but **not publish** dataset (not dataverses)

Dataset / metadata templates

- ❑ One default template per institutional collection
- ❑ Often multiple and more customized templates for sub-collections

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:

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

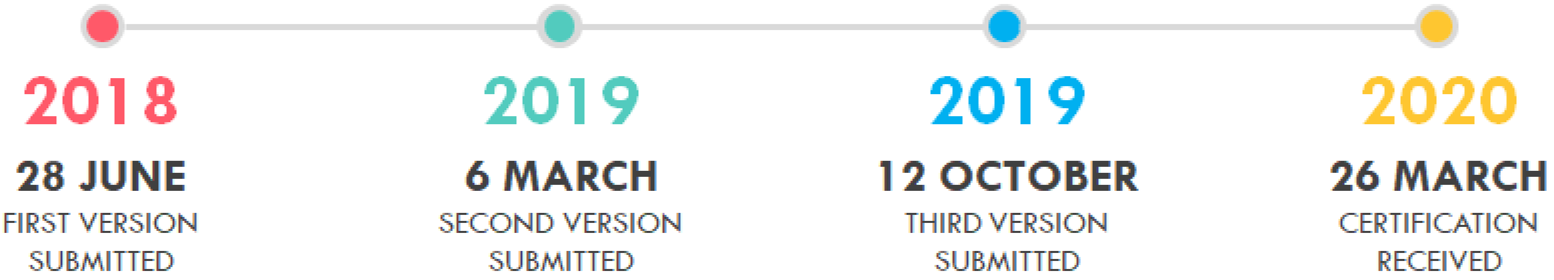
❑ None of us had done this kind of self-assessment before.

❑ 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**.



2018

28 JUNE

FIRST VERSION SUBMITTED

2019

6 MARCH

SECOND VERSION SUBMITTED

2019

12 OCTOBER

THIRD VERSION SUBMITTED

2020

26 MARCH

CERTIFICATION RECEIVED

Main challenges

- ❑ We wanted to certify the **entire repository**. The **complex organisation**, including multiple **institutional collections**, caused some challenges. (cf. different organizational models of DataverseNO and DataverseNL)
- ❑ 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 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.
- ❑ Challenging work, because there were no good existings examples for research data repositories.

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**.

Outline of DAY 1:

❑ PRESENTATION:
Introduction to the Dataverse software and the DataverseNO repository
- Main features and brief history of the Dataverse software
- Main features and brief history of the DataverseNO
- Organization of DataverseNO
- Configuration of the Dataverse software at DataverseNO
- Certification of DataverseNO

❑ **SHORT BREAK (5 MIN.)**

❑ DISCUSSION & ACTIVITIES:
- Questions and comments
- Strengths and weaknesses of Dataverse repositories
- Pros and cons of different organizational models
- Future challenges

Outline of DAY 1:

❑ PRESENTATION:
Introduction to the Dataverse software and the DataverseNO repository
- Main features and brief history of the Dataverse software
- Main features and brief history of the DataverseNO
- Organization of DataverseNO
- Configuration of the Dataverse software at DataverseNO
- Certification of DataverseNO

❑ SHORT BREAK (5 MIN.)

❑ **DISCUSSION & ACTIVITIES:**
- Questions and comments
- Strengths and weaknesses of Dataverse repositories
- Pros and cons of different organizational models
- Future challenges

Questions or comments?

Activities

Go to

https://tinyurl.com/CSUC2021

Strengths and weaknesses/challenges of DataverseNO

Strengths:

❑ Based on **approved technical solutions**

❑ Provides **strong user support** co-located researchers embedded at partner institutions

❑ Part of **strong international collaborative networks**, e.g. Dataverse community (Harvard, DANS, ...), SSHOC, FAIRsFAIR, ...

Challenges:

❑ Many subjects/domains to be covered

❑ Challenging for **small organizations** to provide extensive user curation support

But: What are the alternatives in cases where no other, more appropriate (e.g. domain-specific) repositories are available?

Thank you for your attention!

See you tomorrow!

A couple of things from yesterday...

How to find successful CoreTrustSeal certification applications?

- ❑ CoreTrustSeal homepage >> Certified Repositories
- ❑ Dataverse Software Guide for CoreTrustSeal Certification >> Introduction

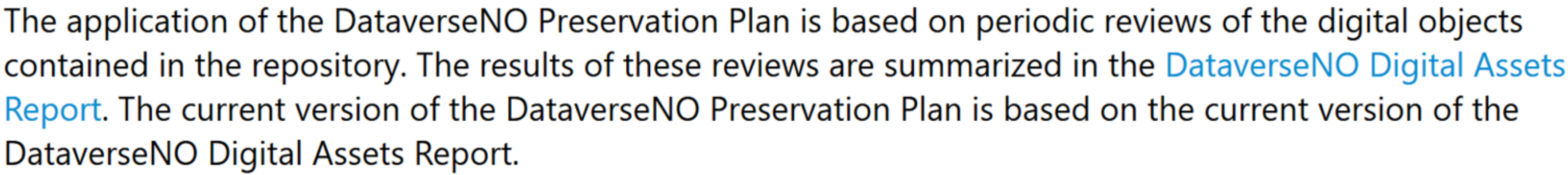
Asset groups in the Preservation Plan

- ❑ See Preservation Plan on info.dataverse.no:
  About >> Policy Framework >>

A couple of things from yesterday...

What are **asset groups** in the Preservation Plan?

- See info.dataverse.no >> About >> Policy Framework >> Preservation Plan



The application of the DataverseNO Preservation Plan is based on periodic reviews of the digital objects contained in the repository. The results of these reviews are summarized in the DataverseNO Digital Assets Report. The current version of the DataverseNO Preservation Plan is based on the current version of the DataverseNO Digital Assets Report.

Outline of the webinar:

- ❑ DAY 1 (May 18): INTRODUCTION AND ORGANIZATIONAL MATTERS
- ❑ **DAY 2 (May 19): DEPOSIT, PUBLICATION, AND CURATION SUPPORT**

Outline of DAY 2:

- ❑ PRESENTATION:
  Introduction to deposit, publication, and curation support in DataverseNO
  - Deposit and publication workflow
  - Deposit and publication support
  - Curation support

- ❑ SHORT BREAK (5 MIN.)

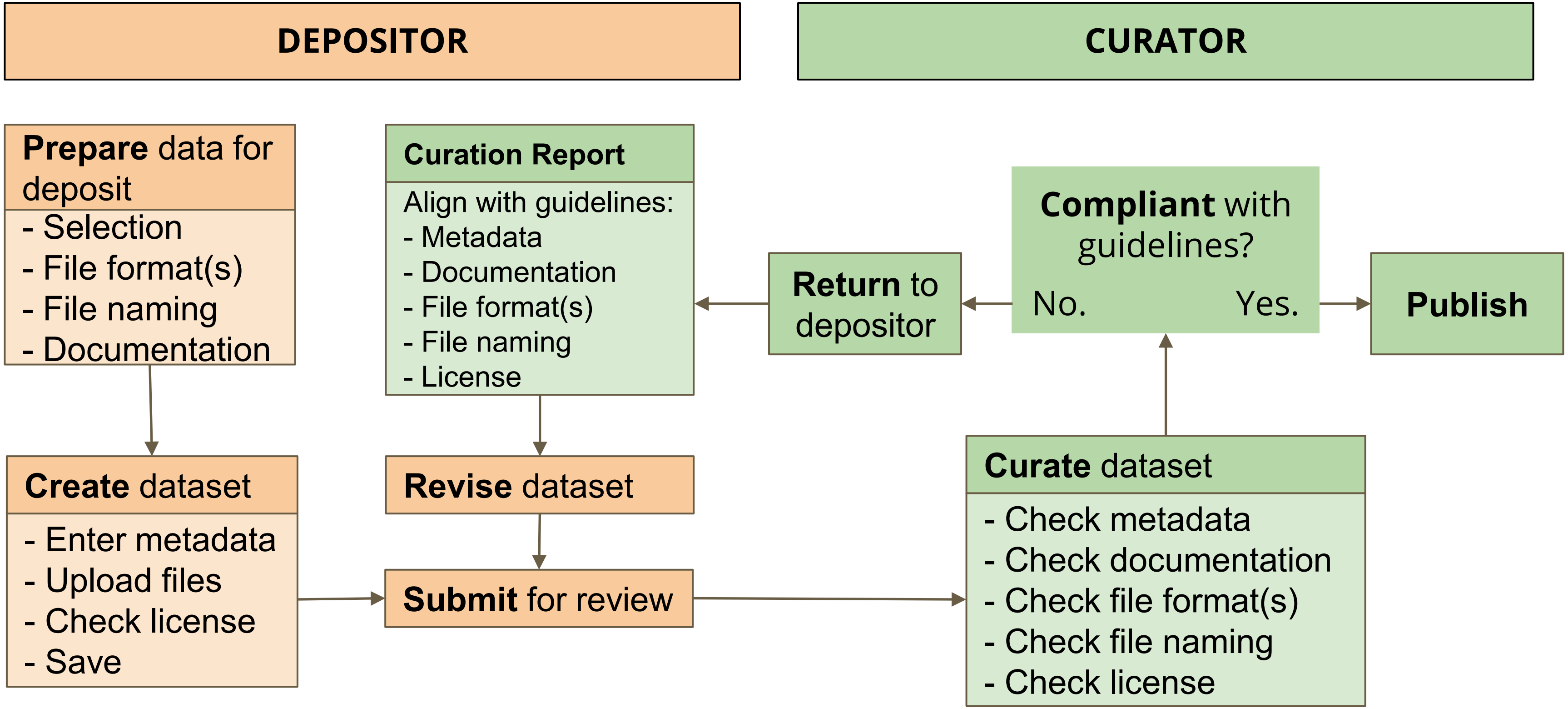
- ❑ DISCUSSION & ACTIVITIES:
  - Questions and comments
  - Challenges for repository managers and curators
  - How to provide scalable deposit support?
  - Collaboration across repositories

Please write your **questions in the chat or Q&A**.

We’ll address them in the **discussion session**.

Deposit and publication workflow in DataverseNO

Deposit, curation, and publication in DataverseNO



DEPOSITOR

CURATOR

Prepare data for deposit
- Selection
- File format(s)
- File naming
- Documentation

Create dataset
- Enter metadata
- Upload files
- Check license
- Save

Curation Report

Align with guidelines:
- Metadata
- Documentation
- File format(s)
- File naming
- License

Revise dataset

Submit for review

Curate dataset
- Check metadata
- Check documentation
- Check file format(s)
- Check file naming
- Check license

Compliant with guidelines?

No.

Yes.

Return to depositor

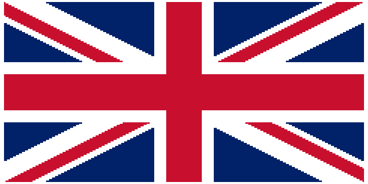
Publish

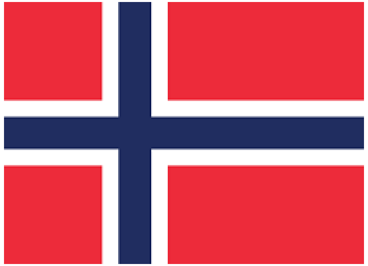
New version of dataset

The same deposit, curation, and publication procedure also applies for publication of a new version of dataset.

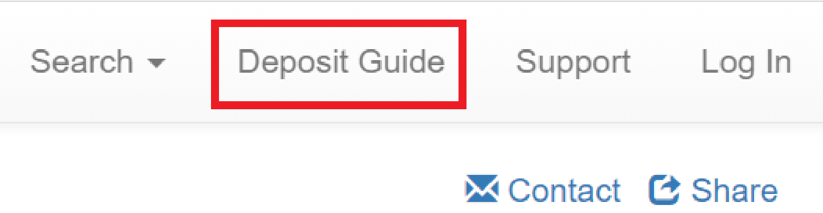
Deposit and publication support in DataverseNO

Deposit Guidelines



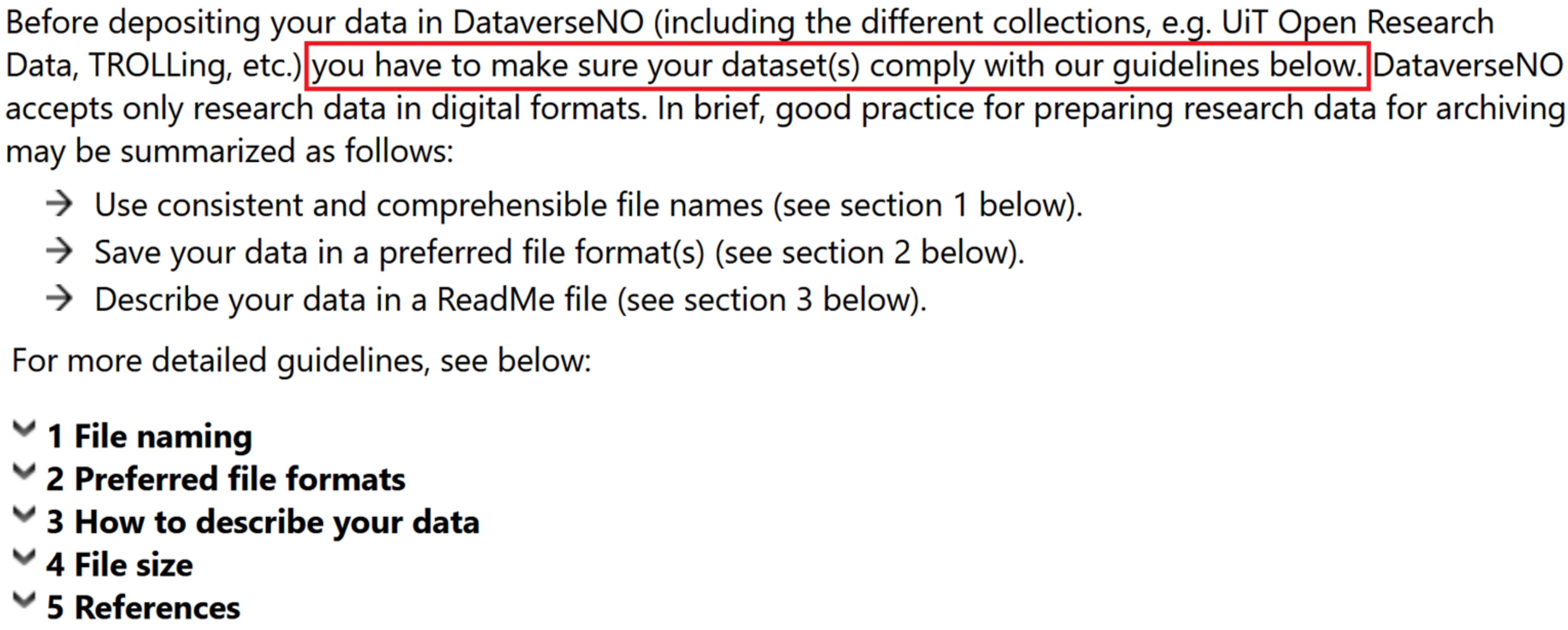


- ❑ Link to Deposit Guidelines from repository:



- ❑ Three main sections:
  - → Prepare your data for depositing
  - → Deposit your data
    - → DataverseNO Deposit Agreement
  - → Refer to your data

Prepare data for deposit



Before depositing your data in DataverseNO (including the different collections, e.g. UiT Open Research Data, TROLLing, etc.) you have to make sure your dataset(s) comply with our guidelines below. DataverseNO accepts only research data in digital formats. In brief, good practice for preparing research data for archiving may be summarized as follows:

- → Use consistent and comprehensible file names (see section 1 below).
- → Save your data in a preferred file format(s) (see section 2 below).
- → Describe your data in a ReadMe file (see section 3 below).

For more detailed guidelines, see below:

1 File naming

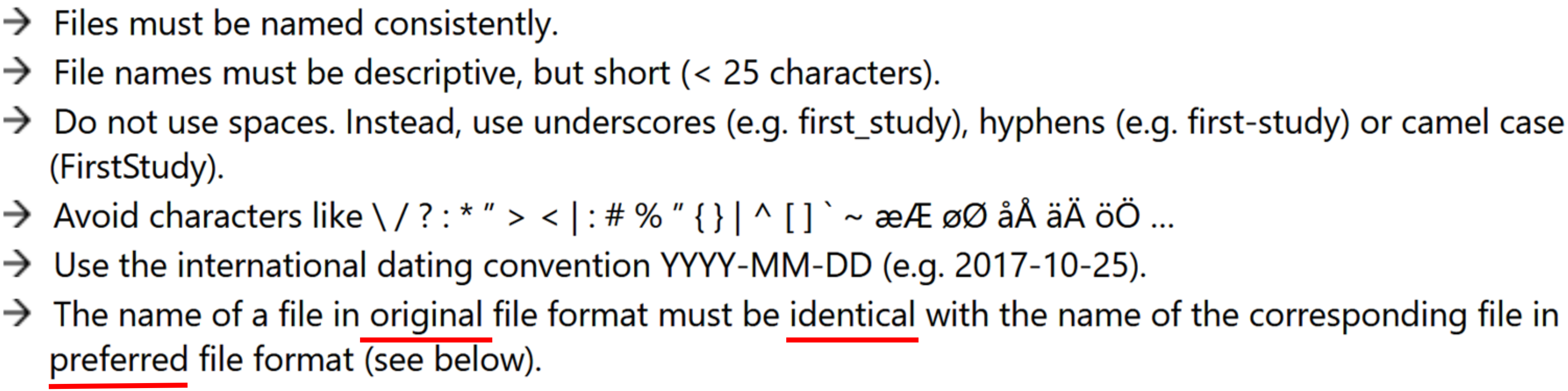
2 Preferred file formats

3 How to describe your data

4 File size

5 References

File naming recommendations



→ Files must be named consistently.

→ File names must be descriptive, but short (< 25 characters).

→ Do not use spaces. Instead, use underscores (e.g. first_study), hyphens (e.g. first-study) or camel case (FirstStudy).

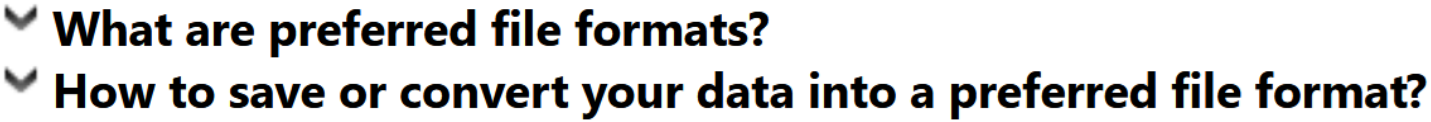
→ Avoid characters like \ / ? : * " > < | : # % " { } | ^ [ ] ` ~ æÆ øØ åÅ äÄ öÖ ...

→ Use the international dating convention YYYY-MM-DD (e.g. 2017-10-25).

→ The name of a file in original file format must be identical with the name of the corresponding file in preferred file format (see below).

❑ The last point is important to enable the repository to create assets reports for the Preservation Plan. See Asset Group 1: Items with only non-preferred file format(s)

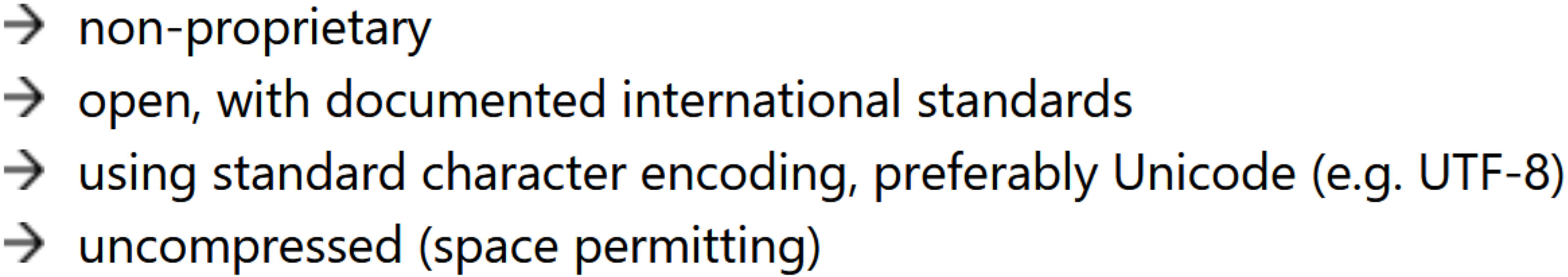
Preferred file formats



- **What are preferred file formats?**
- **How to save or convert your data into a preferred file format?**

What are preferred file formats?

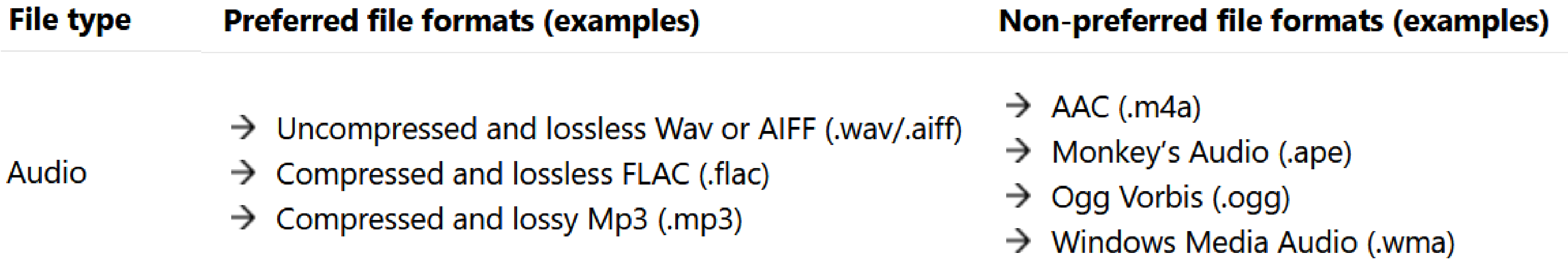
General characteristics:



→ non-proprietary
→ open, with documented international standards
→ using standard character encoding, preferably Unicode (e.g. UTF-8)
→ uncompressed (space permitting)

What are preferred file formats?

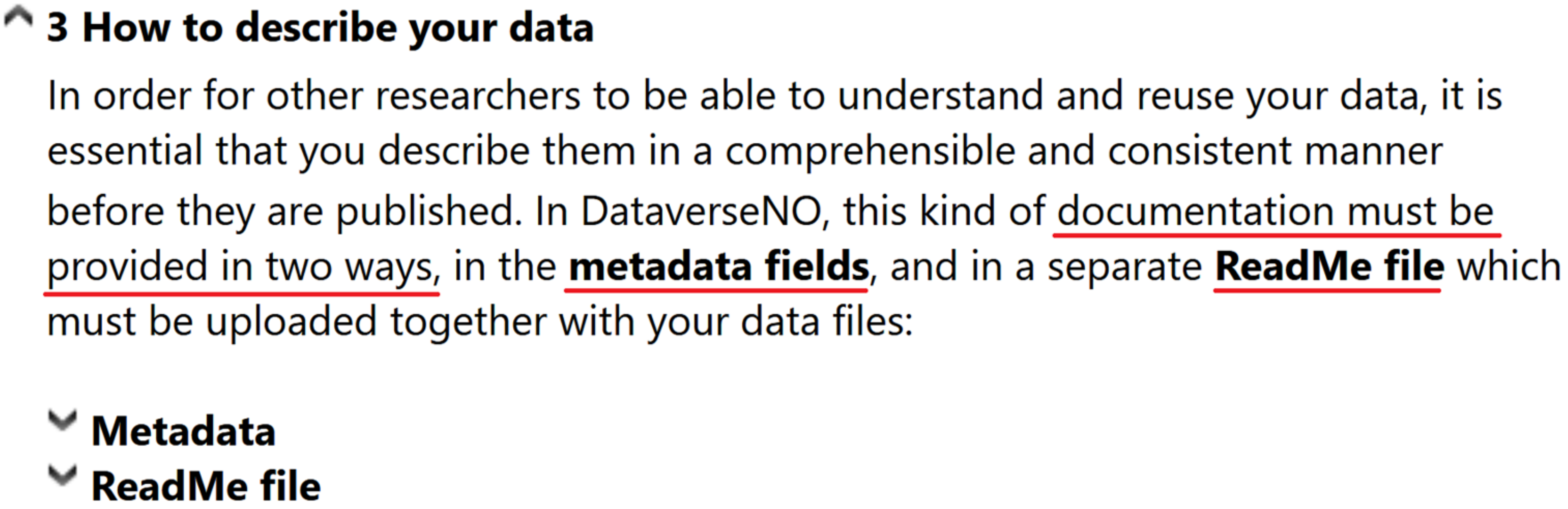
❑ List of preferred file formats:



File type	Preferred file formats (examples)	Non-preferred file formats (examples)
Audio	→ Uncompressed and lossless Wav or AIFF (.wav/.aiff) → Compressed and lossless FLAC (.flac) → Compressed and lossy Mp3 (.mp3)	→ AAC (.m4a) → Monkey's Audio (.ape) → Ogg Vorbis (.ogg) → Windows Media Audio (.wma)

❑ Collaborate on common list for Dataverse repositories?

How to describe the data?



3 How to describe your data

In order for other researchers to be able to understand and reuse your data, it is essential that you describe them in a comprehensible and consistent manner before they are published. In DataverseNO, this kind of documentation must be provided in two ways, in the **metadata fields**, and in a separate **ReadMe file** which must be uploaded together with your data files:

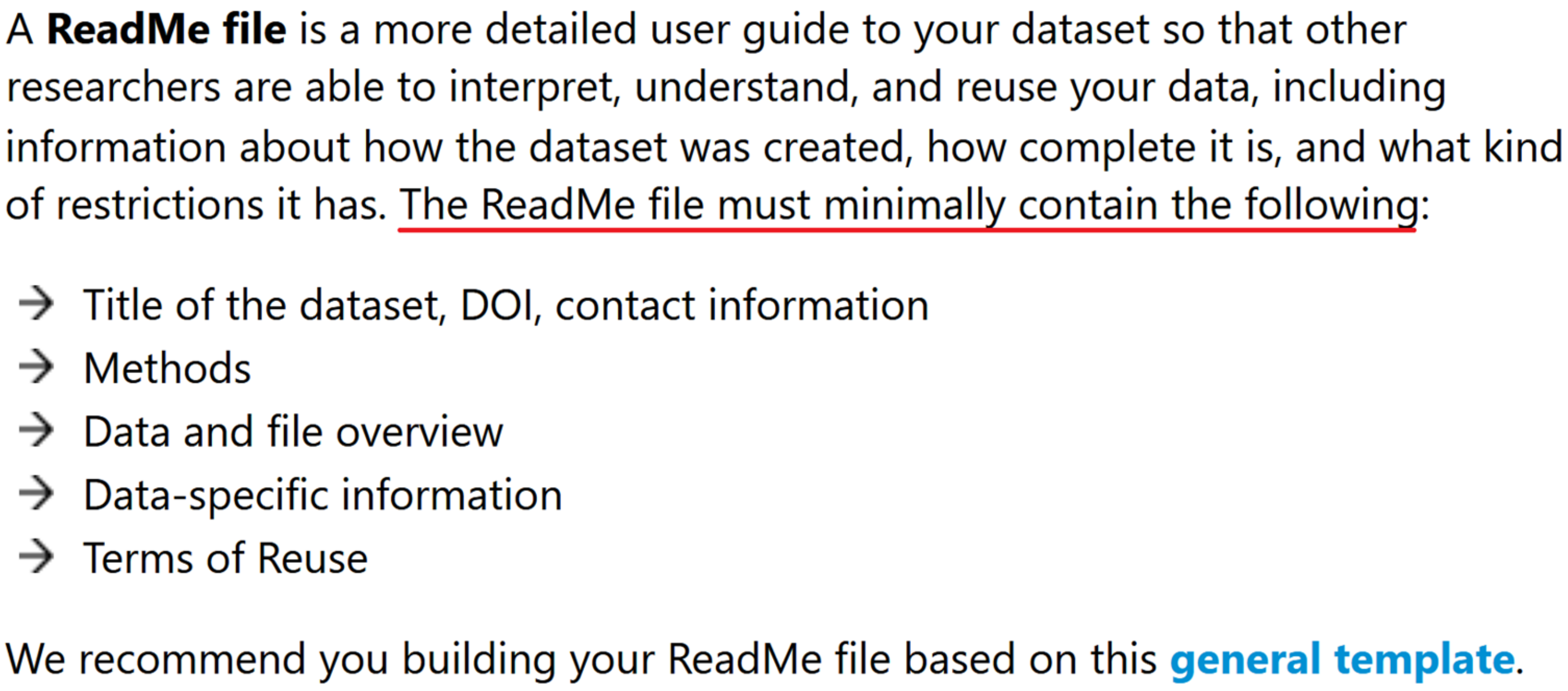
Metadata

ReadMe file

Metadata

Part of the *How to deposit data* section. See more details below.

ReadMe file



A **ReadMe file** is a more detailed user guide to your dataset so that other researchers are able to interpret, understand, and reuse your data, including information about how the dataset was created, how complete it is, and what kind of restrictions it has. The ReadMe file must minimally contain the following:

- Title of the dataset, DOI, contact information
- Methods
- Data and file overview
- Data-specific information
- Terms of Reuse

We recommend you building your ReadMe file based on this **general template**.

ReadMe file template

GENERAL INFORMATION

METHODOLOGICAL INFORMATION

> <Note! It may generally be considered appropriate to have **overlap** in the methods section of a research data README file **with** citation of the **original article**. See Committee on Publication Ethics (COPE) guidance on text recycling: https://...>

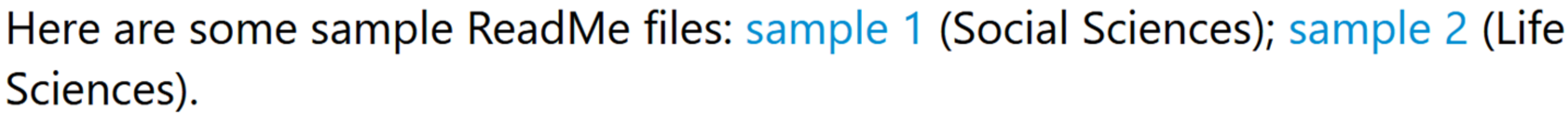
DATA & FILE OVERVIEW

DATA-SPECIFIC INFORMATION FOR: [FILENAME]

SHARING/ACCESS INFORMATION

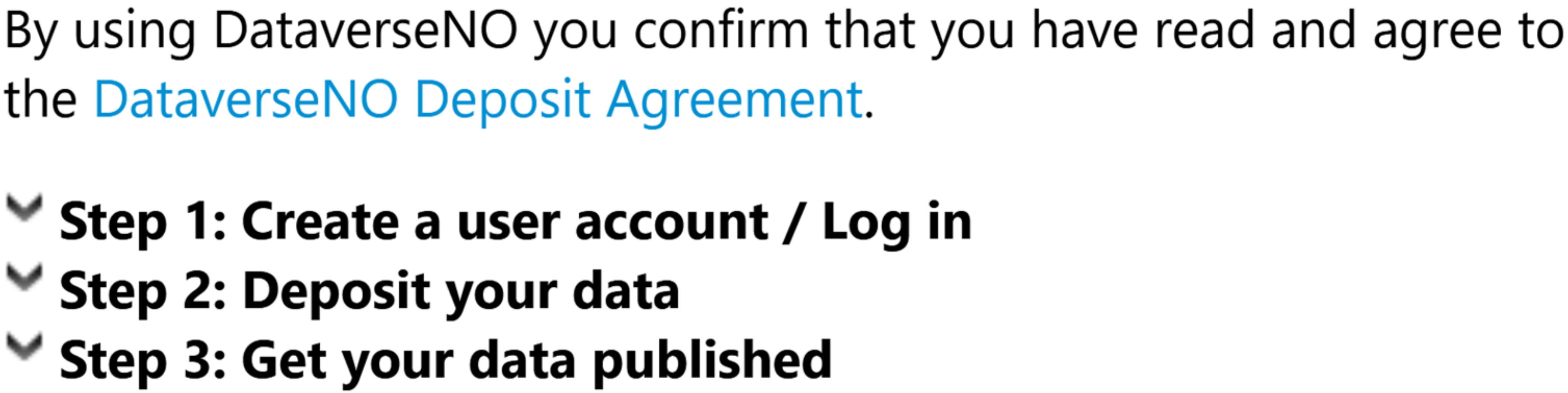
ReadMe file

Refer depositor to authentic sample ReadMe files:



Here are some sample ReadMe files: sample 1 (Social Sciences); sample 2 (Life Sciences).

Deposit data



By using DataverseNO you confirm that you have read and agree to the DataverseNO Deposit Agreement.

- **Step 1: Create a user account / Log in**
- **Step 2: Deposit your data**
- **Step 3: Get your data published**

Step 1: Create a user account / Log in

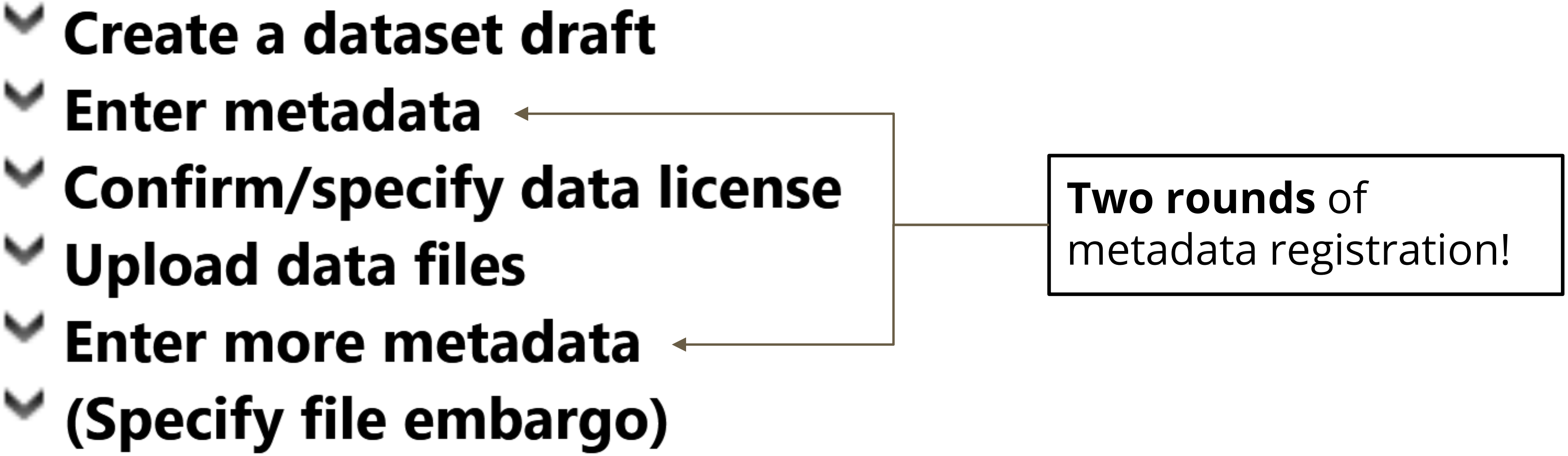
Researchers from **Norwegian research organizations**:

- ❑ Log in with your **institutional credentials** (Single sign-on, Feide)

Other researchers:

- ❑ Sign up for account using a Google form.

Step 2: Deposit your data



- Create a dataset draft
- Enter metadata
- Confirm/specify data license
- Upload data files
- Enter more metadata
- (Specify file embargo)

Two rounds of metadata registration!

Enter metadata

Deposit Guidelines contain **more information** about the following **mandatory (M)** and **recommended (R)** fields:

Round 1:

Citation Metadata:

❑ Title (M)
❑ Author (M), ORCID (R)
❑ Contact (M)
❑ Description (M)
❑ Keyword (M)
❑ Related Publication (R)

Round 2:

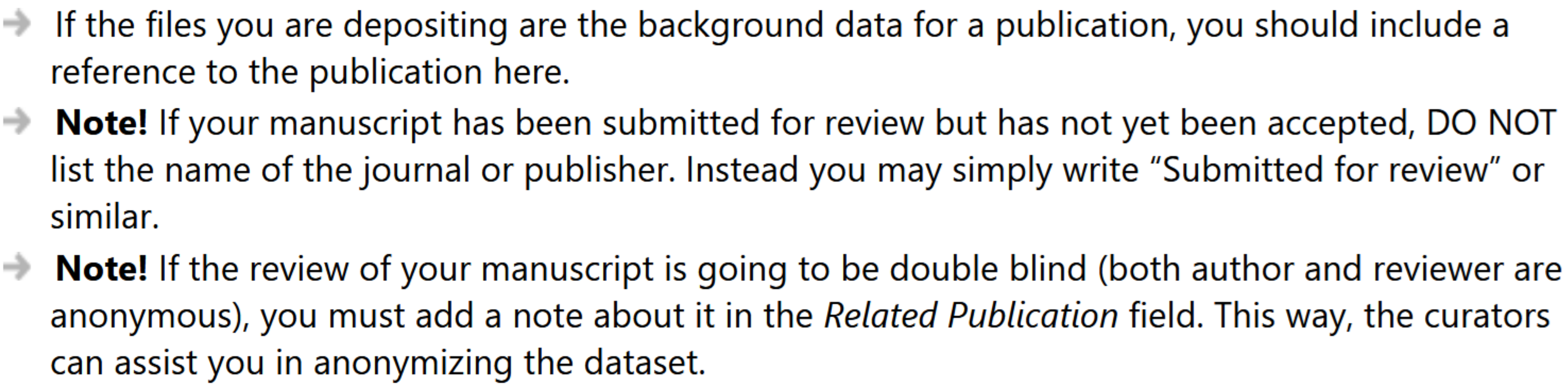
Citation Metadata:

❑ Language (R)
❑ Contributor (R)
❑ Grant Information (R)
❑ Time Period Covered (R)
❑ Date of Collection (R)
❑ Kind of Data (R)
❑ Related Material (R)
❑ Related Dataset (R)
❑ Data Sources (R)

Geospatial Metadata:

❑ Geographic Coverage (R)
❑ Geographic Bounding Box (R)

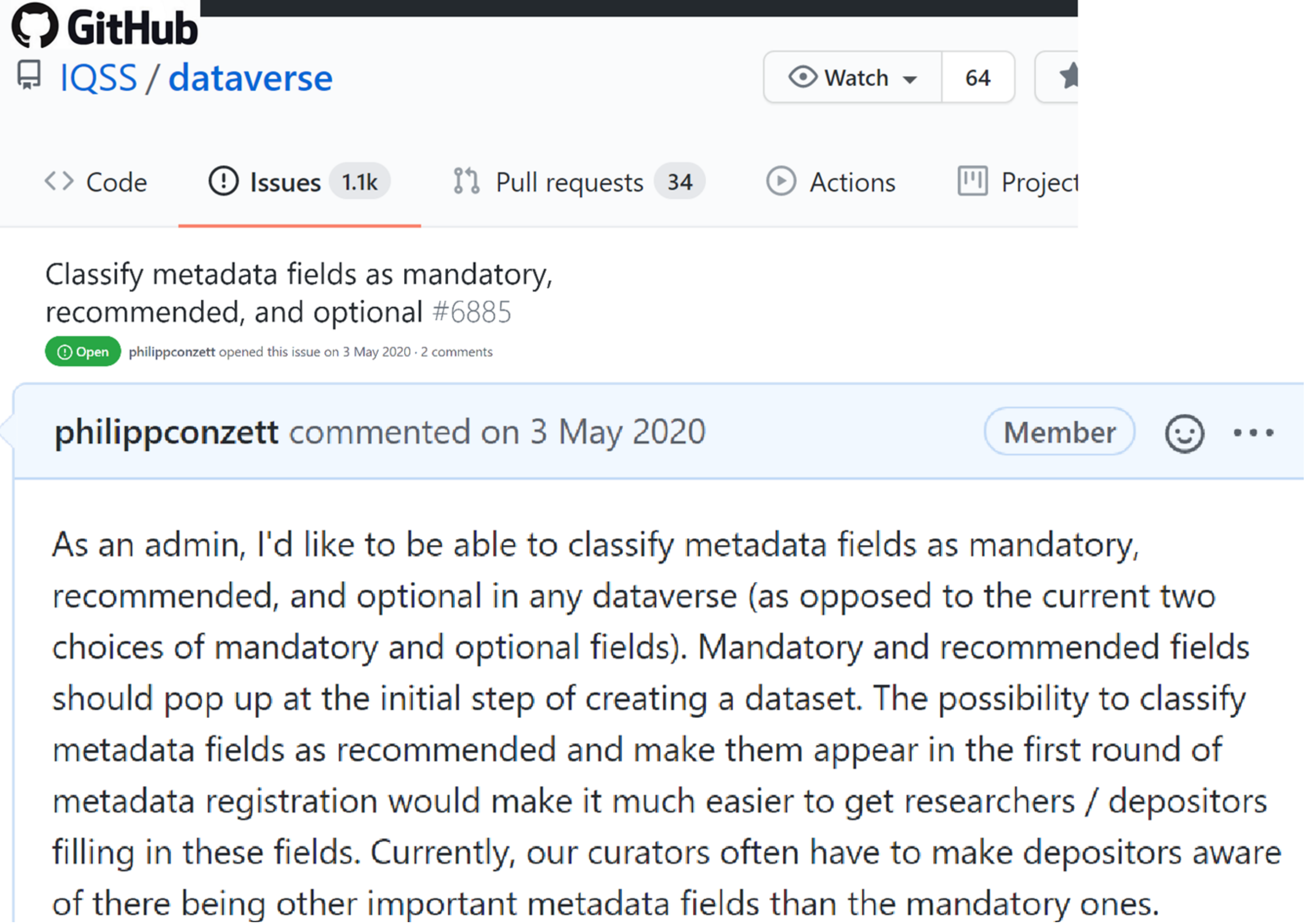
Related Publication field



- If the files you are depositing are the background data for a publication, you should include a reference to the publication here.
- **Note!** If your manuscript has been submitted for review but has not yet been accepted, DO NOT list the name of the journal or publisher. Instead you may simply write “Submitted for review” or similar.
- **Note!** If the review of your manuscript is going to be double blind (both author and reviewer are anonymous), you must add a note about it in the *Related Publication* field. This way, the curators can assist you in anonymizing the dataset.

❏ More details about dataset anonymization in section about curation support.

Proposal for improving metadata registration mode



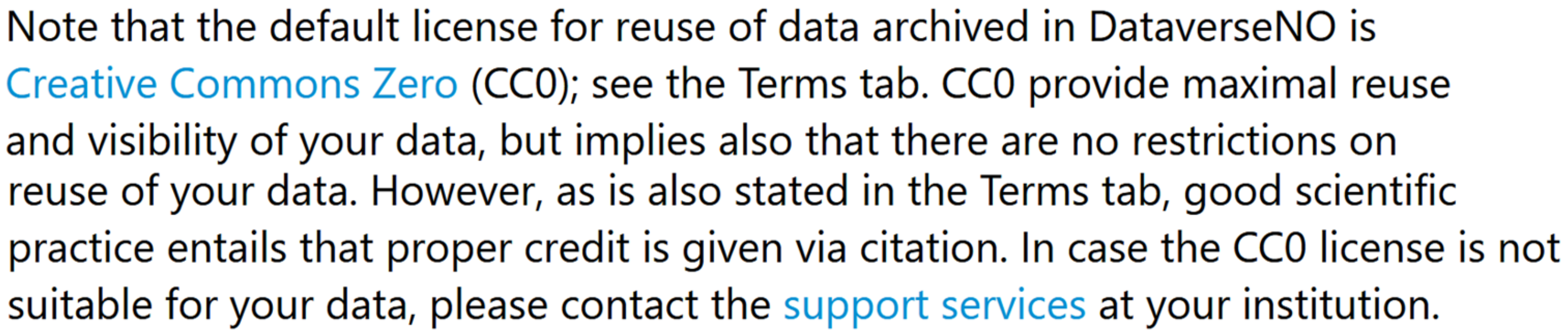
Classify metadata fields as mandatory, recommended, and optional #6885

Open philippconzett opened this issue on 3 May 2020 · 2 comments

philippconzett commented on 3 May 2020

As an admin, I'd like to be able to classify metadata fields as mandatory, recommended, and optional in any dataverse (as opposed to the current two choices of mandatory and optional fields). Mandatory and recommended fields should pop up at the initial step of creating a dataset. The possibility to classify metadata fields as recommended and make them appear in the first round of metadata registration would make it much easier to get researchers / depositors filling in these fields. Currently, our curators often have to make depositors aware of there being other important metadata fields than the mandatory ones.

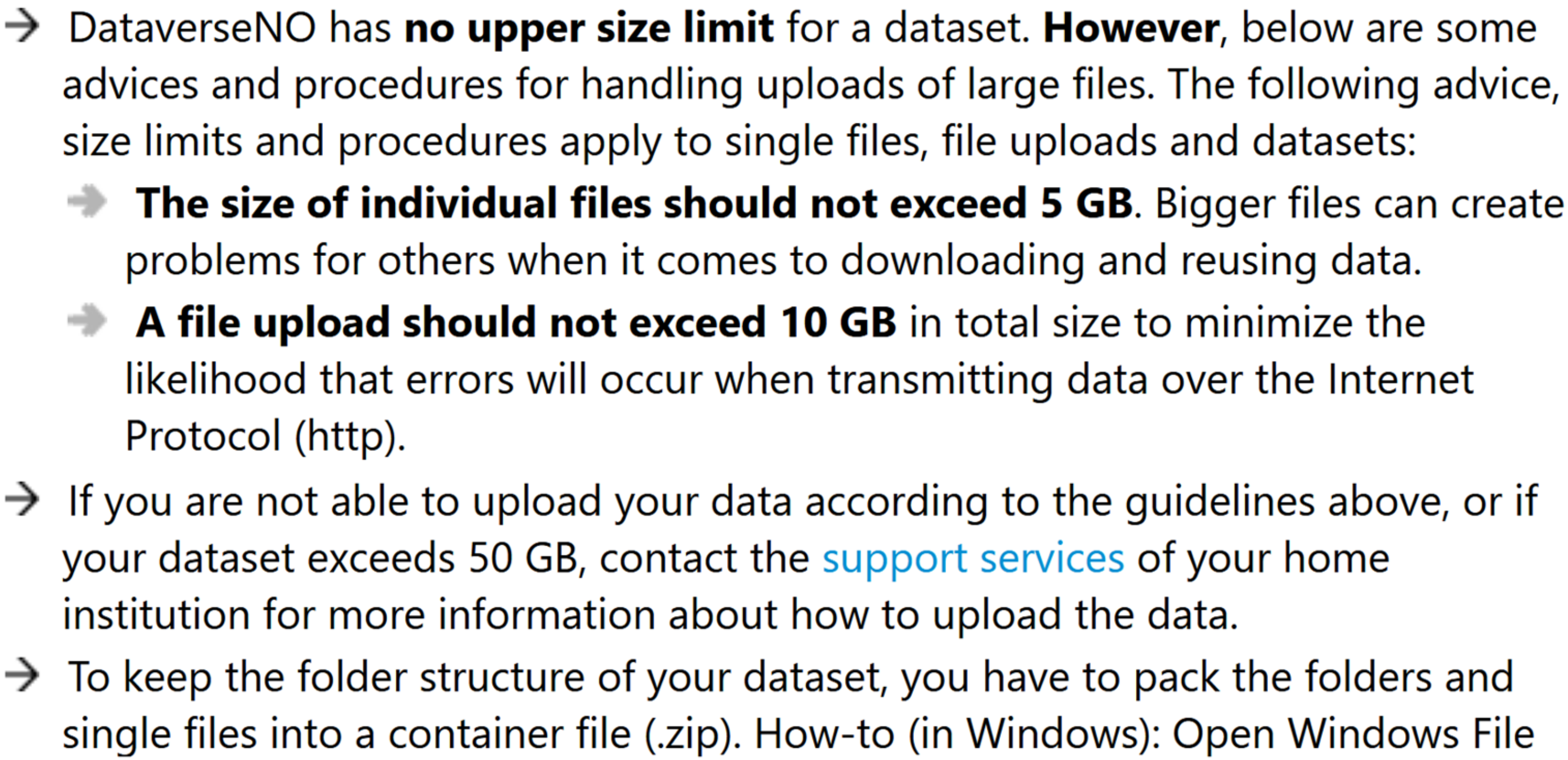
Confirm/specify dataset license and attribution



Note that the default license for reuse of data archived in DataverseNO is Creative Commons Zero (CC0); see the Terms tab. CC0 provide maximal reuse and visibility of your data, but implies also that there are no restrictions on reuse of your data. However, as is also stated in the Terms tab, good scientific practice entails that proper credit is given via citation. In case the CC0 license is not suitable for your data, please contact the support services at your institution.

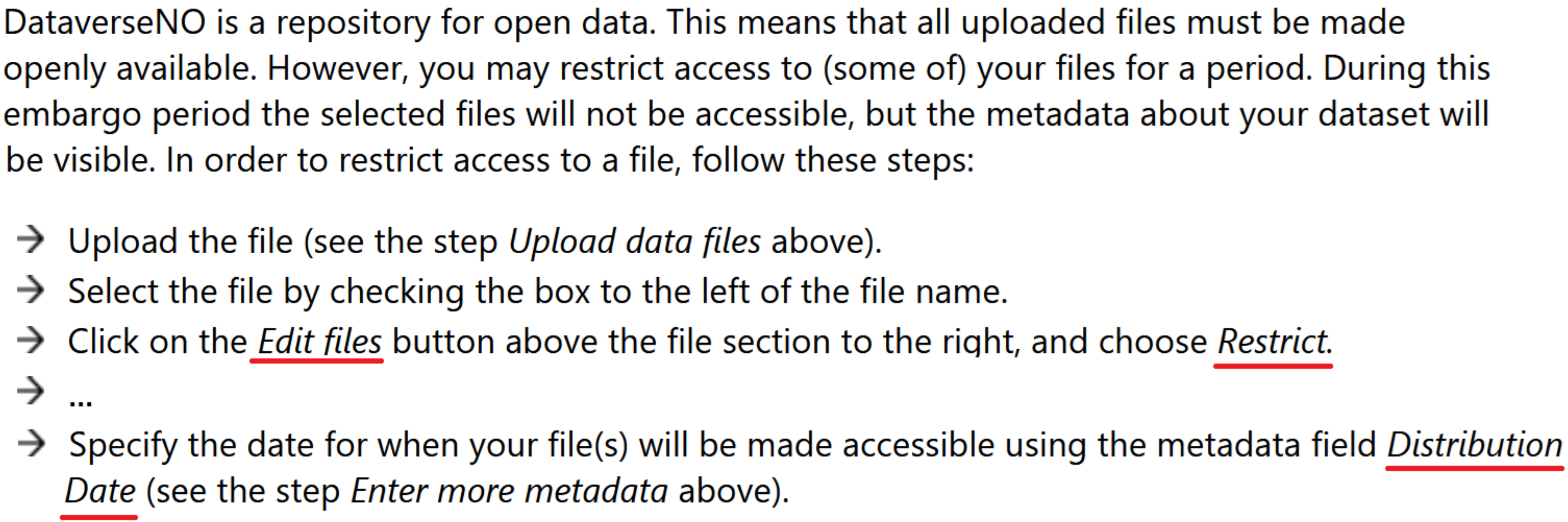
- ❑ The Dataverse community is working on implementation of more standard licenses to be supported by Dataverse software.

Upload files



- DataverseNO has **no upper size limit** for a dataset. **However**, below are some advices and procedures for handling uploads of large files. The following advice, size limits and procedures apply to single files, file uploads and datasets:
  - **The size of individual files should not exceed 5 GB**. Bigger files can create problems for others when it comes to downloading and reusing data.
  - **A file upload should not exceed 10 GB** in total size to minimize the likelihood that errors will occur when transmitting data over the Internet Protocol (http).
- If you are not able to upload your data according to the guidelines above, or if your dataset exceeds 50 GB, contact the support services of your home institution for more information about how to upload the data.
- To keep the folder structure of your dataset, you have to pack the folders and single files into a container file (.zip). How-to (in Windows): Open Windows File

(Specify file embargo)



DataverseNO is a repository for open data. This means that all uploaded files must be made openly available. However, you may restrict access to (some of) your files for a period. During this embargo period the selected files will not be accessible, but the metadata about your dataset will be visible. In order to restrict access to a file, follow these steps:

→ Upload the file (see the step *Upload data files* above).

→ Select the file by checking the box to the left of the file name.

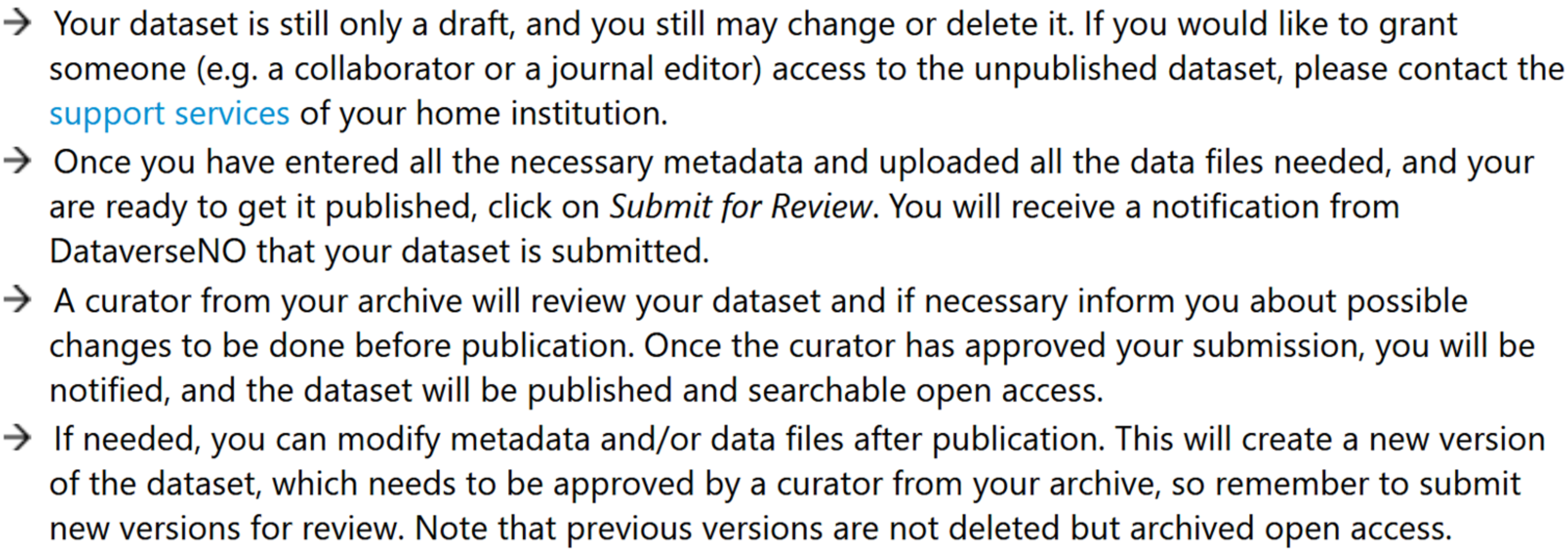
→ Click on the *Edit files* button above the file section to the right, and choose *Restrict.*

→ ...

→ Specify the date for when your file(s) will be made accessible using the metadata field *Distribution Date* (see the step *Enter more metadata* above).

❑ The Dataverse community is working on an embargo functionality.

Get dataset published



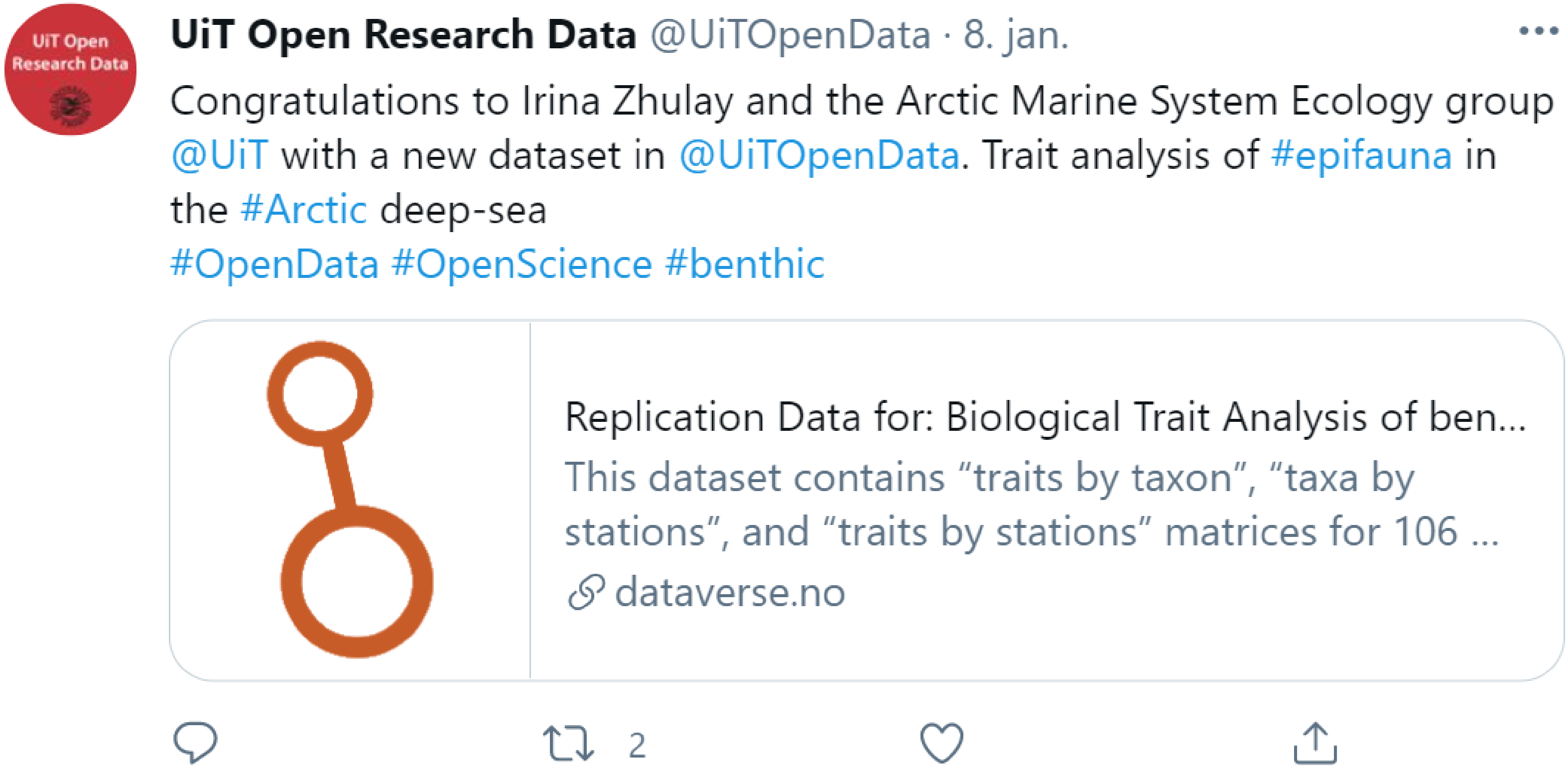
→ Your dataset is still only a draft, and you still may change or delete it. If you would like to grant someone (e.g. a collaborator or a journal editor) access to the unpublished dataset, please contact the support services of your home institution.

→ Once you have entered all the necessary metadata and uploaded all the data files needed, and your are ready to get it published, click on *Submit for Review*. You will receive a notification from DataverseNO that your dataset is submitted.

→ A curator from your archive will review your dataset and if necessary inform you about possible changes to be done before publication. Once the curator has approved your submission, you will be notified, and the dataset will be published and searchable open access.

→ If needed, you can modify metadata and/or data files after publication. This will create a new version of the dataset, which needs to be approved by a curator from your archive, so remember to submit new versions for review. Note that previous versions are not deleted but archived open access.

Promoting published datasets



UiT Open Research Data @UiTOpenData · 8. jan.

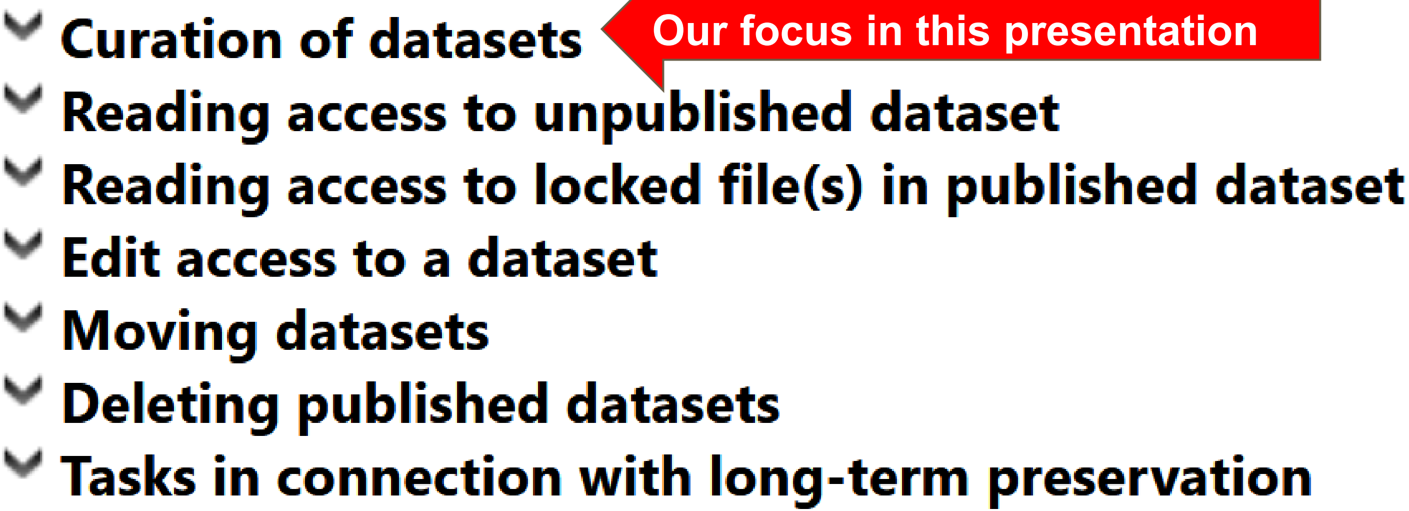
Congratulations to Irina Zhulay and the Arctic Marine System Ecology group @UiT with a new dataset in @UiTOpenData. Trait analysis of #epifauna in the #Arctic deep-sea
#OpenData #OpenScience #benthic

Replication Data for: Biological Trait Analysis of ben...
This dataset contains “traits by taxon”, “taxa by stations”, and “traits by stations” matrices for 106 ...
dataverse.no

Curation support in DataverseNO

Curation Guidelines

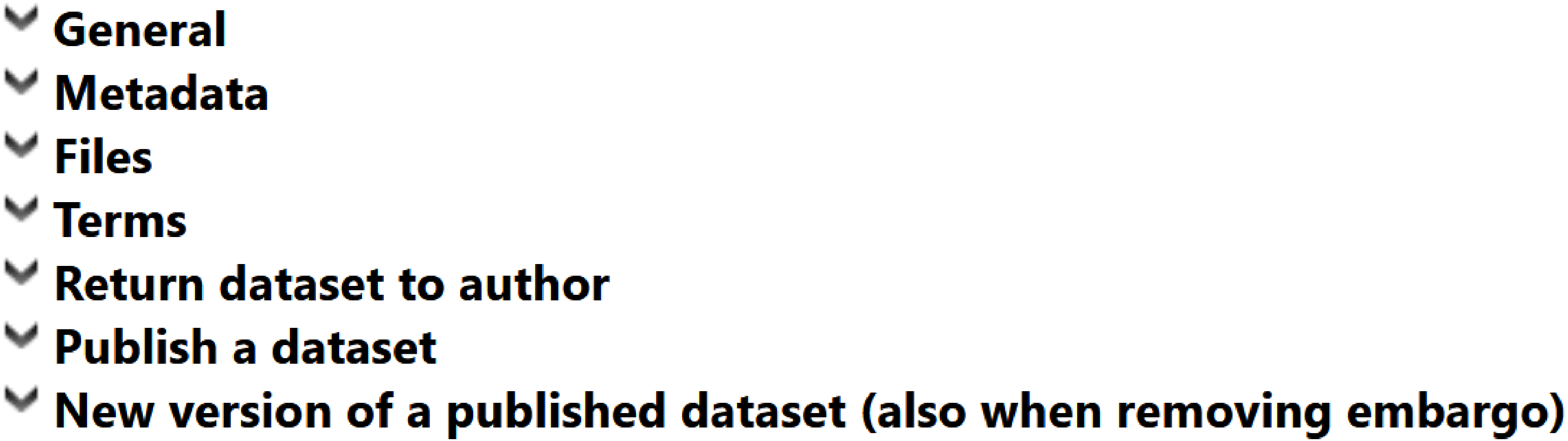
❑ Contain the following main sections:



- Curation of datasets
- Reading access to unpublished dataset
- Reading access to locked file(s) in published dataset
- Edit access to a dataset
- Moving datasets
- Deleting published datasets
- Tasks in connection with long-term preservation

Our focus in this presentation

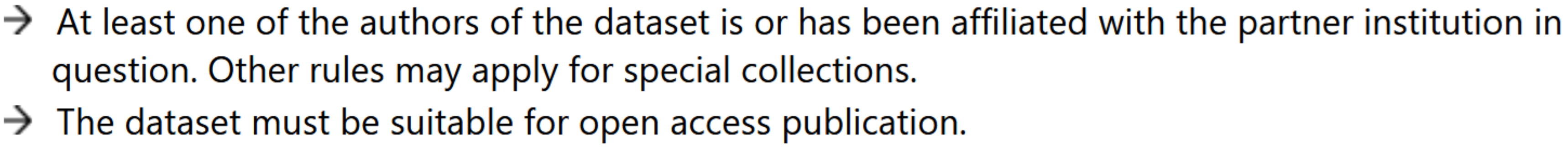
Curation of dataset



- General
- Metadata
- Files
- Terms
- Return dataset to author
- Publish a dataset
- New version of a published dataset (also when removing embargo)

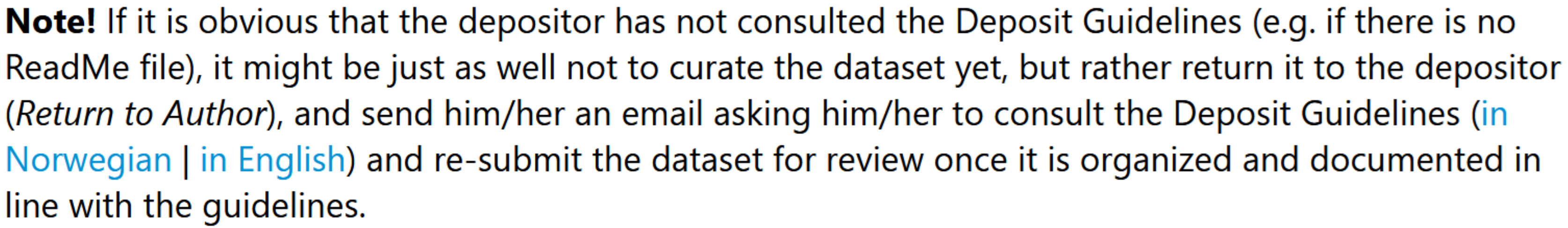
General guidance on curation

- ❑ **Find dataset** to be curated under *Notification* in User Menu.
- ❑ Check **Version tab** to find out whether dataset is new, or new version of previously published dataset.
- ❑ Check whether author and content meet **requirements in DataverseNO Accession Policy**. Most important points summarized:



→ At least one of the authors of the dataset is or has been affiliated with the partner institution in question. Other rules may apply for special collections.

→ The dataset must be suitable for open access publication.

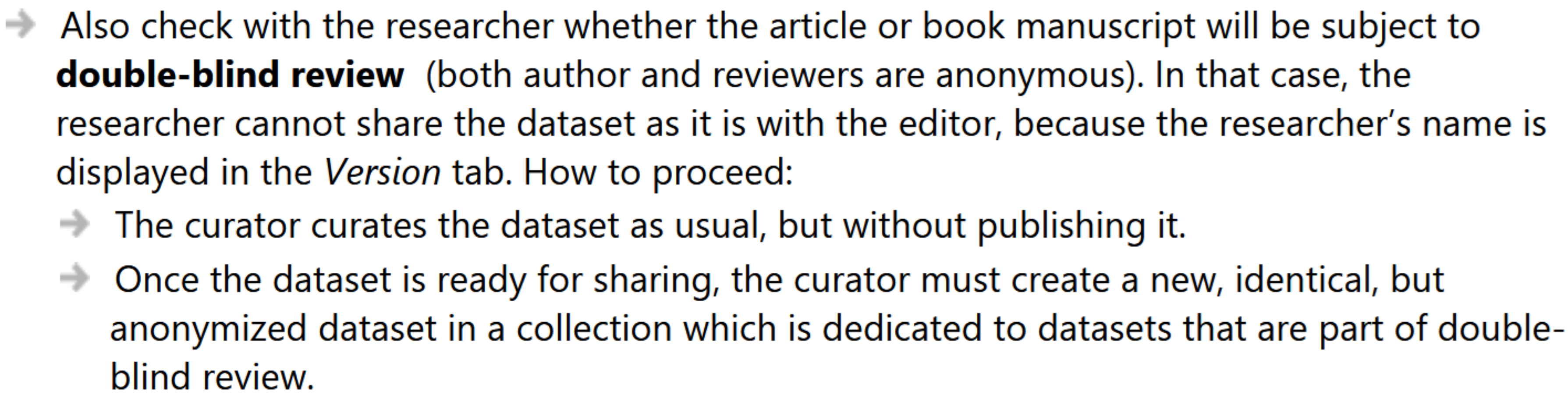


Note! If it is obvious that the depositor has not consulted the Deposit Guidelines (e.g. if there is no ReadMe file), it might be just as well not to curate the dataset yet, but rather return it to the depositor (*Return to Author*), and send him/her an email asking him/her to consult the Deposit Guidelines (in Norwegian | in English) and re-submit the dataset for review once it is organized and documented in line with the guidelines.

Curation of metadata

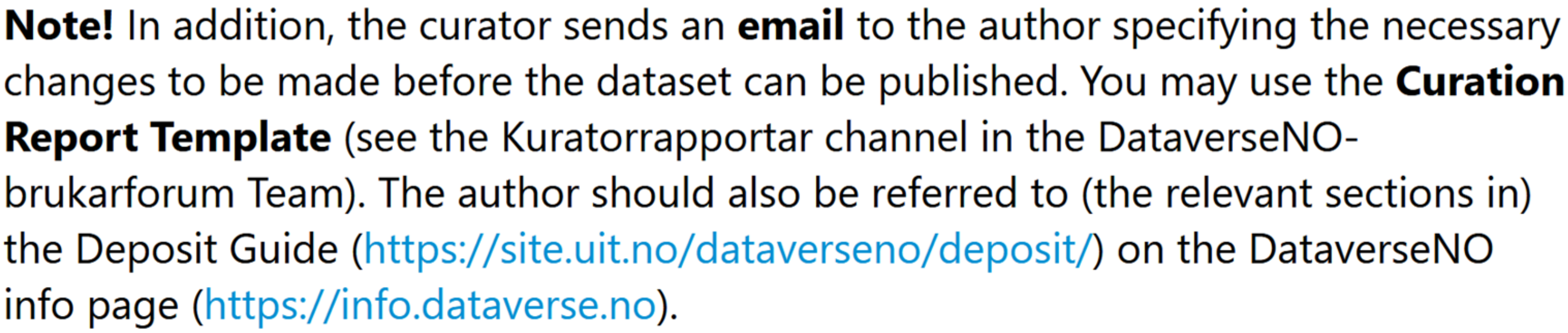
❑ **Basically:** Check whether the depositor has followed the recommendations in the Deposit Guidelines.

❑ **Special attention** to datasets that are going to be part of **double-blind peer review** process:



- Also check with the researcher whether the article or book manuscript will be subject to **double-blind review** (both author and reviewers are anonymous). In that case, the researcher cannot share the dataset as it is with the editor, because the researcher’s name is displayed in the *Version* tab. How to proceed:
  - The curator curates the dataset as usual, but without publishing it.
  - Once the dataset is ready for sharing, the curator must create a new, identical, but anonymized dataset in a collection which is dedicated to datasets that are part of double-blind review.

Return dataset to depositor



Note! In addition, the curator sends an **email** to the author specifying the necessary changes to be made before the dataset can be published. You may use the **Curation Report Template** (see the Kuratorrapportar channel in the DataverseNO-brukarforum Team). The author should also be referred to (the relevant sections in) the Deposit Guide (https://site.uit.no/dataverseno/deposit/) on the DataverseNO info page (https://info.dataverse.no).

DataverseNO Curation Report Template

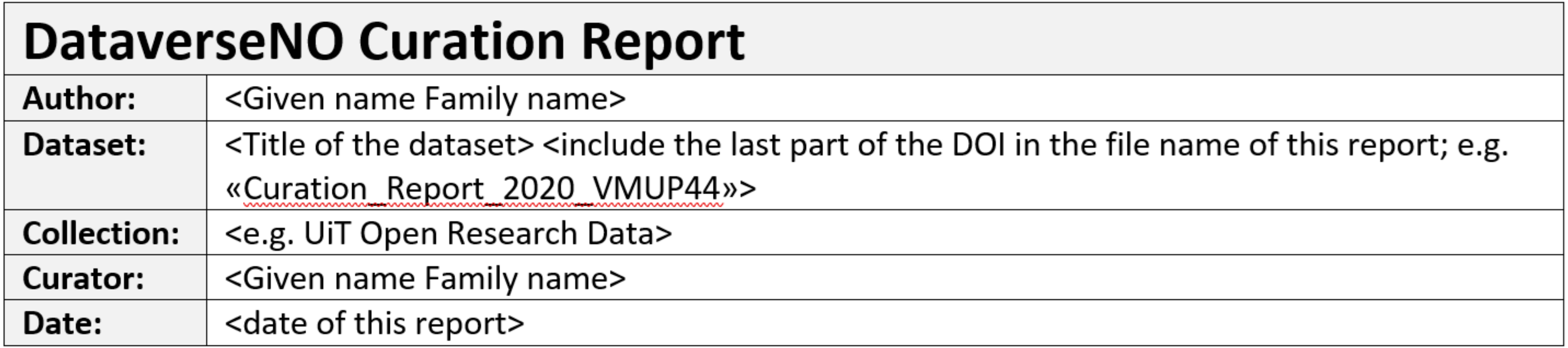
Why to use a **standardized** curation report? >> To make the work of curators easier:

❑ Much of the **information** usually provided in feedback to depositor has to be **repeated in each email**.

❑ Sometimes, some depositors seem to get the impression that the requested changes are “invented” by the individual curator, who is “picky”. A standardized report makes it clear that the changes are necessary because of our guidelines = to make the data as FAIR as possible.

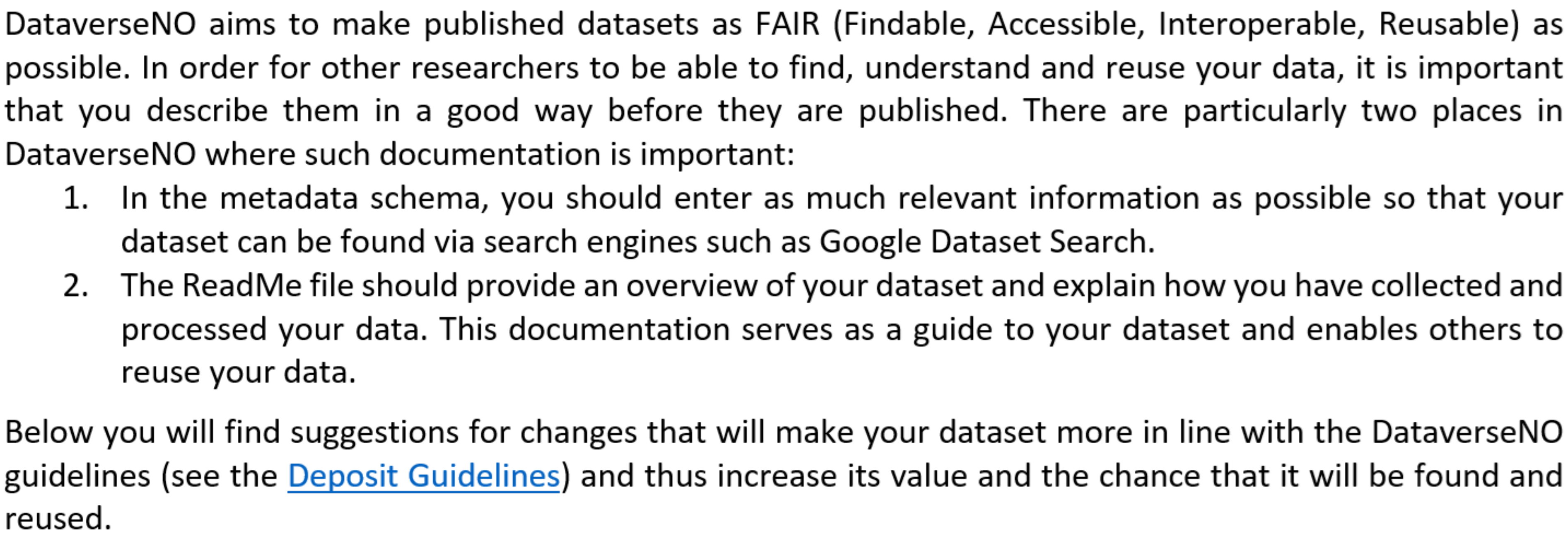
How? >> **Word document** (Norwegian and English version); on Teams

DataverseNO Curation Report Template -- header



DataverseNO Curation Report	
Author:	<Given name Family name>
Dataset:	<Title of the dataset> <include the last part of the DOI in the file name of this report; e.g. «Curation_Report_2020_VMUP44»>
Collection:	<e.g. UiT Open Research Data>
Curator:	<Given name Family name>
Date:	<date of this report>

DataverseNO Curation Report Template -- explain



DataverseNO aims to make published datasets as FAIR (Findable, Accessible, Interoperable, Reusable) as possible. In order for other researchers to be able to find, understand and reuse your data, it is important that you describe them in a good way before they are published. There are particularly two places in DataverseNO where such documentation is important:

1. In the metadata schema, you should enter as much relevant information as possible so that your dataset can be found via search engines such as Google Dataset Search.
2. The ReadMe file should provide an overview of your dataset and explain how you have collected and processed your data. This documentation serves as a guide to your dataset and enables others to reuse your data.

Below you will find suggestions for changes that will make your dataset more in line with the DataverseNO guidelines (see the [Deposit Guidelines]) and thus increase its value and the chance that it will be found and reused.

DataverseNO Curation Report -- sample

Curation training and other support

- UiT provides **training** of collection managers of **new partner institutions**.
- UiT organizes two **annual meetings** where curators from all partner institutions discuss issues relating to curation and collection management.
- Continuous support and discussion in **Teams**. Examples:
  - Questions and answers
  - Sharing of curation reports and other helpful tools and advice
- UiT organizes **workshops and webinars** for collection managers and curators. Examples:
  - January 20-21, 2020: RDA in Norway train-the-trainers workshop for data curators
  - January 23-24, 2020: European Dataverse Workshop 2020
  - March 2, 2021: Webinar on file organization and file formats

Who are the curators?

At the larger universities:

- ❑ Often subject/liaison librarians
- ❑ Many of them with researcher background within the field

At smaller universities/university colleges:

- ❑ Often metadata/senior librarians

At research institutions/centres:

- ❑ ?

Data Curation at UiT The Arctic University of Norway

- Most of the research data management (RDM) support services of the university library at UiT is provided by **subject librarians**.
- In addition our RMD support team includes one Open Access advisor, one metadata librarian, and two IT engineers.
- Currently, we are **18 subject librarians** at UiT, **6 of them** are part of the **RDM support team**.
- The idea is to include more subject librarians in the team as the need arises.
- In addition to their other tasks, the subject librarians are responsible for the following main tasks within RDM support:
  - Teaching RDM courses/webinars
  - Provide guidance on data management plans (DMPs) and other RDM issues
  - Curating datasets within their disciplines

Outline of DAY 2:

- ❏ PRESENTATION:
  Introduction to deposit, publication, and curation support in DataverseNO
  - Deposit and publication workflow
  - Deposit and publication support
  - Curation support

- ❏ **SHORT BREAK (5 MIN.)**

- ❏ DISCUSSION & ACTIVITIES:
  - Questions and comments
  - Challenges for repository managers and curators
  - How to provide scalable deposit support?
  - Collaboration across repositories

Outline of DAY 2:

- ❑ PRESENTATION:
  Introduction to deposit, publication, and curation support in DataverseNO
  - Deposit and publication workflow
  - Deposit and publication support
  - Curation support

- ❑ SHORT BREAK (5 MIN.)

- ❑ **DISCUSSION & ACTIVITIES:**
  - Questions and comments
  - Challenges for repository managers and curators
  - How to provide scalable deposit support?
  - Collaboration across repositories

Questions or comments?

Activities

Go to https://tinyurl.com/CSUC2021

Thank you for your attention!

See you at another Dataverse event!

E.g. at the Dataverse Community Meeting 2021?

References

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

Becker, C., Kulovits, H., Guttenbrunner, 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.

Dataverse. Wikipedia. https://en.wikipedia.org/wiki/Dataverse.

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.