Dataverse Community Meeting 2021
Breakout Session on Curation Workflows
June 17, 2021

Curation Support in DataverseNO

Philipp Conzett
UiT The Arctic University of Norway
ORCID: https://orcid.org/0000-0002-6754-7911
Twitter: @PhilippConzett @DataverseNO #dataverse2021

DataverseNO

Outline of presentation

- ❏ Key facts about DataverseNO
- ❏ Levels of curation
- ❏ Curation workflows
- ❏ Organization of curation support
- ❏ Supporting resources
- ❏ Challenges
- ❏ Desirables

Key facts about DataverseNO

Key facts about DataverseNO

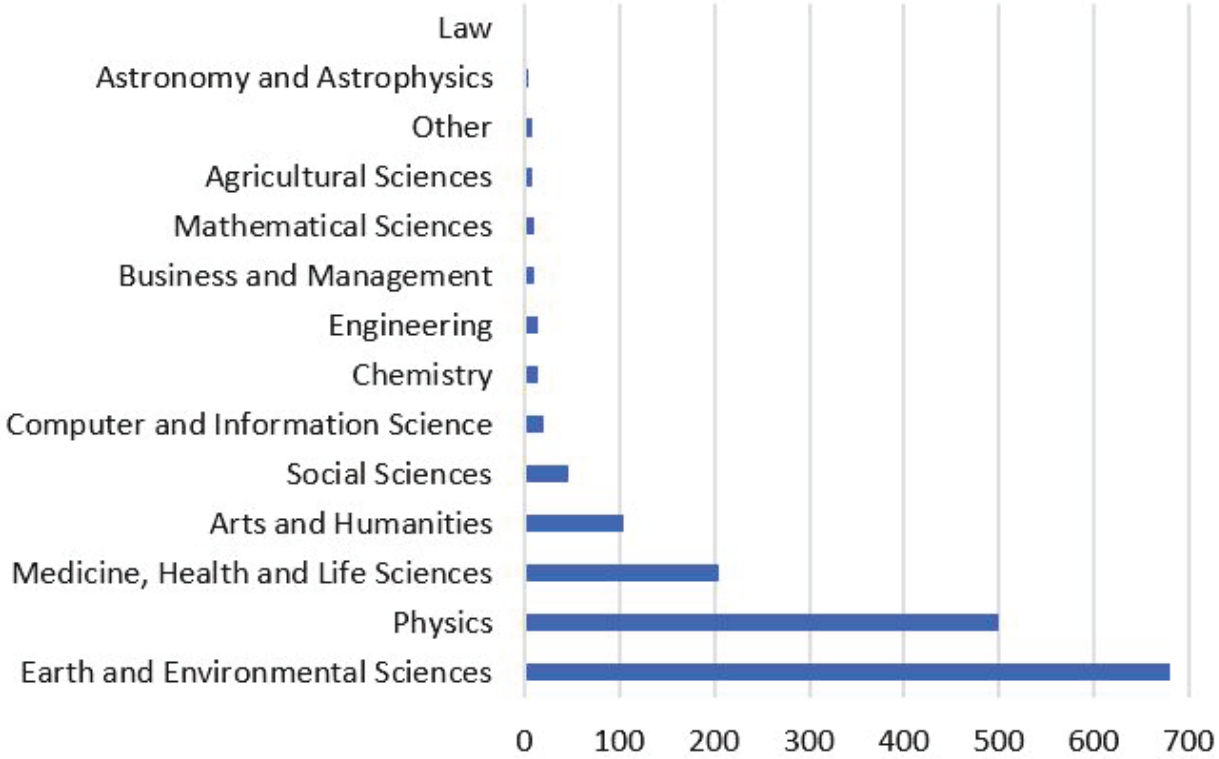
DataverseNO ...

- ❏ is a **national**, **generic** repository for **open** research data from researchers at **Norwegian** research **organizations**;
- ❏ is **curated**, aligned with the **FAIR principles** (cf. Conzett 2020), and **CoreTrustSeal**-certified since 2020;
- ❏ runs on the **Dataverse software**;
- ❏ is operated at **UiT The Arctic University of Norway**.

… a national **repository**

- **Institutional Focused** (cf. Schlatter & Ji, 2021)
- Currently **10 partner institutions** (a new one has just joined us this month…)
- Universities or university colleges
- But also open for (individual) researchers from **other Norwegian research organizations**

… a generic **repository**



- Accepting data from **all domains** of science
- Graph shows distribution of published datasets 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.

Levels of curation

CoreTrustSeal distinguishes between …

… four levels of curation:

A. Content distributed as deposited
B. **Basic** curation – e.g., brief checking, addition of basic metadata or documentation
C. **Enhanced** curation – e.g., conversion to new formats, enhancement of documentation
D. **Data-level** curation – as in C above, but with additional editing of deposited data for accuracy

(From CoreTrustSeal Standards and Certification Board, 2019: 6)

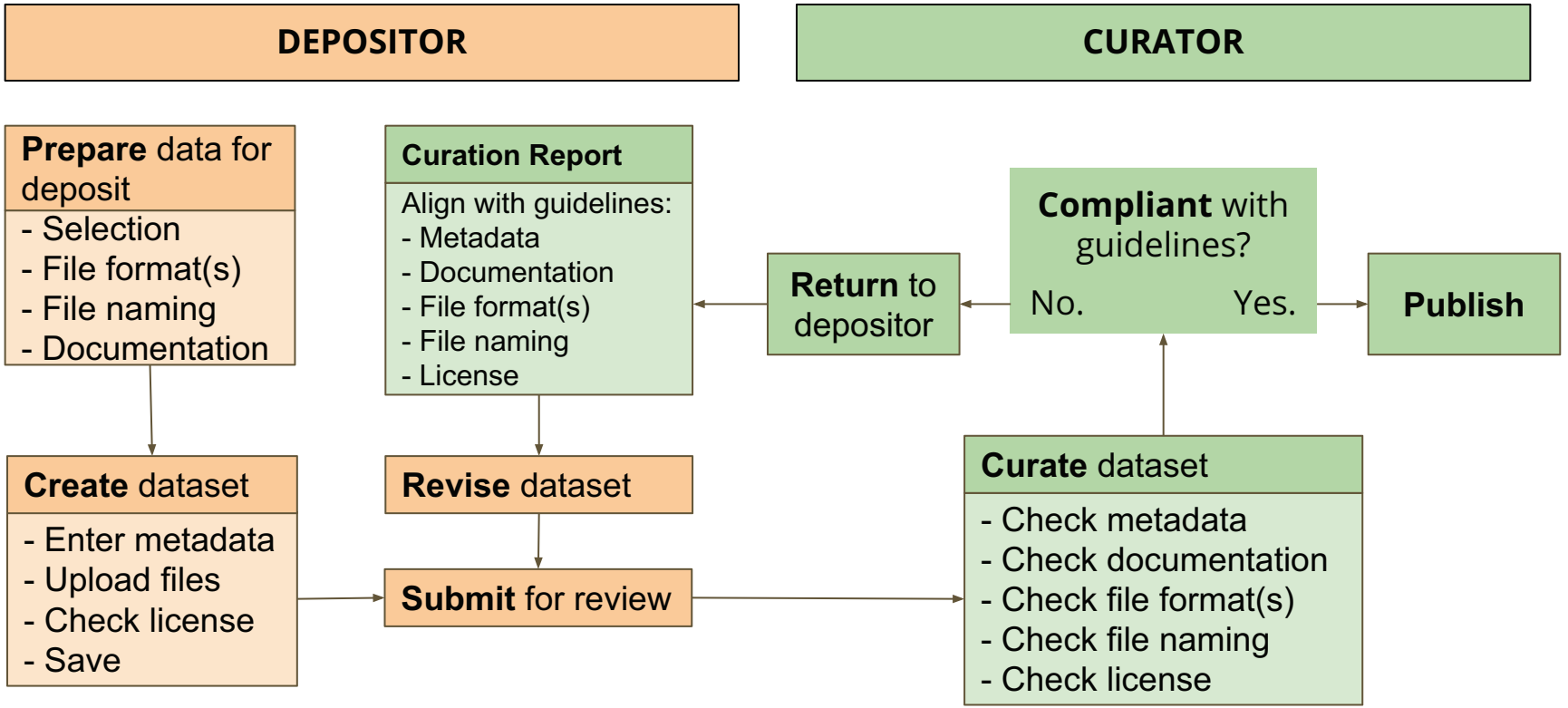
- ❏ We allow level **A** curation if deposited data are compliant with guidelines.
- ❏ We do level **B**, **C** or **D** curation as needed.
- ❏ But level D, i.e. data-level curation, does not include any assessment of the *scientific quality* of the data; we only may comment e.g. missing values in a tabular file, or mismatch between abbreviations as used in a datafile vs. as explained in the documentation.
- ❏ As a main rule **curation through feedback**, i.e. **curators request/suggest** changes, **depositors carry out** changes.
- ❏ Curators may make (minor) changes, which have to be confirmed by depositor.

Curation workflows

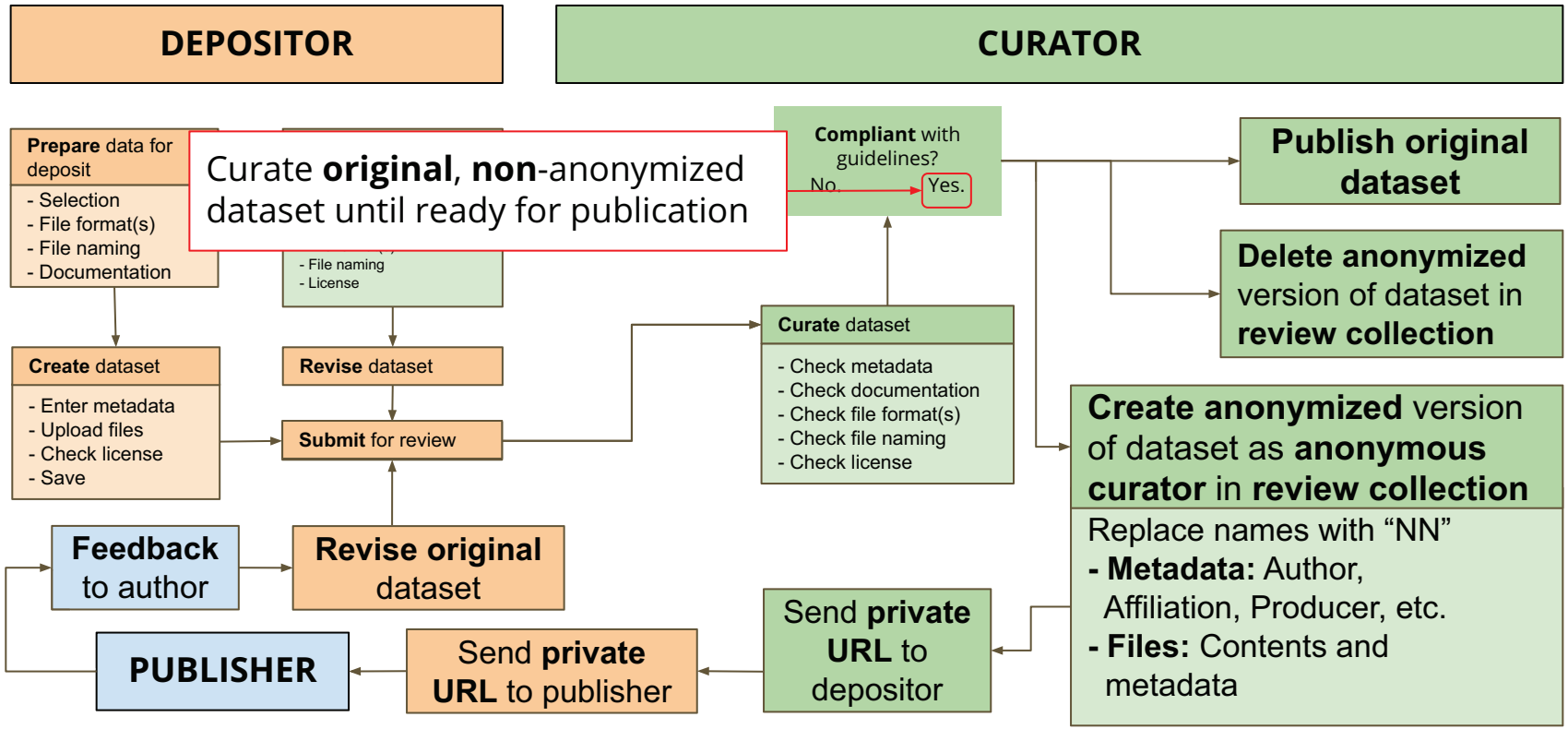
Two main curation workflows

- **Basic** curation workflow
- **Special** curation workflow for datasets to be involved in **double-blind peer review** of publication manuscripts

Basic curation workflow

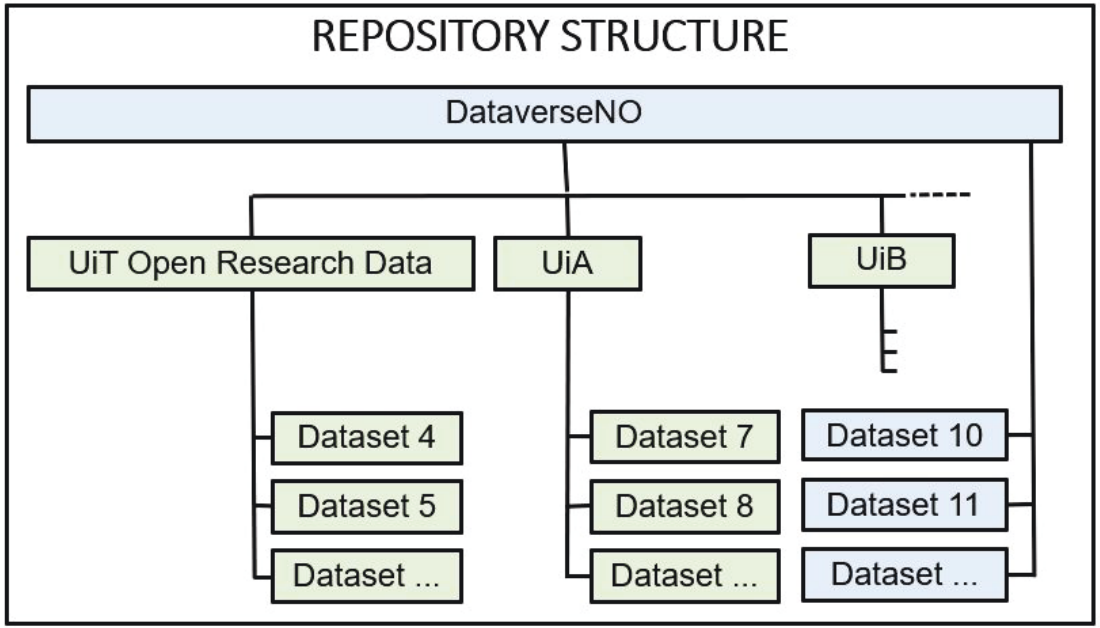


Special curation workflow for double-blind peer review



Organization of curation support

Distributed curation support



- Support staff at **partner organizations** curate datasets within their **institutional collections**.
- Researchers get **local** deposit and publishing **support** at their home institution.

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**:

- ❏ ?

Supporting resources

Three main resources supporting curation

- ❏ **Deposit Guidelines** including README file template (aimed at depositors, but curators need to have in-depth knowledge of this)
- ❏ **Curation Guidelines** including Curation Report Template
- ❏ **Curation Training** and **Curation Network**

>> Brief look at the Curation Guidelines, and Curation Training and Network:

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**

Core tasks = 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)

Curation of dataset

- General
- Metadata
- Files
- Terms

} Basically:
check whether dataset complies with deposit guidelines

- 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 author is **affiliated** with the partner institution in question.
  - Data 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.

Return dataset to depositor

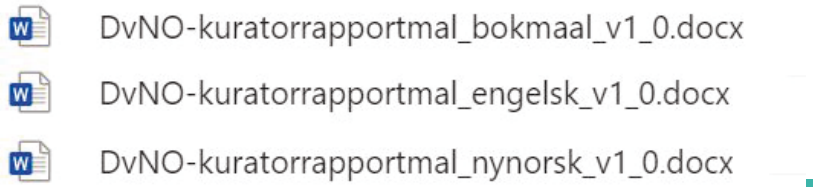
Note! In addition, the curator sends an **email** to the author specifying the necessary changes to be made before the dataset can be published. We recommend you to use the **Curation Report Template** (see the *Kuratorrapportar* channel in the DataverseNO-brukarforum Team). The author should also be 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 use a **standardized** curation report? **Two main reasons**:

- To **make** the **work** of curators **easier**:
  - Much of the **information** usually provided in feedback to depositor has to be **repeated in each email**. This is already included in the template.
  - Depositors may get the impression that the requested changes are "invented" by the individual, "picky" curator.
    A standardized report makes it clear that the changes are necessary because of our guidelines, aiming at making the data as FAIR as possible.
- To **align curation** support **across** institutional **collections** (cf. feedback from CoreTrustSeal certification)

How? >> **Word document** (Norwegian and English version) shared in 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 -- extract of metadata section

METADATA (see the section "Metadata" in the Deposit Guidelines)

Citation Metadata

Title:
If the data form the basis of a publication, you may use the title of the publication and add "Replication Data for: ", "Background Data for: " or a similar prefix in front of the title.

Author – Name:
Use inverted order: Family name, Given name

Author – Identifier:
We recommend adding your ORCID in the *Identifier* field (e.g. 0000-0001-1234-5678). Using an ORCID ensures that your research results are unambiguously linked to you as a researcher. Learn more about ORCID in this video, and get your own ORCID at http://orcid.org/.

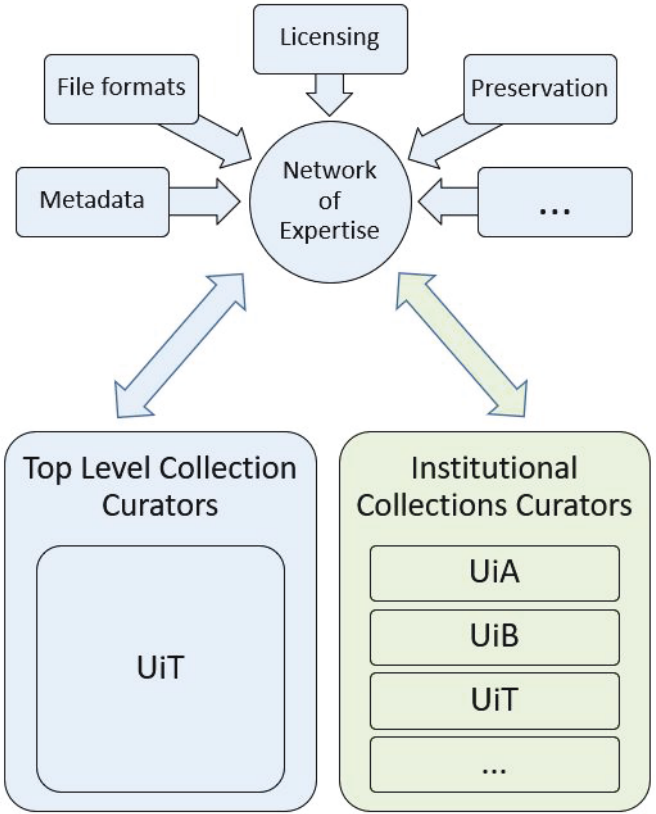
Description:
Next to *Title* and *Keyword*, this is the most important field to make your data findable. Enter as much relevant information as possible. You can also paste the summary you have used in your publication, preferably into a separate *Description* field (click the + sign to the right). If you need to add space between two paragraphs, you can add the HTML tags and around each paragraph.

Three main resources supporting curation

- ❏ **Deposit Guidelines** including README file template (aimed at depositors, but necessary knowledge for curator)
- ❏ **Curation Guidelines** including Curation Report Template
- ❏ **Curation Training and Network**

Curation Network and Training

- UiT provides basic **training** of collection managers and curators at **new partner institutions**.
- UiT organizes two **annual meetings** where curators from all partner institutions discuss issues relating to curation and collection management.
- **Continuous support**, knowledge exchange, 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 2020: RDA in Norway train-the-trainers workshop for data curators
  - January 2020: European Dataverse Workshop
  - March 2021: Webinar on file organization and file formats



Challenges

Alignment of curation support across collections

Main challenge:

- ❏ Need to further **align curation** support **across** institutional **collections**; cf. feedback from CoreTrustSeal.

Main approaches:

- ❏ **Common** Curation **Guidelines** including Curation **Report Template**
- ❏ Plan to **align curation skills** of our support staff with a **national skills framework** being developed by **RDA in Norway**.

Desirables

Some features that would improve curation support

- Include both **mandatory and recommended fields** in the **first round of metadata registration**. Would help us to make depositors register recommended metadata before first curation round.

Some features that would improve curation support (2)

- Common **framework for file formats** (based on suitability for long-term preservation) that individual repositories could endorse in their lists:

Category	Recommendation
A	Preferred for long-term preservation. Recommended whenever possible and appropriate to save data in such a format.
B	Suitable for long-term preservation to only a limited extent. Recommended when alternative A is not possible or appropriate (e.g. due to workload or loss of information).
C	Not suitable for long-term preservation. Accepted only in case alternative A or B are not possible.

Cate-gory	Ownership		Documentation		Coding		Compression		Adoption	
	Open	Proprietary	Open	Closed	Text-based	Binary	No/Loss-less	Yes	Widely used	Rarely used
A	X	(X?)	X		X	(X)	X	(X?)	X	(X)
B		X	X		X	(X)	X	(X?)	X	(X)
C		X		X	X	X	X	X	X	X

(Explanation: X = typical/desirable property of file format in the category at stake; (X) = less typical/desirable property of file format in the category at stake)

Some features that would improve curation support (3)

- More **advanced curation management support** within or integrated with the Dataverse software, including features like
  - Track curation status of datasets in review
  - Integrated feedback within datasets and files (instead of by email)

References

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

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.

Special workflow for double-blind peer review: DataverseNO Curator Guidelines: https://site.uit.no/dataverseno/admin-en/curatorguide/#metadata, see metadata field “Related Publication”.

Thank you for listening!

dataverse.no

@DataverseNO

?

info.dataverse.no