

*EOSC-Nordic FAIRification webinar:
FAIRification STEP 2 – FAIR principle F3
February 3, 2021*

File-level identification support in Dataverse / DataverseNO

Philipp Conzett



UiT The Arctic University of Norway

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

Twitter: @PhilippConzett @DataverseNO

Thanks to

- ❑ the organizers for inviting me to this webinar,
- ❑ the Dataverse community for input to this presentation.

What is DataverseNO?

- ❑ A **national, generic** repository for **open** research data
- ❑ Operated at **UiT The Arctic University of Norway**
- ❑ Currently **9 partner institutions** (universities)
- ❑ Aligned with the **FAIR** principles and **CoreTrustSeal**-certified
- ❑ Runs on the **Dataverse** software

DataverseNO institutions



How does Dataverse / DataverseNO
support file-level identification?

Some preliminary remarks

- ❑ Not sure whether all the features are relevant for the F3 part of FAIR.
- ❑ Including both features that support **human** interaction and features that support **machine** interaction.
- ❑ The overview might still be useful for further discussions about file-level identification support in research data repositories.

How is the support provided?

- ❑ All file-level identification support is provided through the **Dataverse software** (built-in support).
- ❑ DataverseNO runs on the **main distribution** of the Dataverse software (no forks).
- ❑ Some file-level identification support features are **optional at repository level**, i.e. they may be configured (turned on or off) by system admin.
- ❑ If the relevant support feature is activated, file-level identification is either
 - ❑ provided **automatically** (in some cases though dependent on the file type), or it
 - ❑ may be added **manually** by the depositor (and/or curator).

Automatic file-level identification support in Dataverse/DataverseNO

Persistent identification support

Optional at repository level. Activated in DataverseNO.

- ❑ Each file is assigned a **DOI**.
- ❑ Each new version of a file gets its own DOI, consisting of the same DOI **prefix** and **dataset-level suffix**, but a different **file-level suffix**, e.g.:

... <https://doi.org/10.18710/QA8WBZ/7CV2MO>, DataverseNO, V1

... <https://doi.org/10.18710/QA8WBZ/KSOVFW>, DataverseNO, V2

More granular identification support for tabular files

Dependent on file format, e.g. Rdata, Stata, Excel, .csv.

- ❑ Each file is assigned a **Universal Numeric Fingerprint (UNF)**, i.e. semantic checksum which identifies the content of the file independently of the file format.
- ❑ Each file is **ingested at variable level**, meaning that
 - ❑ **variable-level information / metadata** is extracted and stored, including the number of variables and number of observations, and
 - ❑ the variable metadata are **indexed and therefore findable**.

More granular identification support for Flexible Image Transport System (FITS) files

Dependent on file format: .FITS. So far, no examples in DataverseNO.

- ❑ **Metadata** found in the header section are **extracted, aggregated and displayed** in the Astronomy Domain-Specific Metadata of the dataset that the file belongs to.
- ❑ This FITS file metadata is therefore **searchable and browsable** (facets) at the dataset level.

Citation support

- ❑ Each file gets its FORCE11-aligned reference including information about author, publication year, file name, dataset title, file-level DOI, dataset version, and file UNF, e.g.:

Stamm, Johann, 2021, "storj.tab", Programming code for article "Radar imaging with EISCAT 3D", <https://doi.org/10.18710/QRDET2/OZ5XJT>, DataverseNO, V1, UNF:6:VbsDGr5ENhKr98xIVKzNCA== [fileUNF]

Verification support

- ❑ Each file is assigned an **MD5 checksum**.
- ❑ Tabular files (with certain file formats) are assigned a **Universal Numeric Fingerprint (UNF)**.

Media type identification support

- ❑ For files with common file formats, the media type (MIME type) is identified and registered.

Manual file-level identification support

File metadata

Each file may be enriched with **file-level metadata** including



- ❑ a free-text **description**,
- ❑ a **file tag** characterizing the type of file (e.g. data, code, documentation),
- ❑ **provenance information** in the form of an uploaded file in JSON format and following the W3C PROV standard, and – additionally – as a free-text description.

Access restriction

- ❑ Access restrictions may be defined for each file, including terms of access, and terms of reuse.

File hierarchy

- A “file path” can be specified for files. That way, a hierarchy of files can be downloaded, e.g.:

- ▶  1_SizeST
- ▶  2_ZetapotentialST
- ▶  3_FluoST
- ▶  4_CellTesting
- ▶  5_FluorescenceTracing
- ▶  00_ReadMe.txt (14.7 KB)

(Source: <https://doi.org/10.18710/5E8VUI>)

Some issues and desirables

Issues and desirables

- ❑ Dataset-level DOI versioning is not in place yet.
- ❑ Some of the file-level identification information can not be correctly harvested / indexed e.g. through OAI-PMH or Schema.org.
- ❑ Some of the features may not be machine-actionable yet.
- ❑ Need more granular license support, both at dataset and file level.

DataverseNO has been / is being assessed by EOSC-Nordic and FAIRsFAIR and commits to continuously improving its FAIRness.

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)