# Integrating DMP Tools and Data Repositories

Maximilian Moser
Center for Research Data Management, TU Wien, Austria

Invenio Module: https://invenio-madmp.readthedocs.io/en/latest/
Handout: https://bit.ly/32CPAgN

# Background

Software Developer at TU Wien
for the FAIR Data Austria project

Working on an integration between our maDMP tool and Data Repository (Invenio RDM)
for synchronizing information between both tools

Goal: Extract tool-independent framework for integration with a use of maDMPs

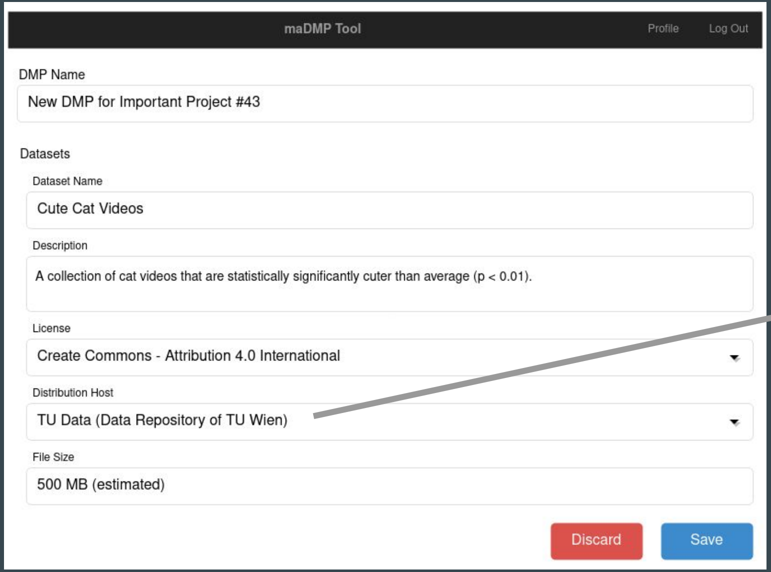
# Use Cases

- Creating new DMPs in a DMP tool triggers creation of deposit drafts with partially pre-filled metadata in a Data Repository
- Creating new deposits in a Data Repository triggers addition of datasets in DMP maintained by a DMP tool
- Updates to dataset fields in either tool should be reflected in the other tool as well

Feasibility has already been tested, now we want to do it *application-independently* in production

- https://projects.iq.harvard.edu/dcm2020/agenda
- https://rd-alliance.org/group/repository-platforms-research-data-ig/post/rda-hackathon-madmps-lets-connect-repository

# Use Case 1 - Automatic Creation of Deposit Drafts



DMP Name

New DMP for Important Project #43

Datasets

Dataset Name

Cute Cat Videos

Description

A collection of cat videos that are statistically significantly cuter than average ($p < 0.01$).

License

Create Commons - Attribution 4.0 International

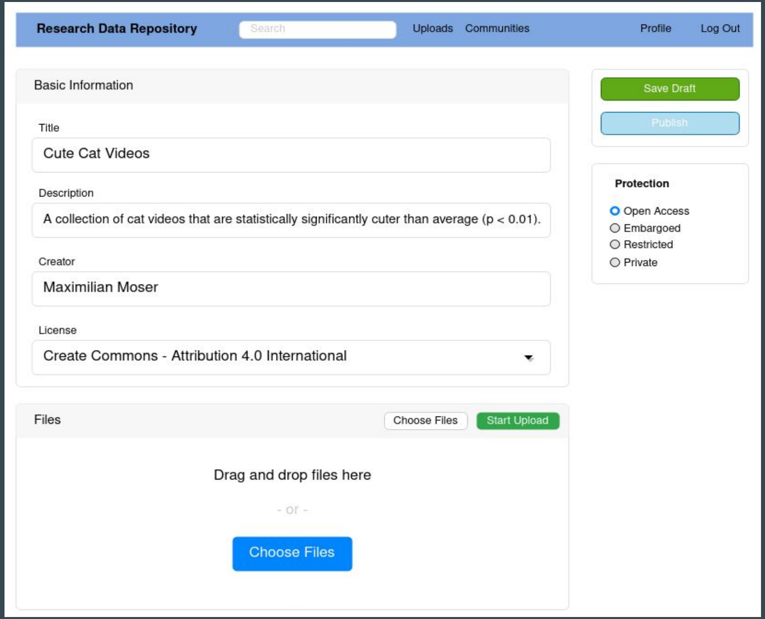
Distribution Host

TU Data (Data Repository of TU Wien)

File Size

500 MB (estimated)

Any Tool for creating DMP, compliant to RDA CS



Basic Information

Title

Cute Cat Videos

Description

A collection of cat videos that are statistically significantly cuter than average ($p < 0.01$).

Creator

Maximilian Moser

License

Create Commons - Attribution 4.0 International

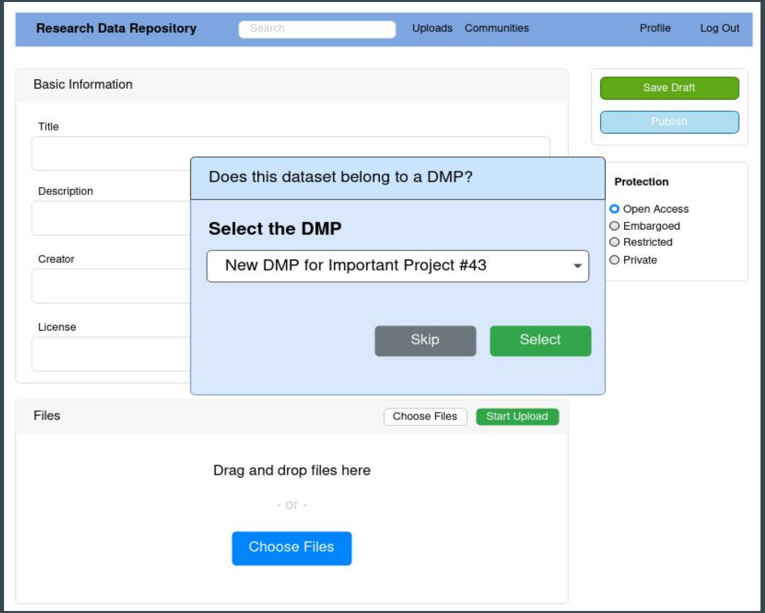
Files

Protection

● Open Access
○ Embargoed
○ Restricted
○ Private

Any Data Repository

# Use Case 2 - Addition of New Datasets



Basic Information

Title

Description

Creator

License

Does this dataset belong to a DMP?

**Select the DMP**

New DMP for Important Project #43

**Protection**

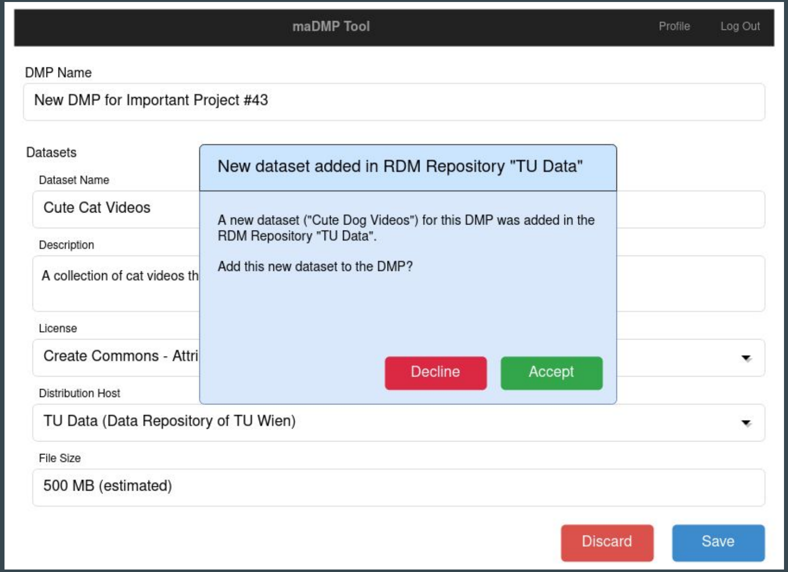
● Open Access
○ Embargoed
○ Restricted
○ Private

Files

Drag and drop files here

- or -

Any Data Repository



DMP Name: New DMP for Important Project #43

Datasets

Dataset Name: Cute Cat Videos

Description: A collection of cat videos th

License: Create Commons - Attri

Distribution Host: TU Data (Data Repository of TU Wien)

File Size: 500 MB (estimated)

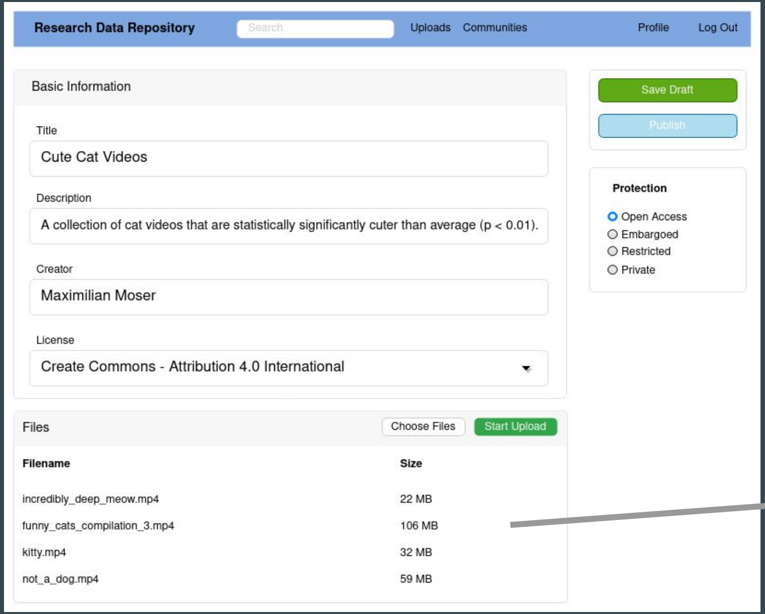
New dataset added in RDM Repository "TU Data"

A new dataset ("Cute Dog Videos") for this DMP was added in the RDM Repository "TU Data".

Add this new dataset to the DMP?

Any Tool for creating DMP, compliant to RDA CS

# Use Case 3 - Information Synchronization



Basic Information

Title: Cute Cat Videos

Description: A collection of cat videos that are statistically significantly cuter than average ($p < 0.01$).

Creator: Maximilian Moser

License: Create Commons - Attribution 4.0 International

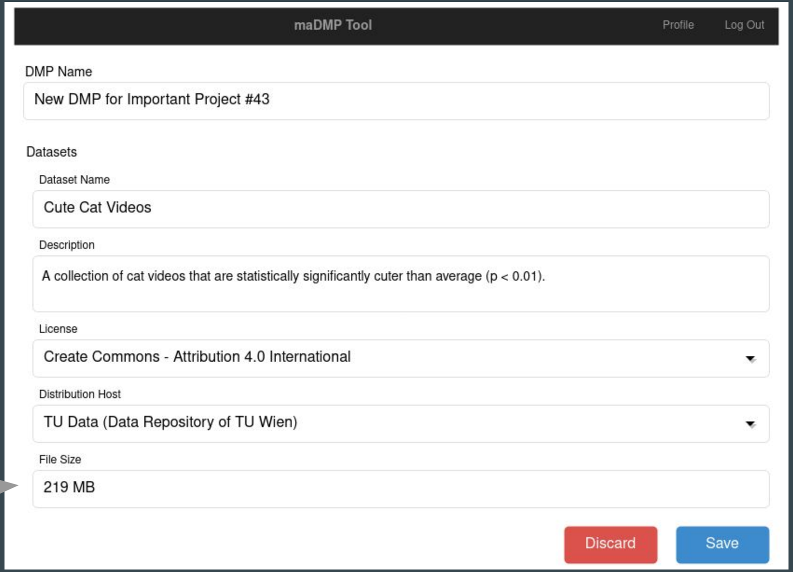
Protection

- ● Open Access
- ○ Embargoed
- ○ Restricted
- ○ Private

Files

| Filename | Size |
|---|---|
| incredibly_deep_meow.mp4 | 22 MB |
| funny_cats_compilation_3.mp4 | 106 MB |
| kitty.mp4 | 32 MB |
| not_a_dog.mp4 | 59 MB |

Any Data Repository



DMP Name: New DMP for Important Project #43

Datasets

Dataset Name: Cute Cat Videos

Description: A collection of cat videos that are statistically significantly cuter than average ($p < 0.01$).

License: Create Commons - Attribution 4.0 International

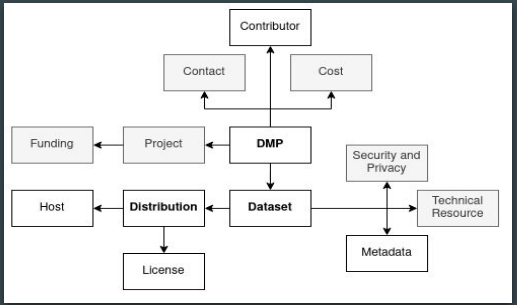
Distribution Host: TU Data (Data Repository of TU Wien)

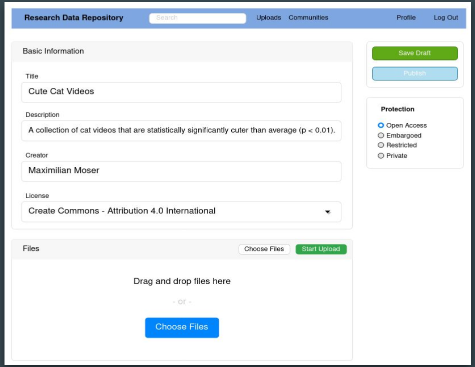
File Size: 219 MB

Any Tool for creating DMP, compliant to RDA CS

# Challenges

- Mapping between maDMP and Data Repositories
  - ○ "how do we translate between two worlds?"
  - ○ e.g. distributions vs. records





Basic Information

Title: Cute Cat Videos

Description: A collection of cat videos that are statistically significantly cuter than average ($p < 0.01$).

Creator: Maximilian Moser

License: Create Commons - Attribution 4.0 International

Protection
- ● Open Access
- ○ Embargoed
- ○ Restricted
- ○ Private

Files

- Identification of specific Dataset Distributions
  - ○ how do you reference something without ID?
  - ○ make it unique: limit number of distributions per host for each dataset to one

# Challenges

- Modification Scope
  - “who is allowed to change what?”
  - e.g. repository can only change its distribution’s properties

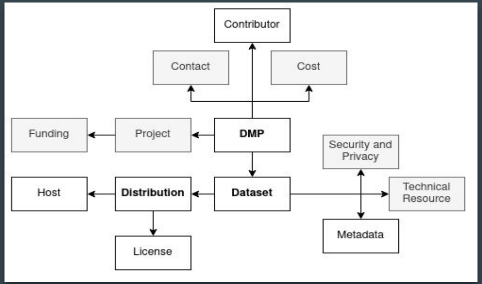
- Request Format
  - “what language do we speak?”
  - JSON Patch vs. JSON

- Ownership management of created deposits?
  - who is responsible, and how should they do it?

```
{
 "op": "replace",
 "path": "/dmp/title",
 "value": "Funny title!"
}
```

or

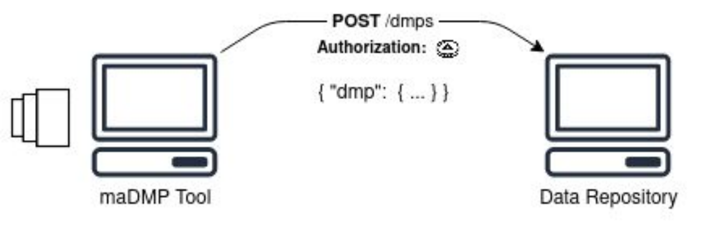
{

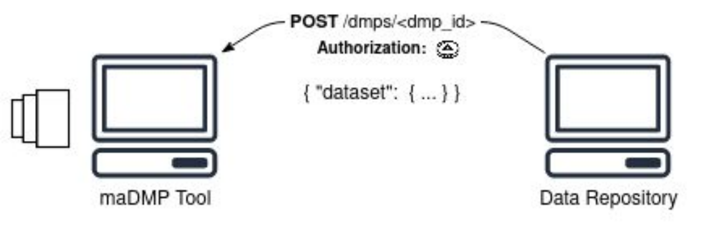


}

# Communication Between DMP Tool and Data Repository

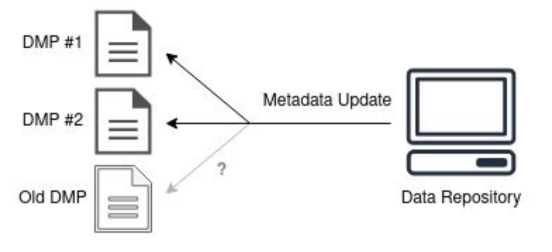
- Request Body: JSON (DMP or Dataset)
  - largely POST or PATCH semantics
- Two-way communication required
- Only authorized tools can participate
  - shared secret can also be used by CRIS system to modify the same set of DMPs
- DMP versioning done in the DMP tool





# Open Questions

- Should metadata changes[1] be reflected in all DMPs (re-)using the dataset?
  - even in DMPs for projects that are already finished?



[1] Record metadata can be edited in Invenio and Zenodo, creating so-called *revisions*

# Conclusions

We found solutions for some challenges, including:

- how do we map the concepts between RDA DMP and Invenio?
- communication: when, what and how?
- limit the modification scope of integrated applications

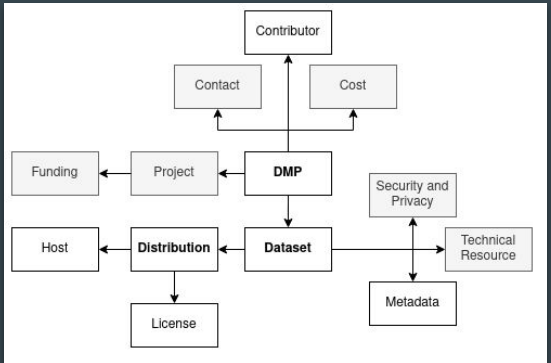
However, we still have some open questions

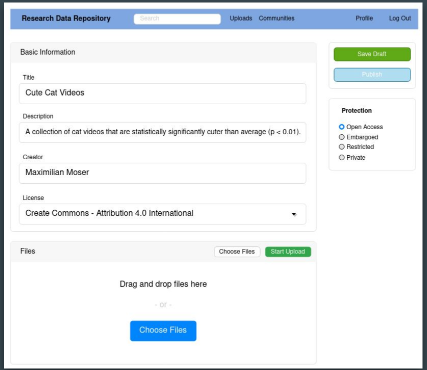
- If you encounter similar engineering challenges, please let us know!

**maximilian.moser@tuwien.ac.at**

# Challenge 1 - Mapping Between maDMP and Data Repos

- many RDA DMP JSON fields are irrelevant for us and can be ignored
- DMPs usually have no “native” counterpart in RDM repositories
  - for the integration though, we should at least remember their IDs
- Datasets are too abstract to be mapped directly to RDM Records
- Distributions, however, essentially *are* concrete dataset deposits





Basic Information

Title: Cute Cat Videos

Description: A collection of cat videos that are statistically significantly cuter than average (p < 0.01).

Creator: Maximilian Moser

License: Create Commons - Attribution 4.0 International

Protection

● Open Access
○ Embargoed
○ Restricted
○ Private

Files

Drag and drop files here

- or -

# Challenge 2 - Identifying Distributions

- Distribution has no required unique identifier in the standard

- access_url is unique, but not always set
  - e.g. when a new DMP is created and distributions are still only planned

- introduce limitation: each dataset has at most one distribution on each host
  - if the same data repository should host multiple distributions of the same dataset (e.g. in different formats), they should be added in the same deposit (i.e. record)
  - makes references by **dataset_id** and **host** unambiguous
  - does not change the current version of the standard

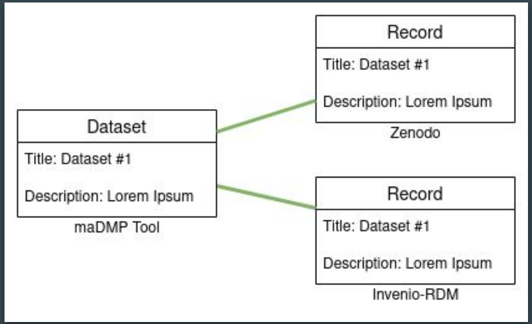
# Challenge 3 - Request Format

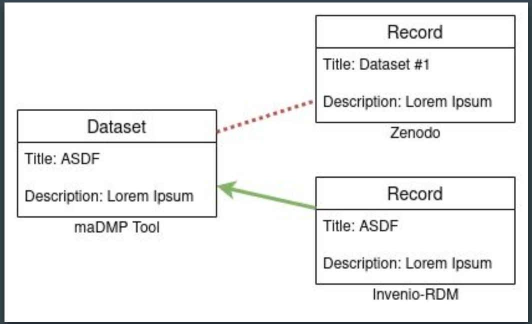
- JSON Patch?
    - uses JSON Pointer to find items
    - we would require something more like XPath

- RDA DMP Common Standard JSON
    - using custom semantics
        - largely PATCH semantics
        - some exceptions, e.g. list of datasets
            - find items by their ID, track new/removed/updated datasets

# Challenge 4 - Modification Scope

- Updates from Data Repositories can only change properties for their Distributions
    - exceptions: dataset_identifier and metadata (both pure additions, though)
    - otherwise, conflicts could arise with non-integrated data repositories:


Record
Title: Dataset #1
Description: Lorem Ipsum
Zenodo
Dataset
Title: Dataset #1
Description: Lorem Ipsum
maDMP Tool
Record
Title: Dataset #1
Description: Lorem Ipsum
Invenio-RDM


Record
Title: Dataset #1
Description: Lorem Ipsum
Zenodo
Dataset
Title: ASDF
Description: Lorem Ipsum
maDMP Tool
Record
Title: ASDF
Description: Lorem Ipsum
Invenio-RDM