# Archiving datasets in Areda: a guide

Version 2.6, July 2023

*Important: Areda is currently only accessible with support of a data steward. Please check the instructions to see if your dataset meets the requirements for archiving in Areda at the moment as we improve the system. If you have any questions about this notice or Areda, please contact the data steward in your faculty.*

This is a guide for archiving the datasets of your research in Areda, the UT data archive. The guide contains the following information:

For general information about data archiving at the UT, see the RDM website.

## I. Important to know before using Areda

- Areda should **not** be used for datasets that still might change (dynamic data), or need to be used regularly.
- For the moment, there is a maximum upload of 1 TB.
- Access to datasets is at research group level: all group members have access to the datasets in the group's bucket (folder). Groups and members are based on the HR system. Students do not have access rights to Areda. See section III for more information about access to archived datasets.
- In case of uploading personal or other confidential data, access has to be managed by using encryption. Areda offers an encryption instruction. Encryption keys need to be securely managed in the research group, preferably based on an internal key management policy or guideline.
- Areda should **not** be used for archiving digital informed consent forms or pseudonymization keys. These must be stored encrypted and separately from the anonymized (or pseudonymized) data, for instance on the p-drive.
- Although in the group's bucket a folder structure (by creating a path) is possible, it is advised to use it as a single archive (database) of zip-files containing a structured set of data files of a research project.
- For making overviews and searching of datasets, use the UT Research Information Portal.

## II. Four steps to be taken when archiving datasets

The process of archiving datasets in Areda consists of the following steps:

1. Preparation: organizing the data files, writing documentation in a README file, creating an archive file, using encryption if needed

2. Uploading the archive file(s) to Areda
3. Adding metadata and README file in UT Research Information System (Pure)
4. Metadata review and final check

Make sure that step 2, 3 and 4 are completed within 4 weeks. Datasets in the research group’s *intake* bucket (see below) will be deleted automatically after 31 days.

## 1. Preparation

Archiving of datasets demands some preparation: selecting and organizing the files with data and related materials, writing data documentation and creating an archive file, including encryption in case of personal or other confidential data.

### Organize files

With regard to selecting and organizing the files with data and related materials to be preserved, pay attention to the following:

- Be sure that the data and related materials (software, models, scripts, code books, etc.) are properly selected and complete.

- In general, only archive anonymized data. If necessary, pseudonymized personal data can be archived in Areda, but the pseudonymization key file must be encrypted and stored *outside* Areda, for instance in a special folder on the p-drive.

- In a data file only include data; put figures and/or analyses derived from/based on these data in separate files.

- Especially for convenient reuse, consider aggregating data into fewer, larger files, rather than many small ones. It is more difficult and time consuming to manage many small files and easier to maintain consistency across data sets with fewer, larger files. It is also more convenient for other users to select a subset from a larger data file than it is to combine and process several smaller files. On the other hand, very large files may exceed the capacity of some software packages. Some examples of ways to aggregate files include by data type, location, time period, measurement platform, investigator, method, or instrument.

#### *File naming*

For files with data and related materials, as well as folders and the archive file (see below: Create an archive file), consider the following naming conventions:

- Use a unique and meaningful name, with a clear and consistent structure, e.g. ProjectName_YYYYMMDD_ContentDescription. There is a maximum of 255 characters.

- Do ***not*** use spaces.

- Do ***not*** use & ” ’ \ / ? ! { } ( ) * + [ ] | : ; @ $ # % ^ ~` <space><newline><tab>.

- Do ***not*** include any information related to an identified or identifiable natural person.

### Write dataset documentation

Good documentation is essential for verification and possible reuse of the research data. Therefore, before you archive the datasets write or update this documentation in a README file (example, see also Making data FAIR). **It is recommended to have this documentation reviewed by a colleague in the research group who is *not* involved in your research**. Key question for this review can be whether it would be possible to verify, reuse or reproduce the data based on this documentation.

## Create an archive file

Create an archive file (zip or tar) of the selected files with data and related materials, **including the reviewed README file with documentation**. The zip or tar file can be compressed to save space and upload/download time. If your dataset is very large (over 100GB), consider creating two or more zip/tar files, each of these containing a set of files that logically belong together (if possible). Put the archive files in one folder which you can upload to Areda.

### *Tar*

The de facto standard for making an archive file is tar. Tar is available for all platforms, using the command prompt.

Under Windows you can make a tar file in the directory d:\data\mydataset as follows:

Press Start, type cmd and open the Command Prompt app.

Type the commands. (As an example your dataset is stored at drive d, in the directory 'data'.)

```
cd \data

tar -cvf mydataset.tar mydataset
```

With tar the default is *not* to compress. If you need compression, add the option 'z' to the last command in the following way:

```
tar -czvf mydataset.tgz mydataset
```

For compressed tar files the extension .tgz is used.

For more options when creating a tar file, see this manual.

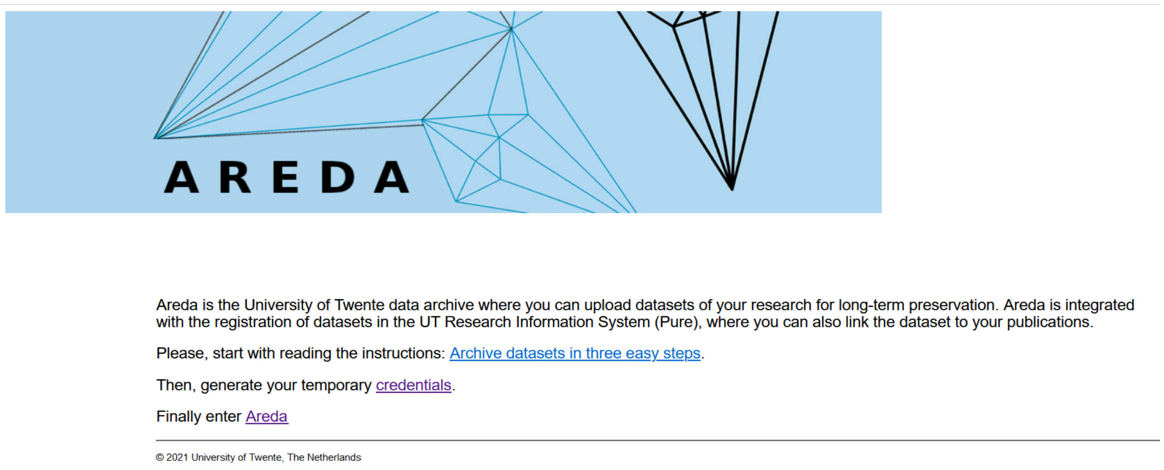
### *Zip*

You can also create an archive file making use of 7-zip. If you want to use another zip tool, make sure that it guarantees creating files that can be properly unzipped also after many years. For this reason, tools like WinZip or the built-in zip utility in Windows are ***not*** a good alternative.

Zip tools, when using the default options, always compress your data. However, compression is not always recommended, because for larger datasets this takes extra CPU time when creating the archive file. Moreover, individual files that are already compressed (e.g. jpg, mpg, mp3, flac) can hardly be compressed any further with zip.

## 2. Uploading (encrypted) archive files

Go to Areda to upload the archive file(s). You will see the following screen:



Areda is the University of Twente data archive where you can upload datasets of your research for long-term preservation. Areda is integrated with the registration of datasets in the UT Research Information System (Pure), where you can also link the dataset to your publications.

Please, start with reading the instructions: Archive datasets in three easy steps.

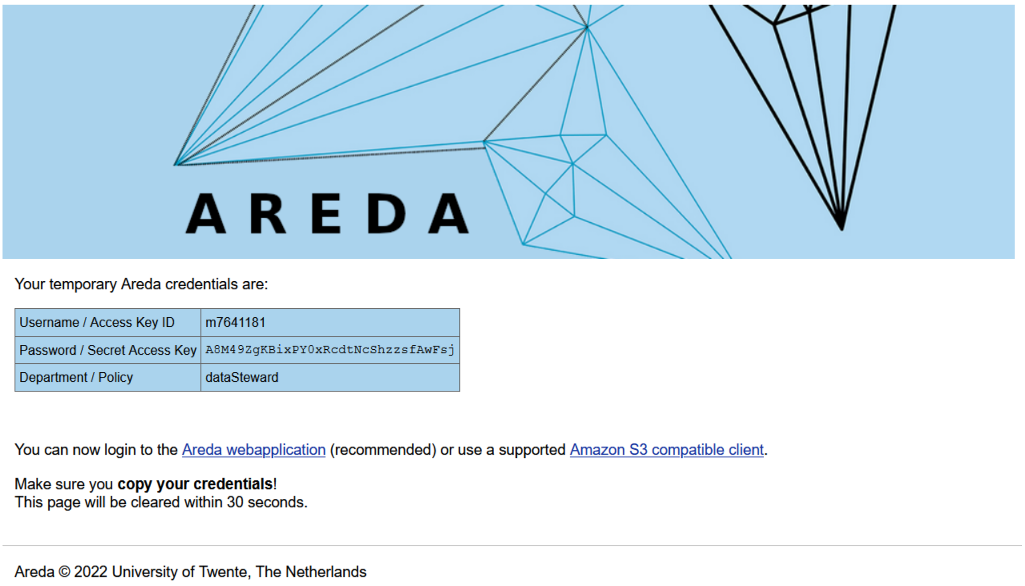
Then, generate your temporary credentials.

Finally enter Areda

© 2021 University of Twente, The Netherlands

Click on credentials and login with your UT account.

In the next screen (see below), copy the Password/Secret Access Key and go back to the Areda webapplication via the recommended link.



Your temporary Areda credentials are:

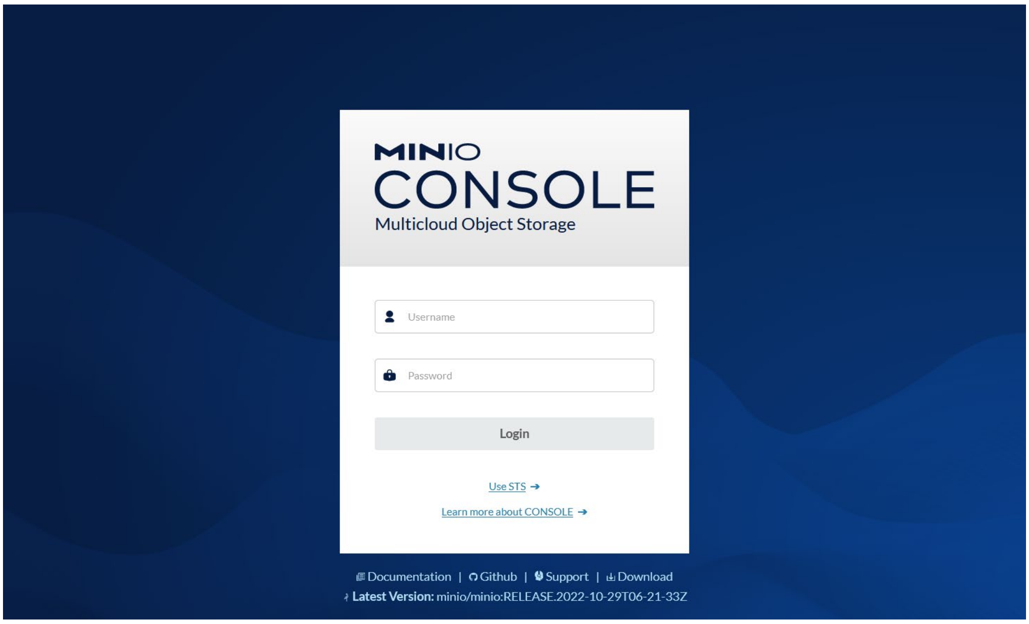
| | |
|---|---|
| Username / Access Key ID | m7641181 |
| Password / Secret Access Key | A8M49ZgKBixPY0xRcdtNcShzzsfAwFsj |
| Department / Policy | dataSteward |

You can now login to the Areda webapplication (recommended) or use a supported Amazon S3 compatible client.

Make sure you **copy your credentials**!
This page will be cleared within 30 seconds.

Areda © 2022 University of Twente, The Netherlands

Click on the Areda-link and you will see the screen below.

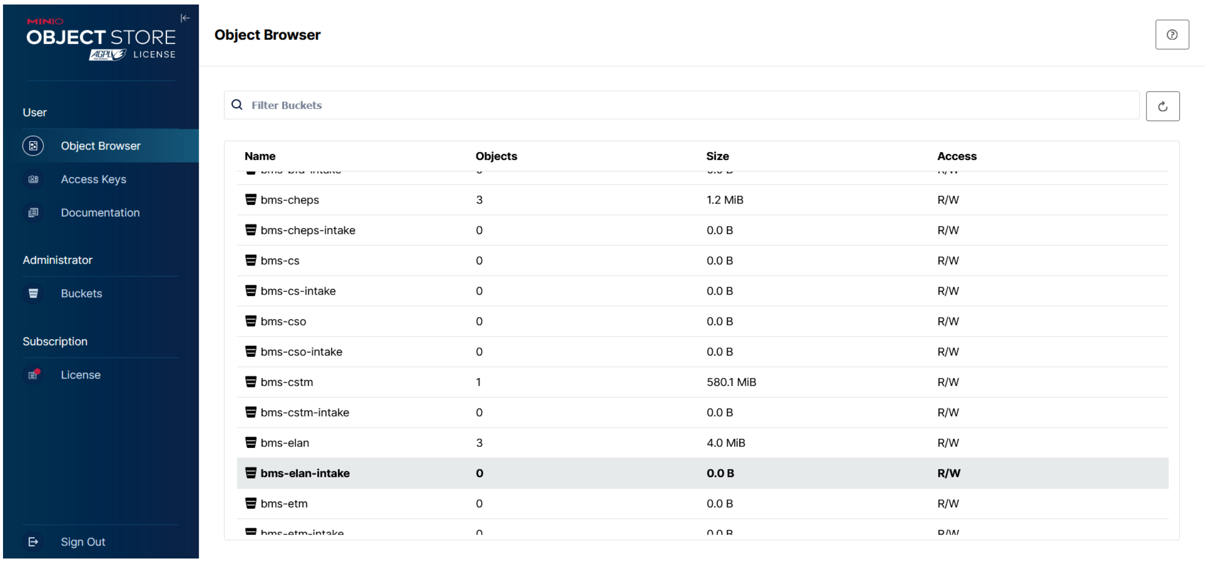

MINIO
CONSOLE
Multicloud Object Storage
Username
Password
Login
Use STS
Learn more about CONSOLE
Documentation | Github | Support | Download

Latest Version: minio/minio:RELEASE.2022-10-29T06-21-33Z

Enter your Username / Access Key ID (m-number) and paste the Password / Secret Access Key.

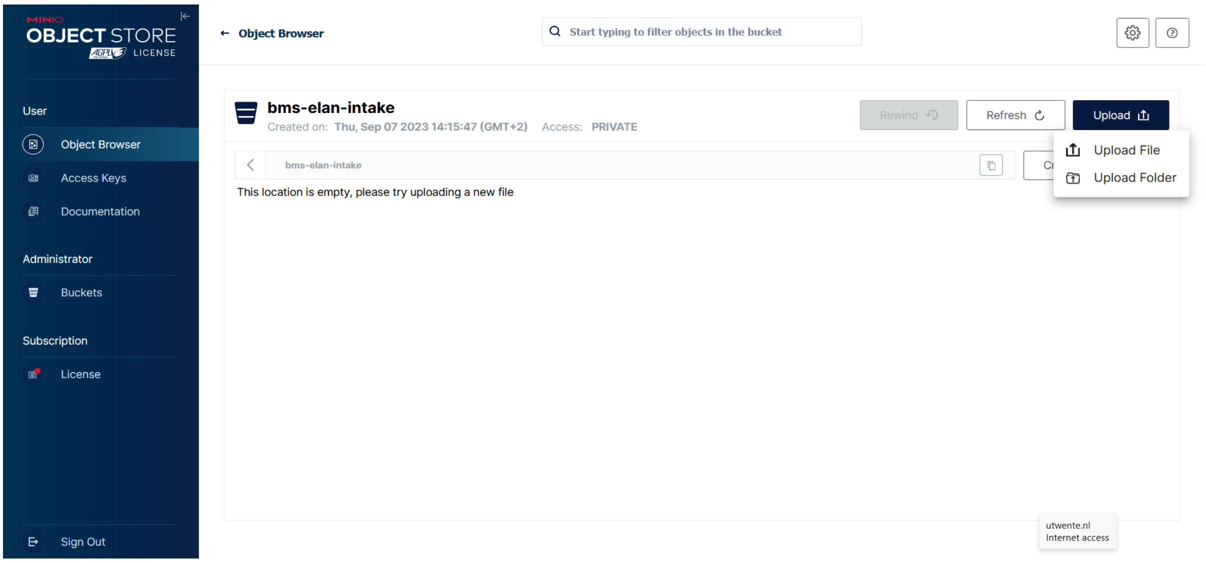
If you encounter any problem with logging in, please contact the data steward in your faculty.

After logging in you will see the two buckets of your organizational unit, like in the screen below.



| Name | Objects | Size | Access |
|---|---|---|---|
| bms-cheps | 3 | 1.2 MiB | R/W |
| bms-cheps-intake | 0 | 0.0 B | R/W |
| bms-cs | 0 | 0.0 B | R/W |
| bms-cs-intake | 0 | 0.0 B | R/W |
| bms-cso | 0 | 0.0 B | R/W |
| bms-cso-intake | 0 | 0.0 B | R/W |
| bms-cstm | 1 | 580.1 MiB | R/W |
| bms-cstm-intake | 0 | 0.0 B | R/W |
| bms-elan | 3 | 4.0 MiB | R/W |
| **bms-elan-intake** | **0** | **0.0 B** | **R/W** |
| bms-etm | 0 | 0.0 B | R/W |

Upload the zip or tar file(s) or folder you created, to the bucket ending in '-intake'. Therefore, press the browse button and you will see the screen below where you can choose to upload a file or folder.



bms-elan-intake
Created on: Thu, Sep 07 2023 14:15:47 (GMT+2) Access: PRIVATE

This location is empty, please try uploading a new file

**Important: Uploading large files can take hours, so make sure your PC or laptop does not go to standby, sleep or hibernate mode during the upload.**

As long as the archive file is in the intake-bucket, you can replace (overwrite) or remove it from the intake-bucket. Be aware that all members of the group have full access which means they can read, write and delete all objects (files) in this bucket, so be careful not to delete files from other members. The files in the final group’s bucket are read-only.

## 3. Adding metadata and README file

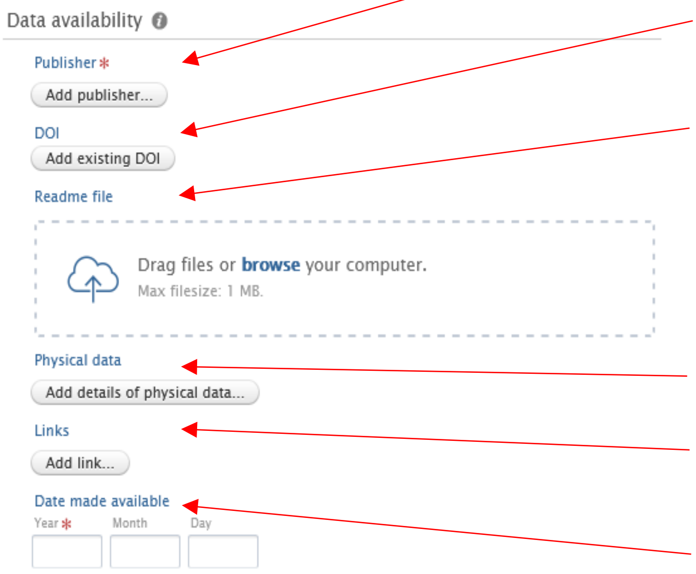
Add the metadata of the uploaded dataset (archive file) and a copy of the reviewed README file in the UT Research Information System (Pure).

**Please note that the README file with the documentation should be available in the archive file in Areda, as well as in Pure.**

For help use the Quick Reference Card (QRC) about registration of datasets in Pure.

Check the following sections with additional information about some specific fields.

## Data availability



Choose ‘Areda (University of Twente)’, so **not** ‘University of Twente’.

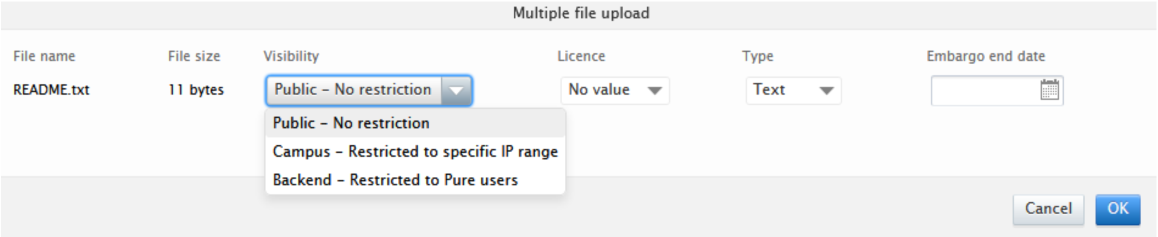
In case of a published dataset, add the DOI issued by the data repository.

Upload a copy of the README file containing the documentation of the dataset. Default the README file is public, but if needed, you can change this when uploading (see below). You can also add a licence.

Skip the physical data field.

Leave the field ‘Links’ empty. This will be filled in by the data steward.

Fill in the date that the dataset has been deposited in a data repository. If it has not been published in a data repository, please fill in the date you uploaded it in Areda.

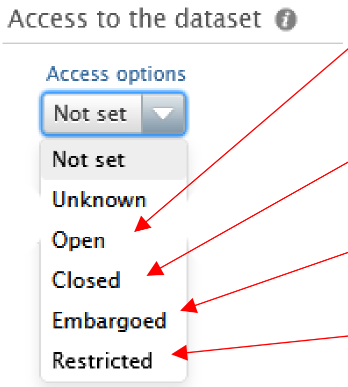


Multiple file upload

| File name | File size | Visibility | Licence | Type | Embargo end date |
|---|---|---|---|---|---|
| README.txt | 11 bytes | Public - No restriction | No value | Text | |

Public - No restriction
Campus - Restricted to specific IP range
Backend - Restricted to Pure users
Cancel
OK

## Access to the dataset



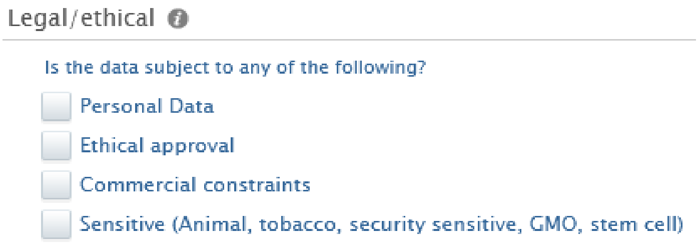
Choose ‘open’ when you also publish or has published the dataset openly accessible in a trusted data repository.

If data in Areda are encrypted, choose ‘closed’ or, in case of temporary non-disclosure, ‘embargoed’.

Involved third parties may demand an embargo period.

If data cannot be published and will only be archived in Areda, where it is only accessible to all group members (default), choose ‘restricted’.

### Legal/ethical



Legal/ethical

Is the data subject to any of the following?

□ Personal Data

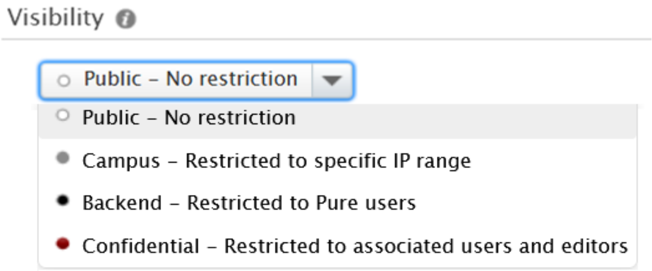
□ Ethical approval

□ Commercial constraints

□ Sensitive (Animal, tobacco, security sensitive, GMO, stem cell)

Access to the data may be restricted, for instance because of privacy. Indicate the reason for this restriction (see also 'Access to the dataset').

### Visibility



Visibility

○ Public – No restriction

○ Public – No restriction

● Campus – Restricted to specific IP range

● Backend – Restricted to Pure users

● Confidential – Restricted to associated users and editors

This sets the visibility of the *description* of a dataset. It is general policy that the description of datasets is publicly visible in the UT Research Information portal.

After completing this metadata form you can press the Save-button for validation. If you want to enter more details about this publication at a later date, select the workflow state 'Entry in progress' in the footer of the window.

## 4. Metadata review and final check

After you have registered a dataset in the previous step and submitted it for validation, the data steward in your faculty will review the metadata and contact you in case any correction or addition is needed. When metadata are complete and correct, the data steward will

- move the archive file from the intake bucket to the final bucket of your research group,
- insert the persistent link of the dataset in the 'Links' field in UT Research Information System and
- send you an email about the result.

Finally, you can decide to do a last check: Download the dataset from Areda using the link in the UT Research Information System and check if the file is equal to the original.

# III. Access to archived datasets

After review by the data steward the bucket with the name of your research group (without the ending "-intake") will contain the archived datasets. Non-encrypted files are accessible (read-only) to all members of the research group. The persistent link in the UT Research Information System contains the path to the dataset in the bucket of the research group. Encrypted files can be accessed by group members having a key (see Encryption).

For non-group members, also external UT, access can be provided by sending a unique, temporary link which you can generate in Areda (Click on the file name and choose 'Share', fill in an expiration time, copy the link to the clipboard and send it). Be aware that access to datasets should be in accordance with data policy of the research group or higher organizational entity, legal requirements (e.g. GDPR) and agreements with third parties involved.

## IV. Searching and overviews of datasets

Searching datasets in Areda can be done in the UT Research Information Portal. To get an overview of the datasets of a research group, use searching or filtering by research unit and datasets.

If you want a report (Excel file, Word file) of all datasets of a research group or faculty, please send an email to ris@utwente.nl.