LONG TERM DATA STORAGE WITH RDMS

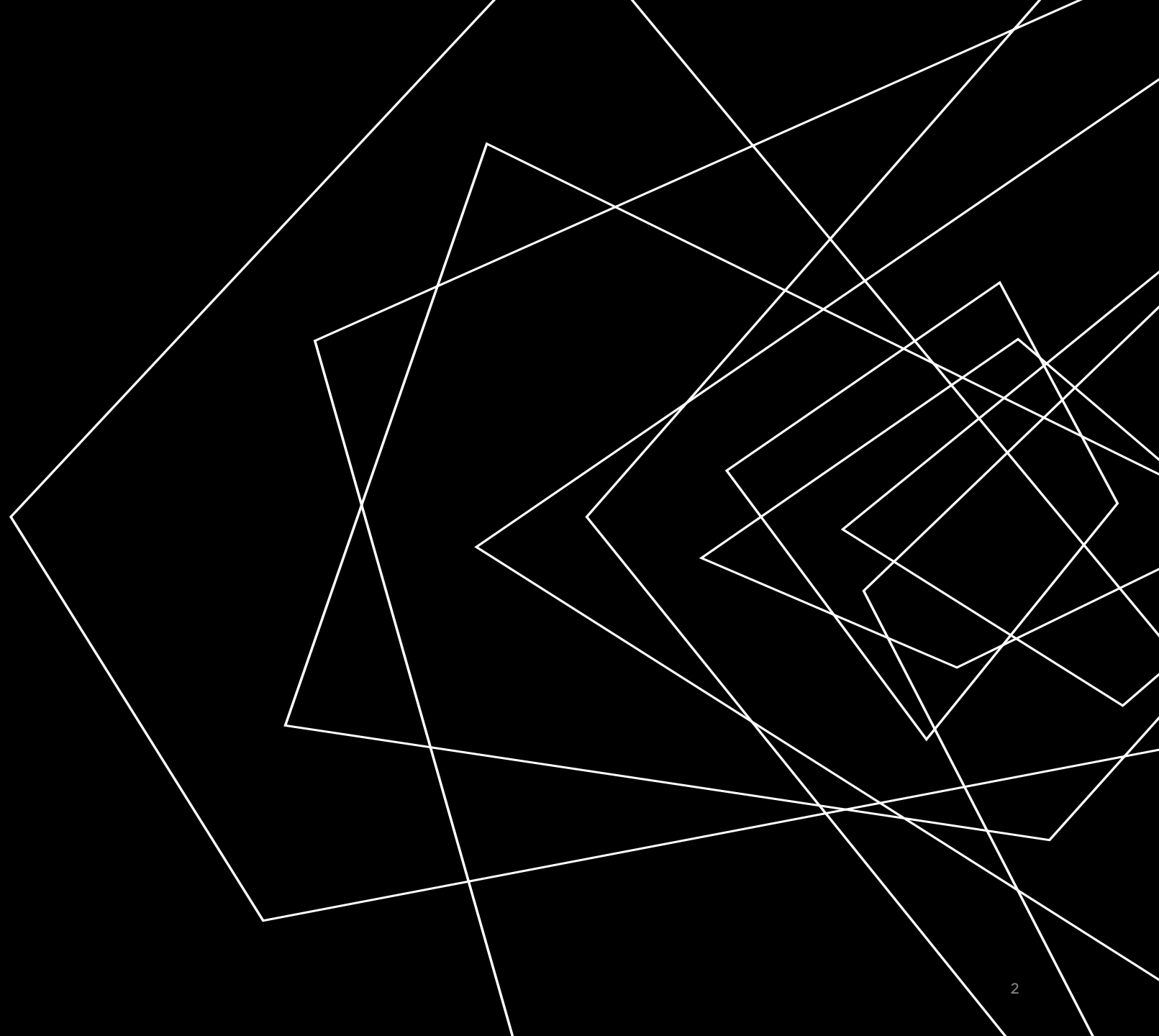
BURCU BEYGU KOOPMANS

09-10-2024

university of groningen

OUTLINE

- Short term vs long term data storage
  - Existing solutions at the university
- What is the RDMS?
- How to access the RDMS
  - Discovering web interface
- How to transfer data to the RDMS
- Different environment in the RDMS
  - How to share data within RDMS?
  - How to work with metadata?
  - How to implement automatized policies?
- Roles and permissions
- How to perform search in the RDMS
- Future



SHORT-TERM VS LONG-TERM DATA STORAGE

Short-Term Data Storage	Long-Term Data Storage
Short-term storage refers to keeping data that is actively being used during the course of a research project. It typically involves temporary or working storage solutions designed to accommodate data in its raw or semi-processed state..	Long-term storage is focused on the preservation and accessibility of research data for future reference, reuse, or replication of research findings. It often involves the use of repositories and archival services to ensure the data remains usable and protected over extended periods.
Purpose: To support active research activities such as data analysis, simulations, or experimental runs	**Purpose:** To ensure that research data is accessible for verification, replication of results, or future studies.

SHORT-TERM VS LONG-TERM DATA STORAGE

EXISTING SOLUTIONS

Short-Term Data Storage	Data Sharing Options for Short Term Data Storage
• UWP: X & Y drives • LWP Home & Project drives • Hábrók /scratch & /projects directories • VRW Home & Group drives	• Unishare • SURF Drive
Purpose: These are environments that allow to perform computation and fast access to data.	**Purpose:** These environments facilitate data sharing between non-RUG users, don't offer computation functionality but data is stored for short term.
For further details refer to the IT Solution page of the DCC	

LONG-TERM DATA STORAGE

EXISTING SOLUTION

Research Data Management System (RDMS)

WHAT IS RDMS?

Long-term data storage where:

- You can implement policy-based automation..
- You can manage metadata templates & extract metadata.
- You can search for your data, save your search and export it.
- You can manage group and personal environments.
- You can assign users different roles and permissions.
- You can access and share data within RUG anytime.
- You can see activity auditing (in progress).
- You can publish your metadata in DataverseNL (in progress).

WHAT IS RDMS?

- RDMS user manual or RDMS wiki:

https://wiki.hpc.rug.nl/rdms/rdmsproject/start

(You can just Google as RDMS wiki)

- Contact us:

rdms-support@rug.nl

BEST PRACTICE

Step 1

- I **don't need to frequently** access, process or analyze data. I published an article/poster/thesis/book/etc.
- I don't need to access the data frequently, maybe in the future I may reuse the data or share it for research purposes.
- I need a storage where even if I leave my supervisor/group leader/ data manager can always find & access the data.
- I need a storage to archive our group/lab/personal data.

Step 2

- Decide on **which data** should be archived.
- Estimate the **amount of data** you need to archive.
- Decide on how many Team Drive you need.
- Decide in who should have which **roles** and permissions and for how long.
- Decide on whom shall be the **main contact point**?
- Ask for help from the RDMS team.

WHAT IS RDMS?

Data management platform that uses iRODS to manage and store data in the backend.

- iRODS (Integrated Rule-Oriented Data System) is an open-source data management software platform that enables organizations to manage, share, and preserve large amounts of distributed data across different storage systems.
- It provides a virtual file system that abstracts data from the underlying storage, allowing for better control, automation, and organization of data at scale.
- iRODS is widely used in research, industry, and government sectors to ensure compliance with data management policies, automate workflows, and manage data lifecycles efficiently.
- Dutch universities and SURF often use iRODS based solutions as their data storage and archiving solution. They develop systems such as RDMS as well based on iRODS e.g: Yoda
- University of Groningen is part of the international iRODS consortium together with other four Dutch organizations; SURF, Utrecht University and Maastricht University.

WHAT IS RDMS?

Since RDMS uses iRODS there some new concepts:

Zone: Zone is a virtual environment designed to manage data storage, users, and resources**.** Each Zone has a unique name. Currently there are two RDMS zones: RUG_DEFAULT and SD_Zone. Advantage of having multiple zones are to assignee specific environments to institutes, big groups etc to optimize the performance.

Data object: Data objects in iRODS represent files, but with enhanced flexibility through metadata, replication, and access control. The virtualized storage allows users to work with their data regardless of the underlying physical location.

Collection: All Data Objects stored in an iRODS system are stored in some Collection, which is a logical name for that set of Data Objects. A Collection can have sub-collections, and hence provides a hierarchical structure. An iRODS Collection is like a directory in a Unix file system (or Folder in Windows), but is not limited to a single device or partition.

WHAT IS RDMS?

Data storage (as of 2024)

- Data is replicated and stored in two different physical location.
  - Kept for 10 years as the university data policy requires.

- Sensitive data must be stored in Sensitive Data Zone.

- Replication is copying live data from 1 site to another, against:
  - system-failure
  - datacenter-failure (fire, etc.)

- Data is backed-up
  - Retention period: 30 days
  - Active versions: 2
  - Deleted versions: 2

HOW TO ACCESS THE RDMS?

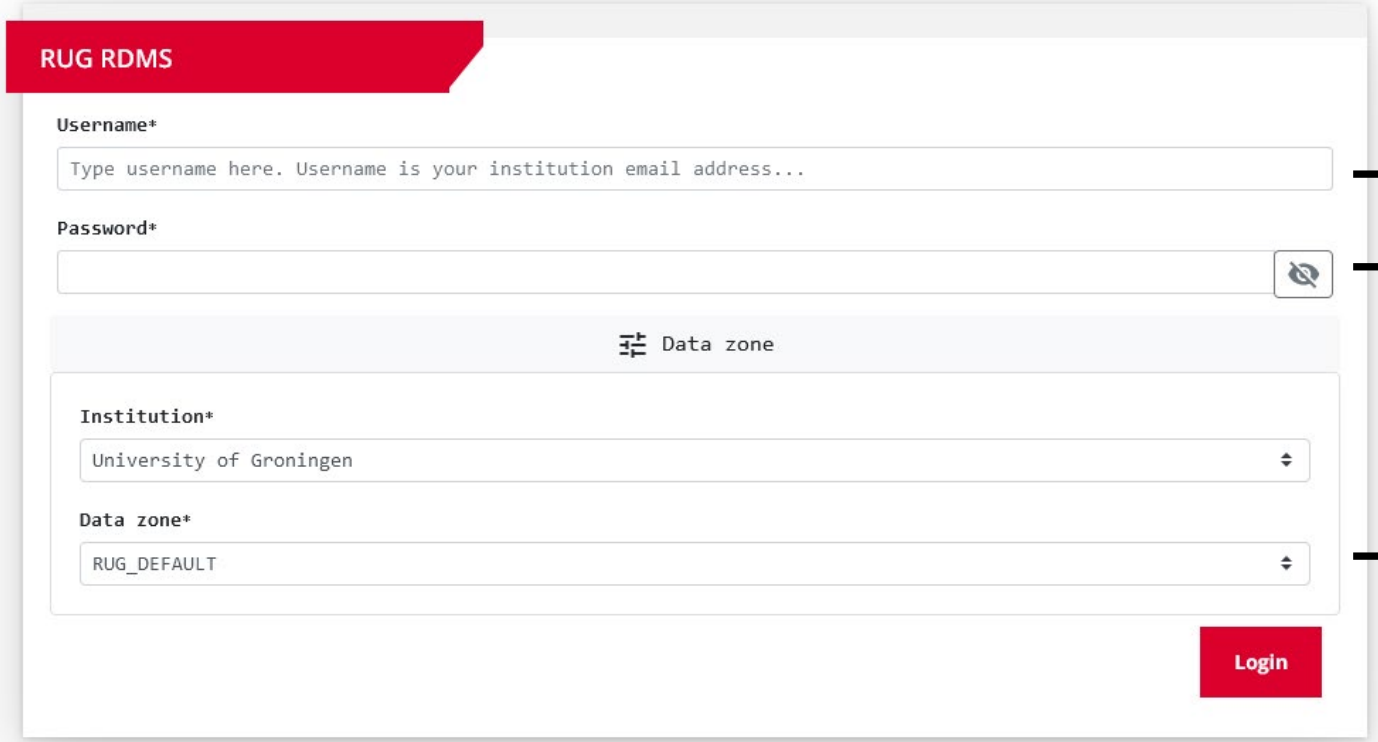
1) Web Interface
2) iCommands

Note: You need VPN if you are outside of the RUG network

HOW TO ACCESS THE RDMS?

WEB INTERFACE

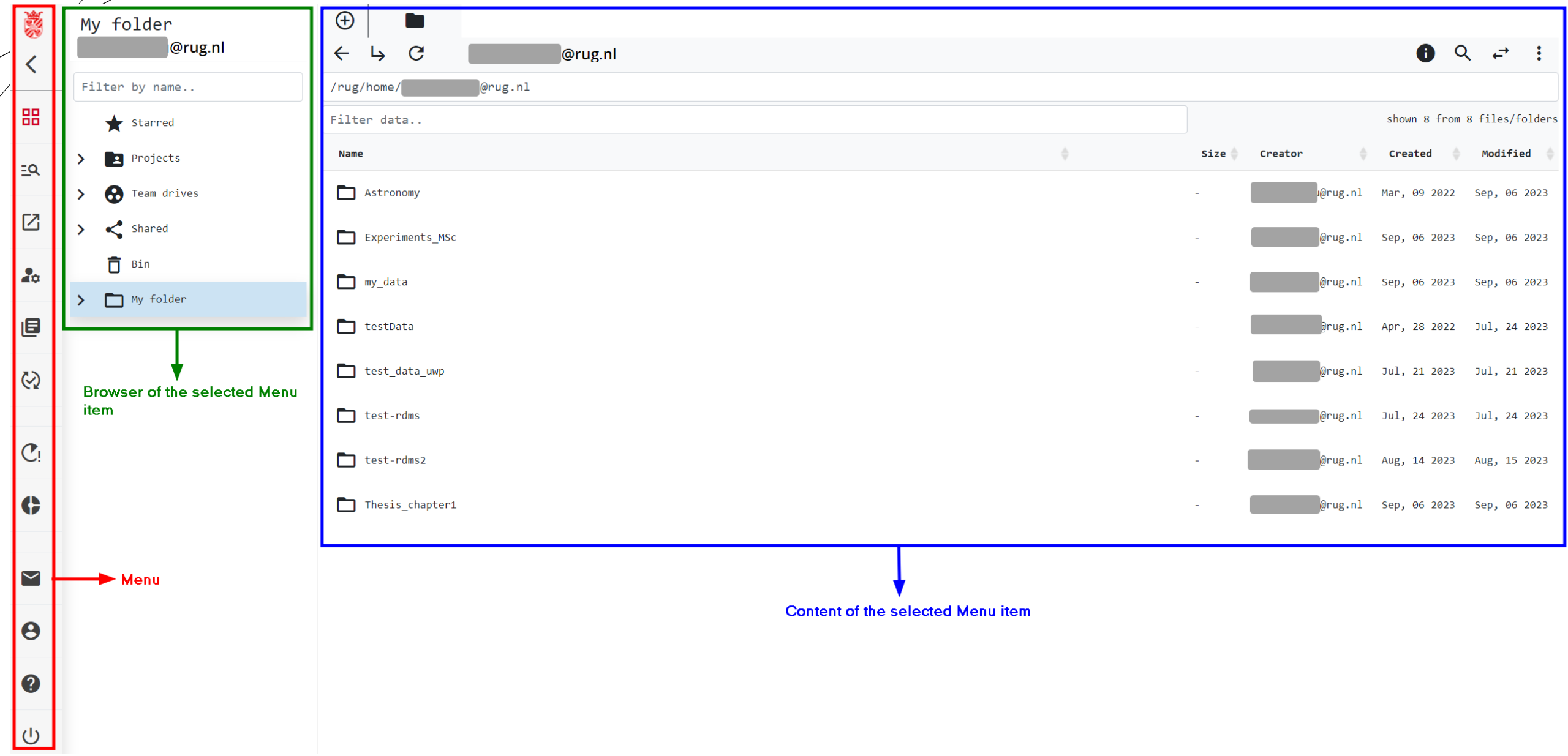
Log in via: https://research.web.rug.nl/rdmswebapp

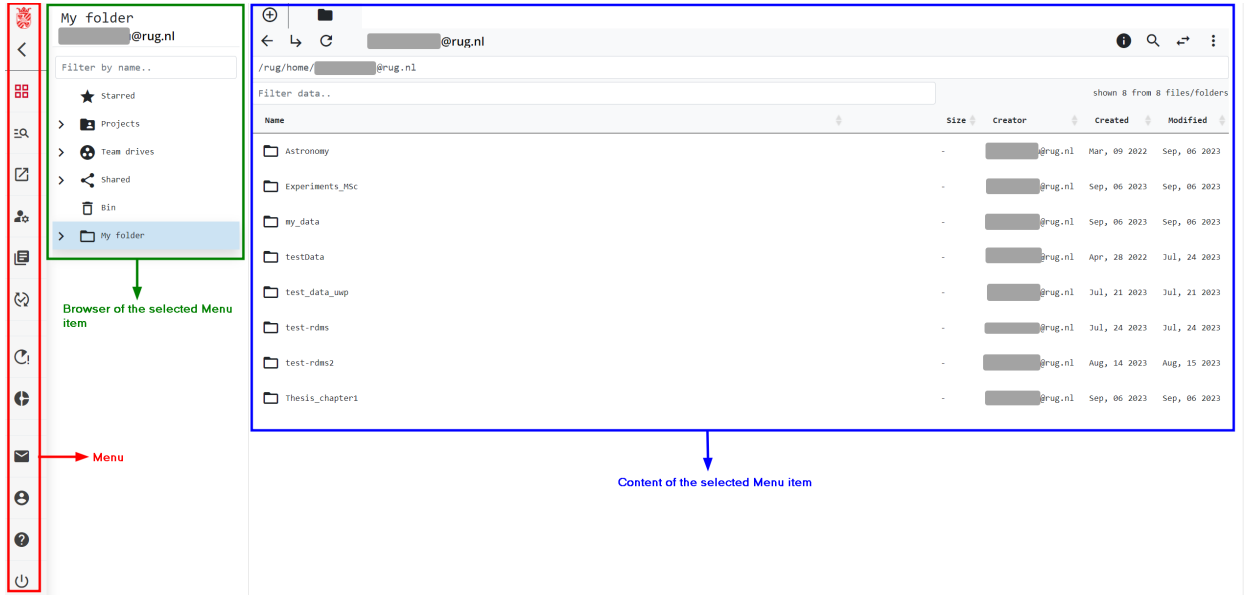


→ RUG email address

→ Password

→ Zone name


My folder
[redacted]@rug.nl
Filter by name..
Starred
Projects
Team drives
Shared
Bin
My folder
Browser of the selected Menu item
Menu
[redacted]@rug.nl
/rug/home/[redacted]@rug.nl
Filter data..
shown 8 from 8 files/folders
Name
Size
Creator
Created
Modified
Astronomy - [redacted]@rug.nl Mar, 09 2022 Sep, 06 2023
Experiments_MSc - [redacted]@rug.nl Sep, 06 2023 Sep, 06 2023
my_data - [redacted]@rug.nl Sep, 06 2023 Sep, 06 2023
testData - [redacted]@rug.nl Apr, 28 2022 Jul, 24 2023
test_data_uwp - [redacted]@rug.nl Jul, 21 2023 Jul, 21 2023
test-rdms - [redacted]@rug.nl Jul, 24 2023 Jul, 24 2023
test-rdms2 - [redacted]@rug.nl Aug, 14 2023 Aug, 15 2023
Thesis_chapter1 - [redacted]@rug.nl Sep, 06 2023 Sep, 06 2023
Content of the selected Menu item

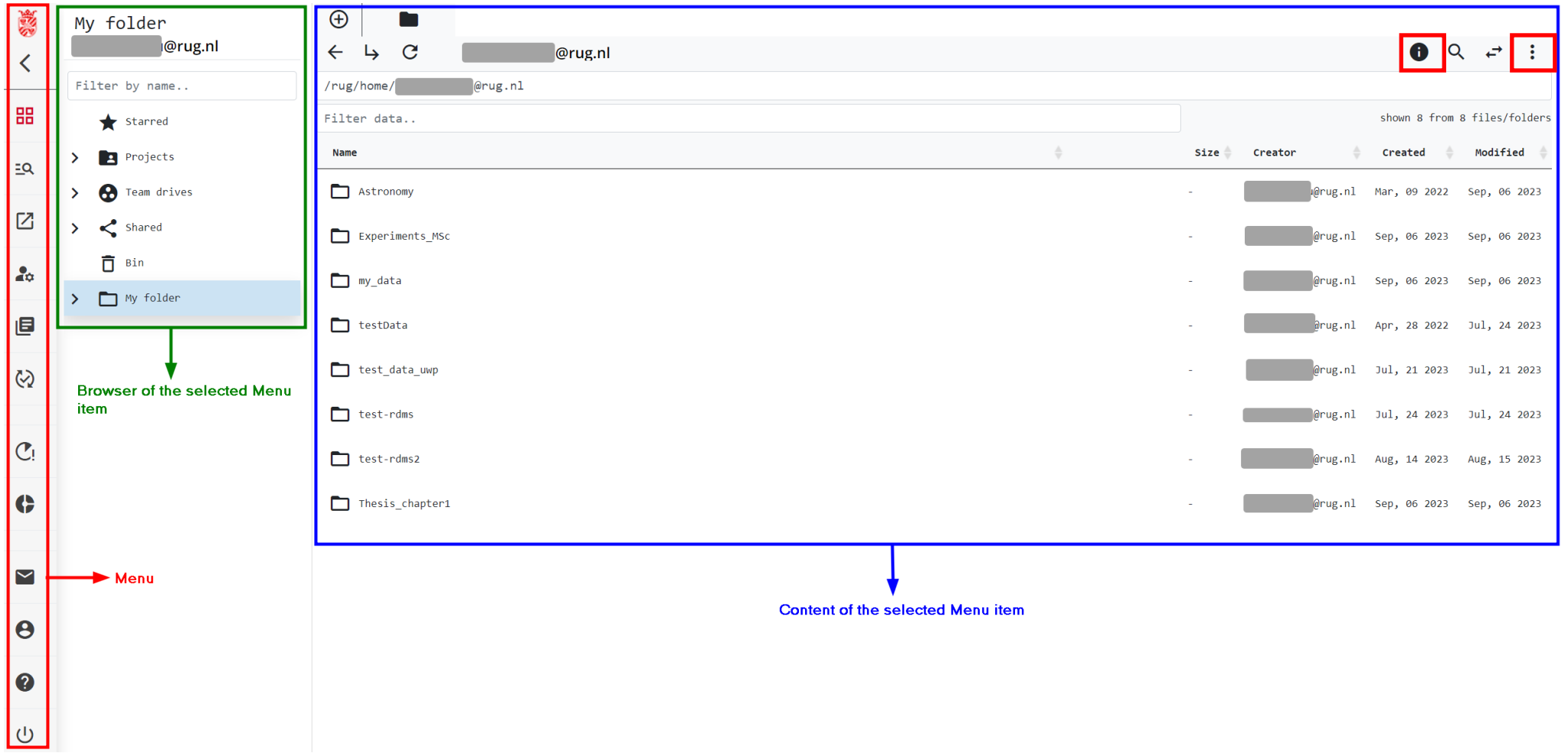


Via the web interface:

- Managing user roles and permissions
- Sharing data within the RUG
- Policy implementation
- Create metadata templates, append metadata to files and folders and extract metadata
- Creating Groups, Team Drives and Projects
- Managing Archiving and Publishing (not yet enabled) workflows
- Uploading data
- .....

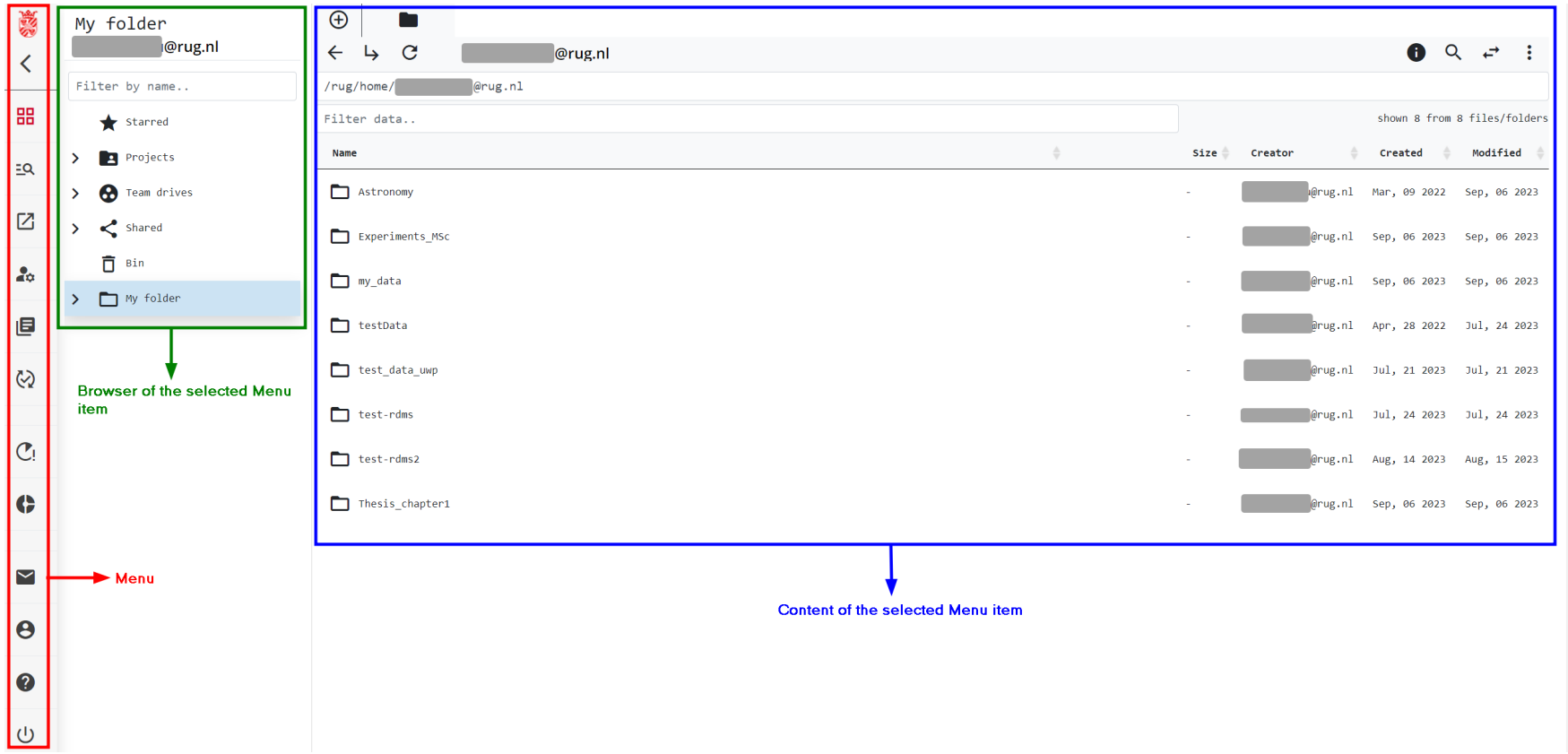
Two functionality buttons that you will see in every level. They will display actions such as:

- Add user, group, move to bin, create folder, rename, upload/download file, add, extract metadata, etc.



Useful hotkeys

dbn: Create a new collection (folder)
dbu: Upload a new data object (file)
dbm: Move the selected objects or collections
dbl: Link a collection
ctrl + h: Display the help window



HOW TO ACCESS THE RDMS?

ICOMMANDS

- You need a shell/terminal where you can install iCommands
- iCommands are Unix utilities that give users a command-line interface to operate on data in the iRODS system. iCommands provide the most comprehensive set of client-side standard iRODS manipulation functions.
- iCommands are already installed in the **LWP** and **Hábrók**. From the LWP it is also **possible** to access X & Y drive. This way data transfer or other data management tasks can be performed on the data on X & Y drive.

RDMS Command Line Workshop October 23rd, 13:00

HOW TO ACCESS THE RDMS?

ICOMMANDS

Step 1:

- Create a ~\.irods directory in your home directory:

```
$ mkdir ~/.irods
```

- Create an irods_environment.json file in that directory with a text editor of your choice.

```
$ cd ~/.irods
$ nano irods_environment.json
```

HOW TO ACCESS THE RDMS?

ICOMMANDS

Step 2:

- Configure the irods_environment.json file $ cd ~/.irods

```
{
    "irods_authentication_scheme": "pam_password",
    "irods_client_server_negotiation": "request_server_negotiation",
    "irods_client_server_policy": "CS_NEG_REQUIRE",
    "irods_default_resource": "rootResc",
    "irods_encryption_algorithm": "AES-256-CBC",
    "irods_encryption_key_size": 32,
    "irods_encryption_num_hash_rounds": 16,
    "irods_encryption_salt_size": 8,
    "irods_host": "store.data.rug.nl",
    "irods_port": 1247,
    "irods_ssl_verify_server": "cert",
    "irods_user_name": "<emailadress>",
    "irods_zone_name": "rug"
}
```

→ Your email address

HOW TO ACCESS THE RDMS?

ICOMMANDS

Step 3:

- Each time you want to interect with RDMS, you need to initiate iCommands:

$ iinit
Enter your current PAM password: Type in your password

- You are ready to use iCommands

$ ipwd
$ ils
$ itree
$ icd <new_path>

HOW TO ACCESS THE RDMS?

ICOMMANDS

Accssing X or Y drive within LWP via iCommands

Even though the UWP is a Windows OS, you can still use the iCommands to benefit from its various functionalities, speed and efficiency. The Y: drive is available under /media/ydrive

$ iinit
$ cd /media/ydrive/<your folder>

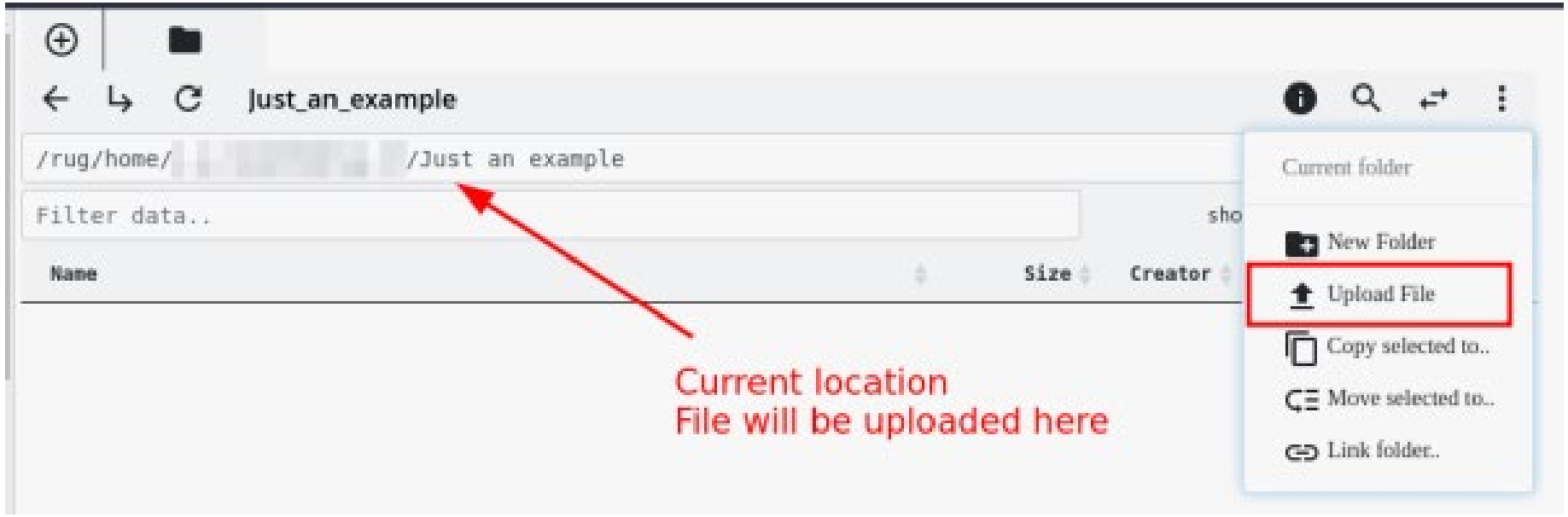
- You are ready to use iCommands

HOW TO TRANSFER DATA TO THE RDMS?

- Web Interface
- iCommands
- File transfer applications
  - CyberDuck
  - WinSCP

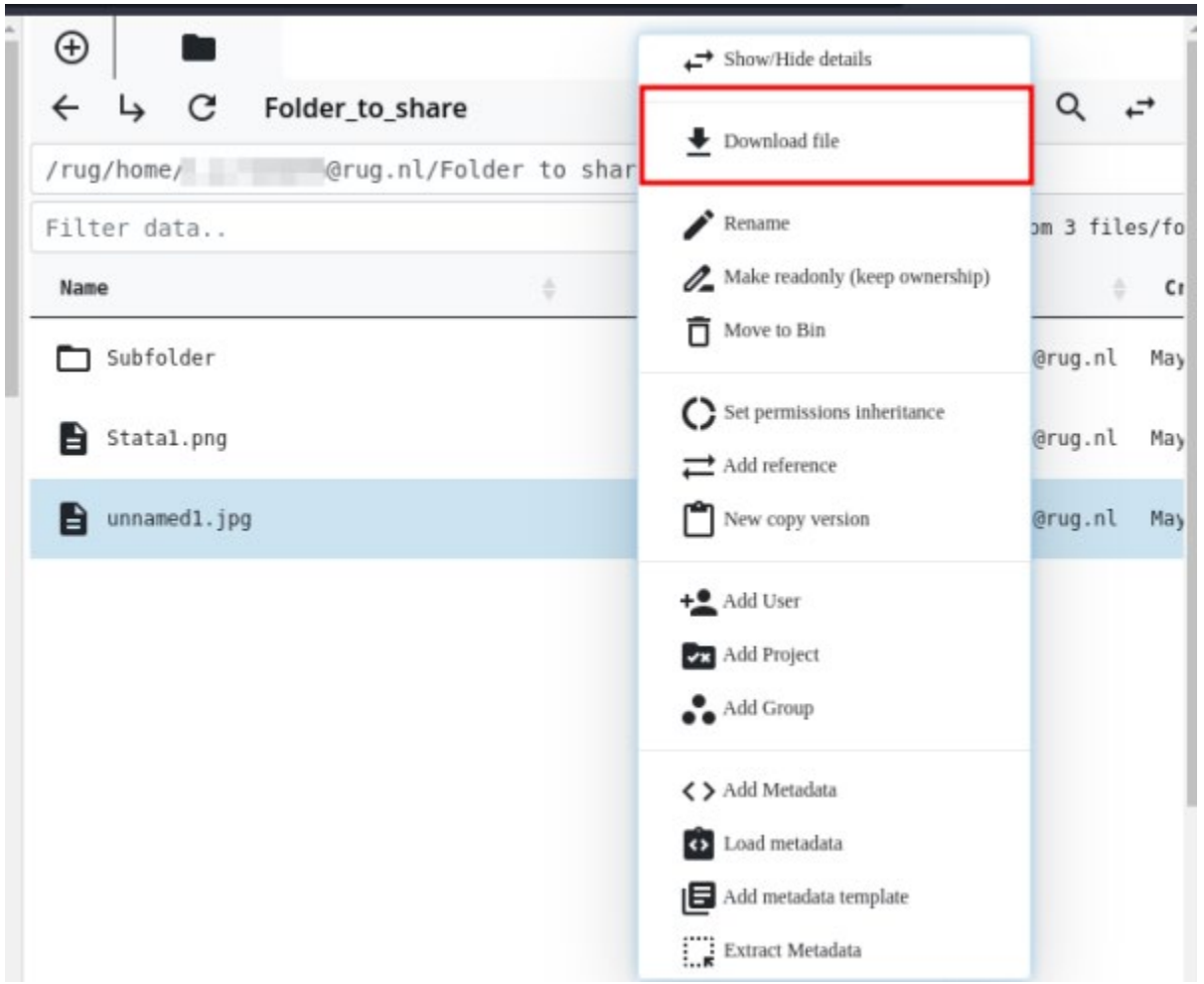
HOW TO TRANSFER DATA TO THE RDMS?

- Web Interface
  - Only individual files
  - Data size < 50 MB



HOW TO TRANSFER DATA TO THE RDMS?

- Web Interface
  - Only individual files
  - Data size < 50 MB



HOW TO TRANSFER DATA TO THE RDMS?

- iCommands

```
# Transfer a file to the RDMS
$ iput test.txt /rug/home/rdms-testers@rug.nl/some_example_folder
$ ils /rug/home/rdms-testers@rug.nl/some_example_folder
/rug/home/rdms-testers@rug.nl/some_example_folder:
  test.txt
```

If you want to transfer a whole folder structure, the additional -r flag has to be used.

```
$ iput -r Testdir_L1
Running recursive pre-scan... pre-scan complete... transferring data...
```

If you want to see the transfer rate, you can use the additional -P flag.

```
$ iput -P 1GB_Testfile_1
0/1 -  0.00% of files done   0.000/953.674 MB -  0.00% of file sizes done
```

HOW TO TRANSFER DATA TO THE RDMS?

- Download data from RDMS:

$ iget [flags] <source> <destination>

- Synchronize data:

```
# Synchronize the content of local folder to a folder on the RDMS:
$ irsync -r <source> i:<dest>

Synchronize the content of remote folder on the RDMS with a local folder:
$ irsync -r i:<source> <dest>

Synchronize the content of two remote RDMS folders:
$ irsync -r i:<source> i:<dest>
```

HOW TO TRANSFER DATA TO THE RDMS?

- File transfer applications
  - CyberDuck: You can use two protocols WebDAV or iRODS native

- What is a protocol?

Data transfer protocols refer to the standardized methods used to securely move data between various data sources and applications.

Examples:

- the HTTP (HyperText Transfer Protocol) is used for transferring hypermedia documents on the web,
- FTP (File Transfer Protocol) is used for transferring files from one host to another,
- WebDAV is an extension of HTTP, used for collaborative file editing and management over the web .....

HOW TO TRANSFER DATA TO THE RDMS?

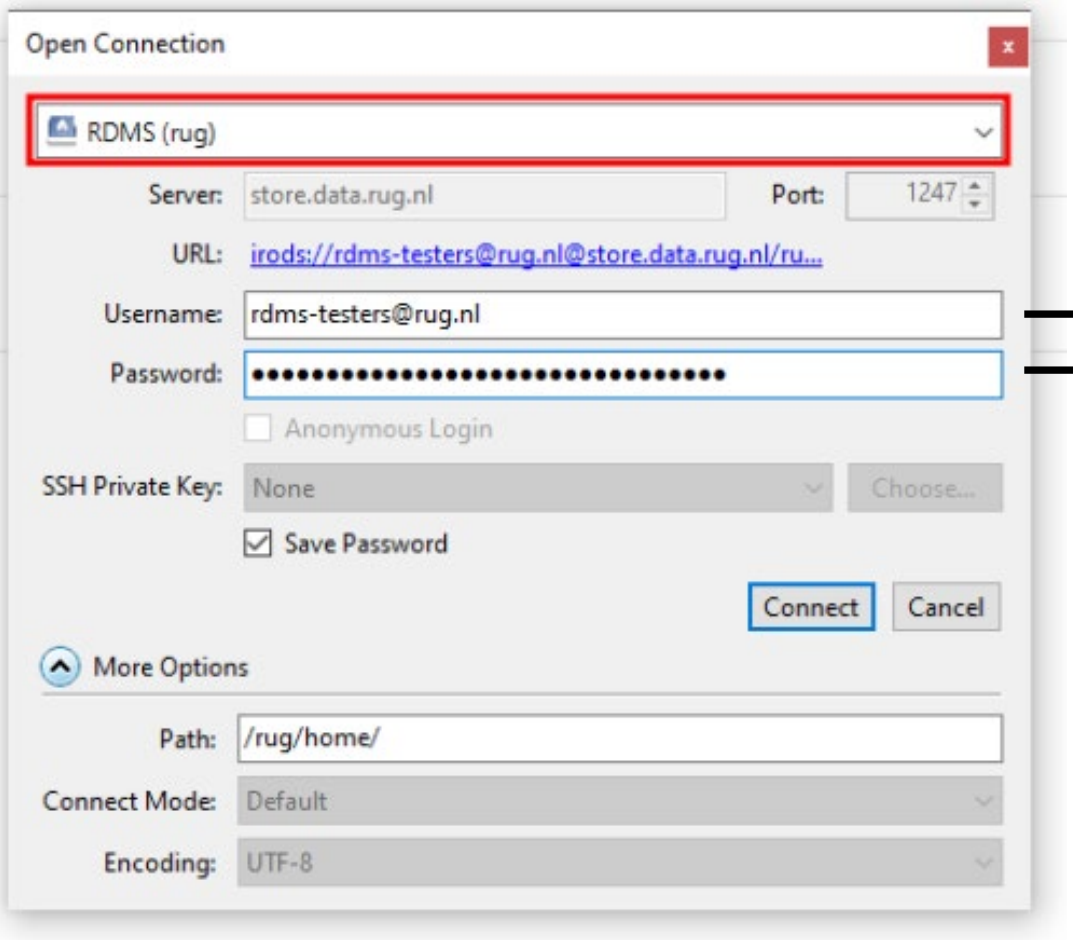
- File transfer applications
  - Cyberduck: You can use two protocols WebDAV or iRODS native

Step 1: Go to RDMS wiki and Cyberduck chapter

Click on and download the profile. Add the extension *.cyberduckprofile*

HOW TO TRANSFER DATA TO THE RDMS?

Step 2: Go to the location you downloaded the file and double-click. Cyberduck should be automatically loaded.



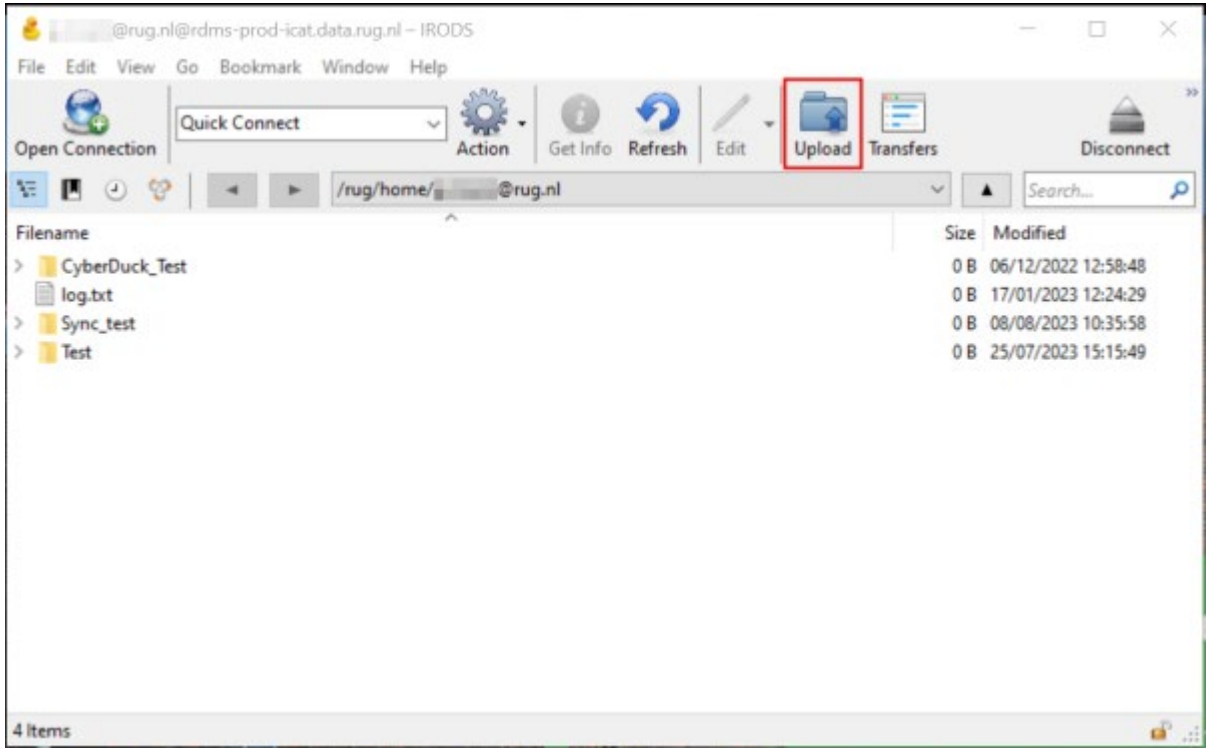
Everything should be filled exactly the same except your own email address & password.

→ Your email address

→ Your password

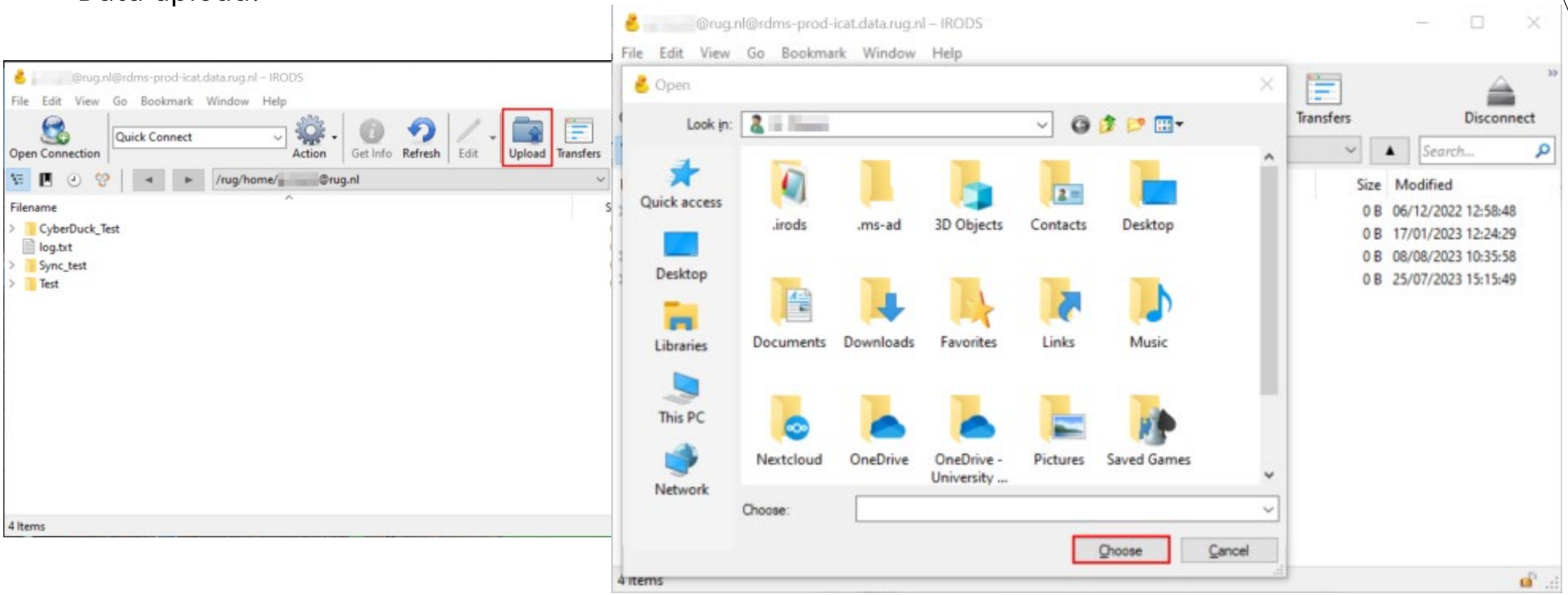
HOW TO TRANSFER DATA TO THE RDMS?

Data upload:



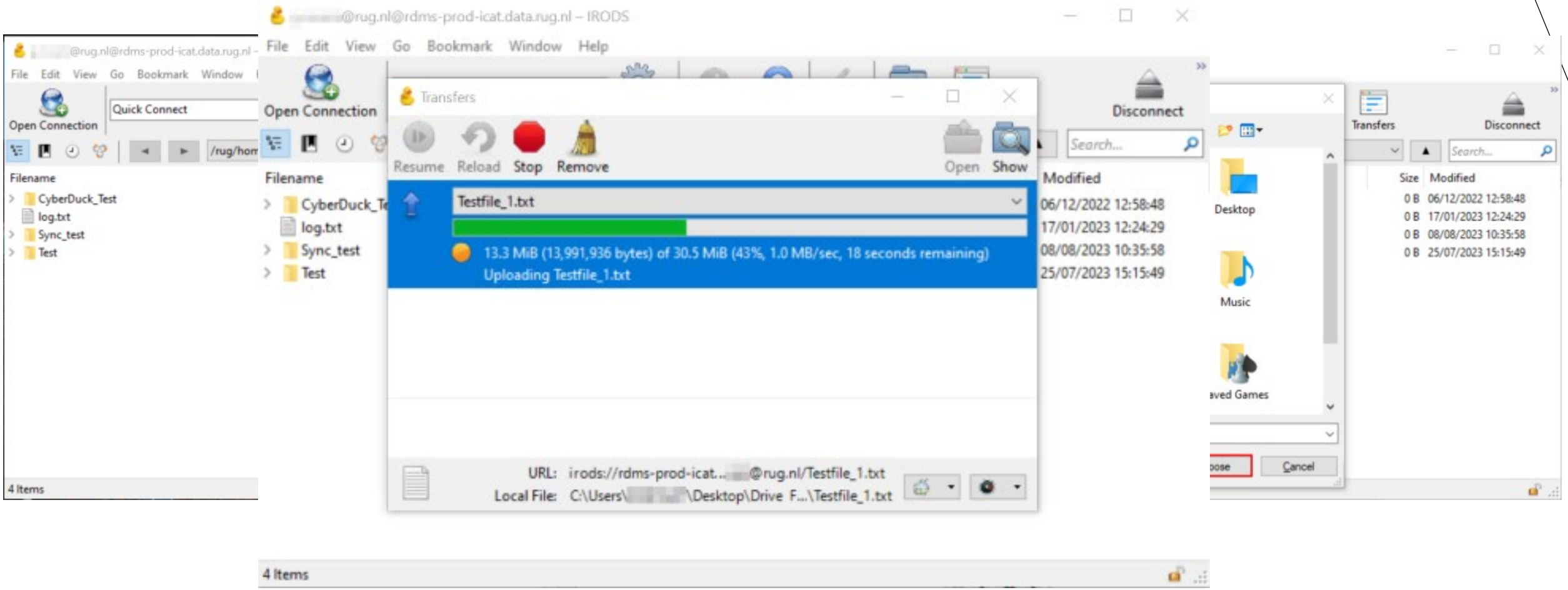
HOW TO TRANSFER DATA TO THE RDMS?

Data upload:



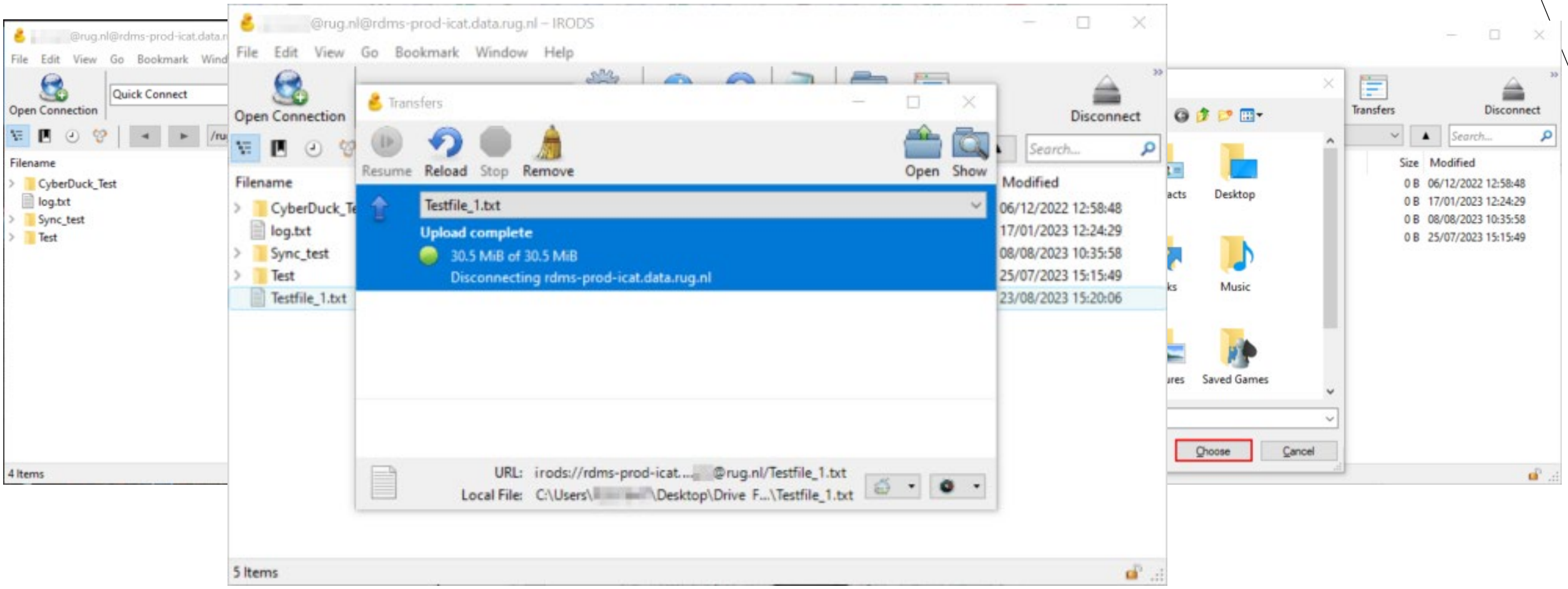
HOW TO TRANSFER DATA TO THE RDMS?

Data upload:



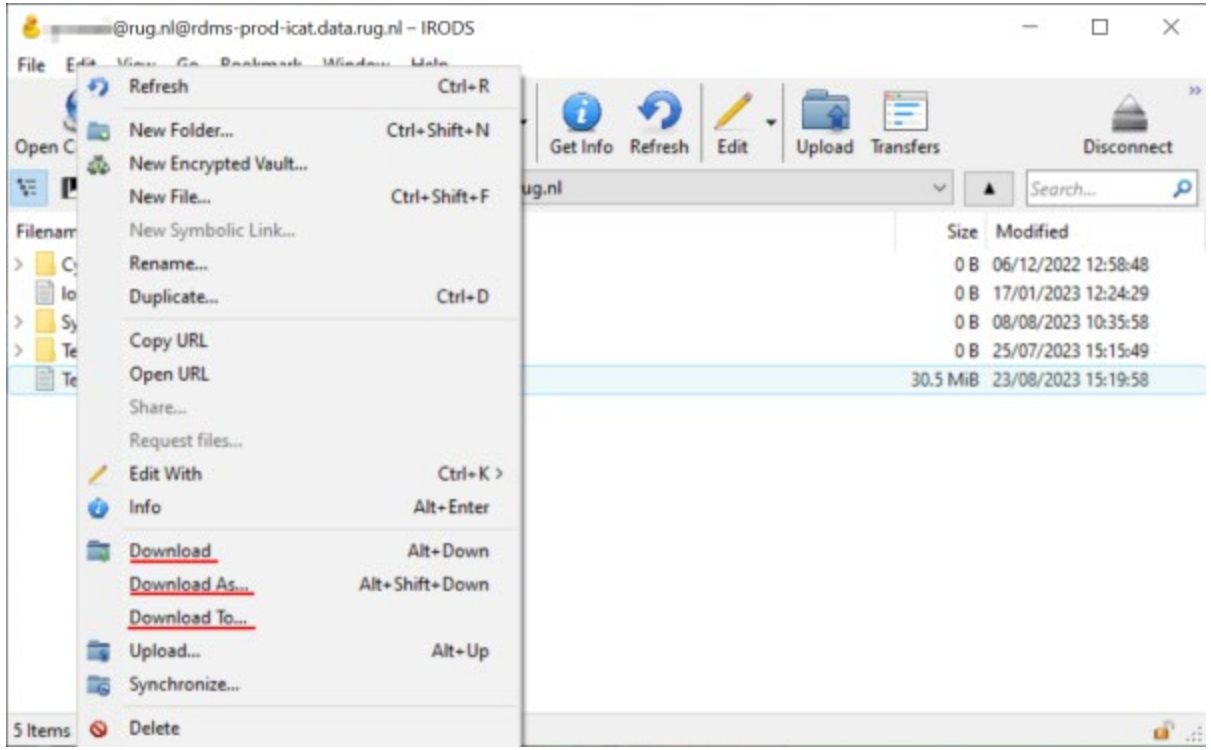
HOW TO TRANSFER DATA TO THE RDMS?

Data upload:



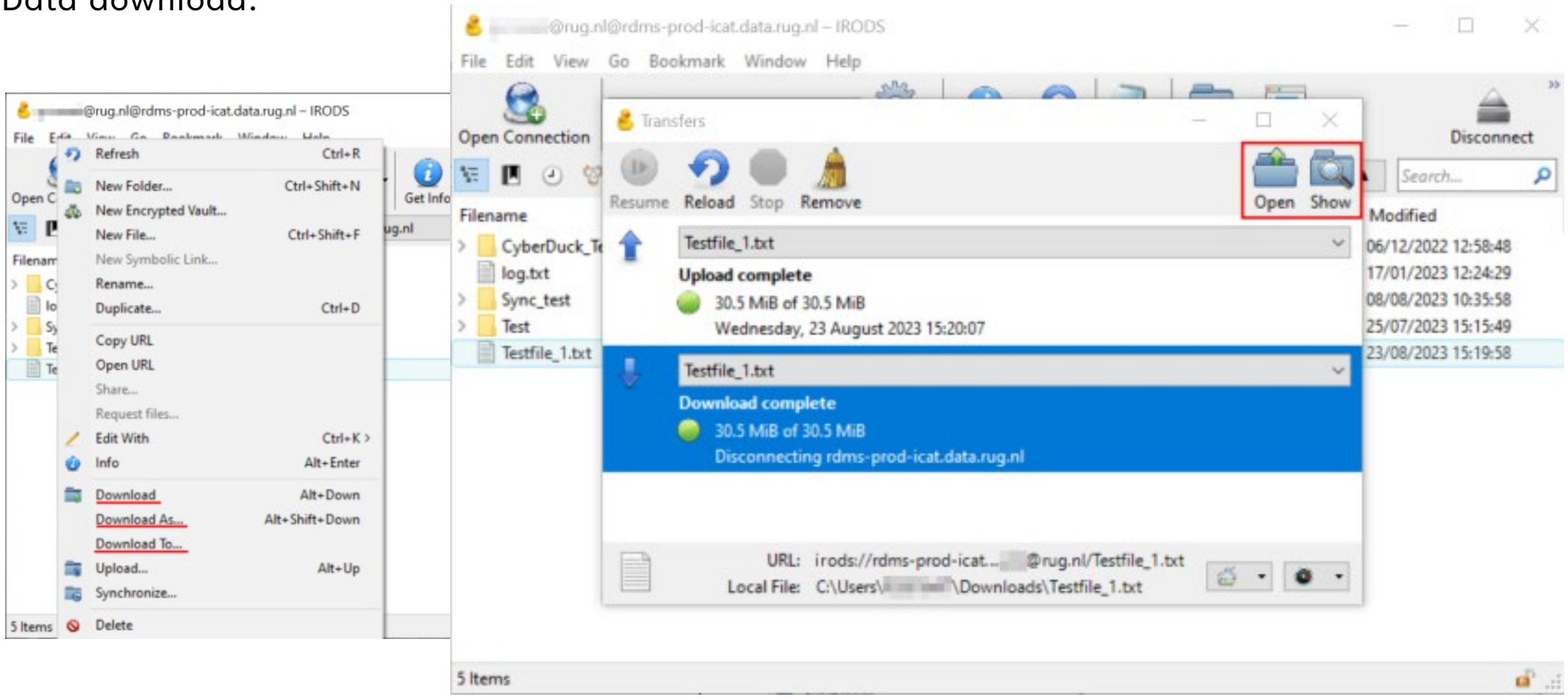
HOW TO TRANSFER DATA TO THE RDMS?

Data download:



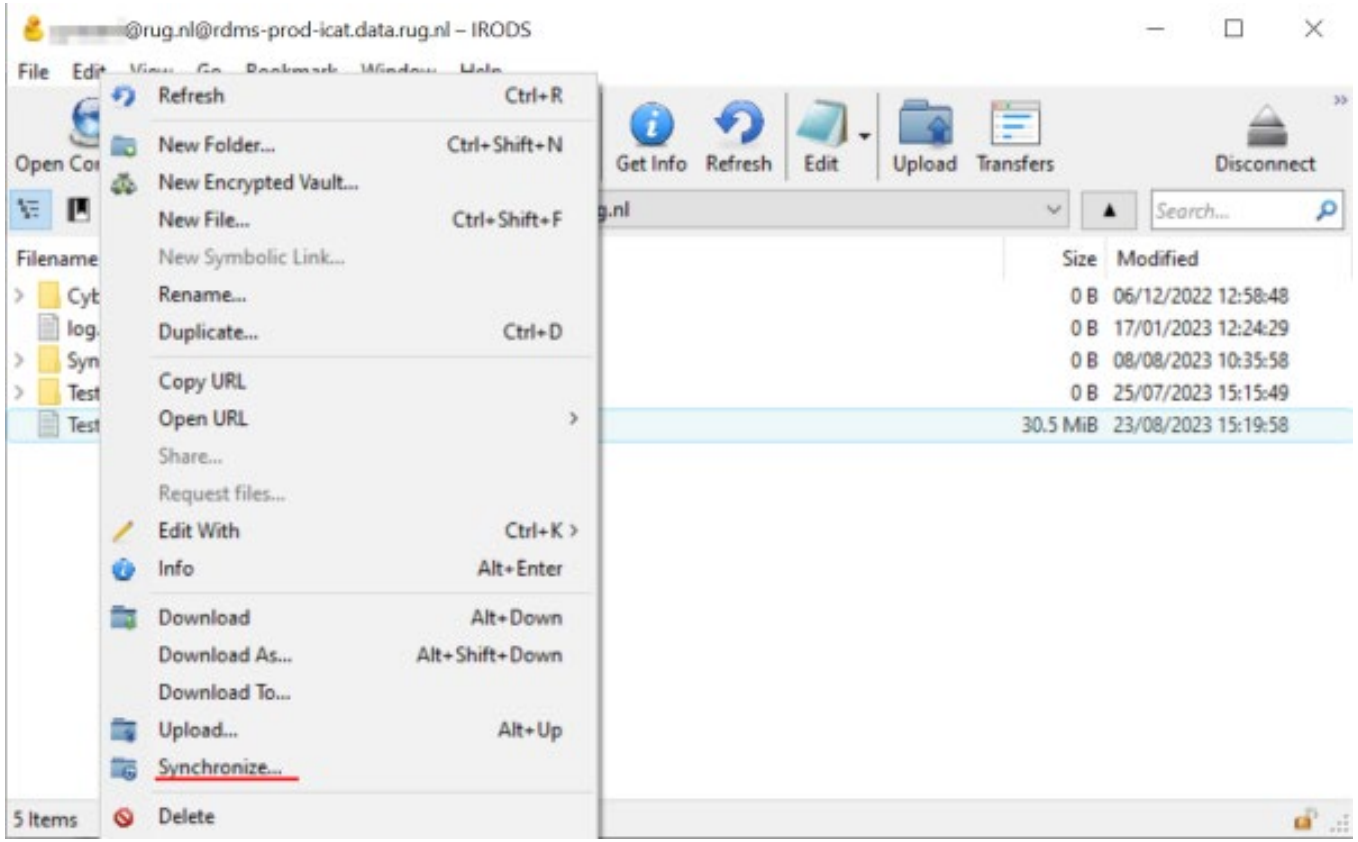
HOW TO TRANSFER DATA TO THE RDMS?

Data download:



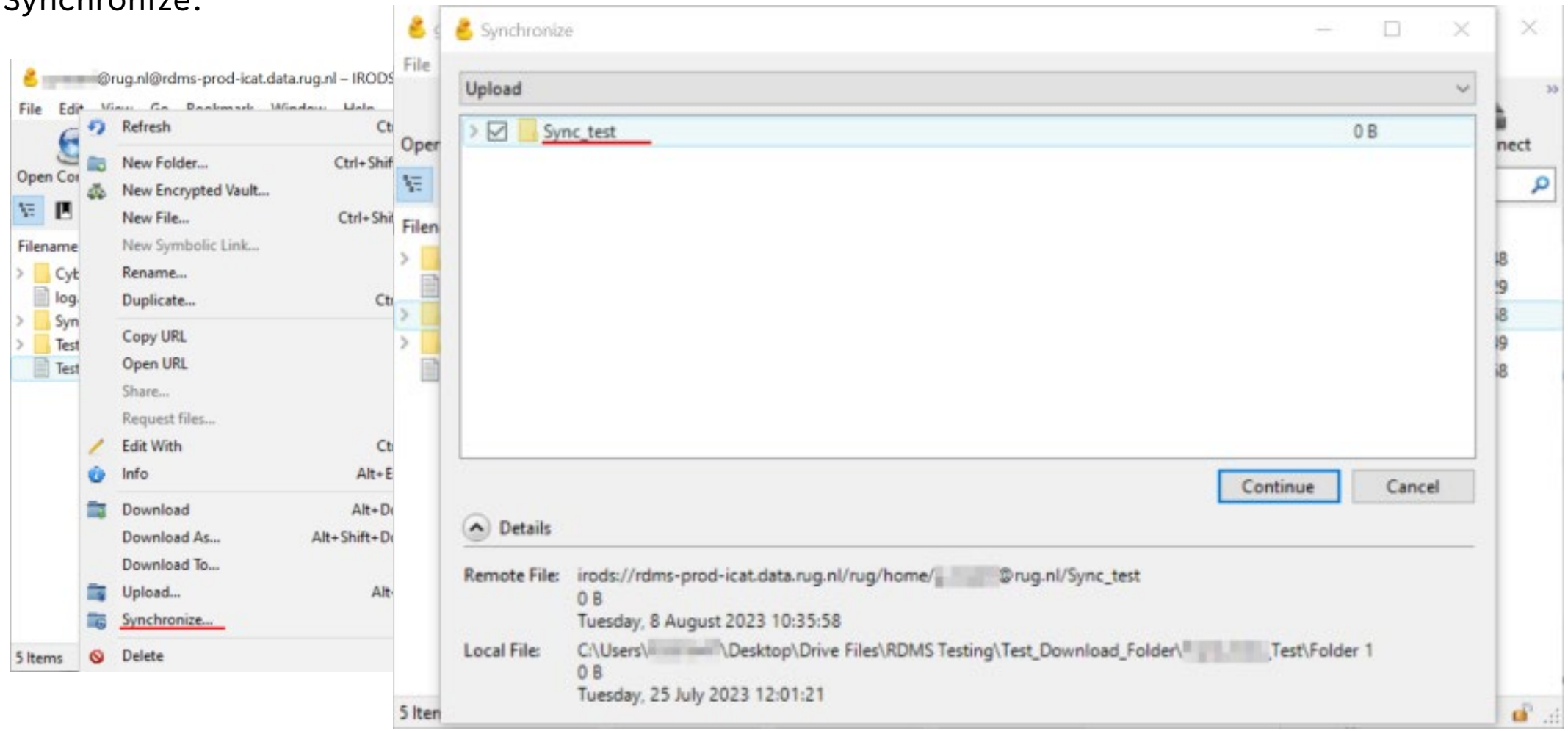
HOW TO TRANSFER DATA TO THE RDMS?

Synchronize:



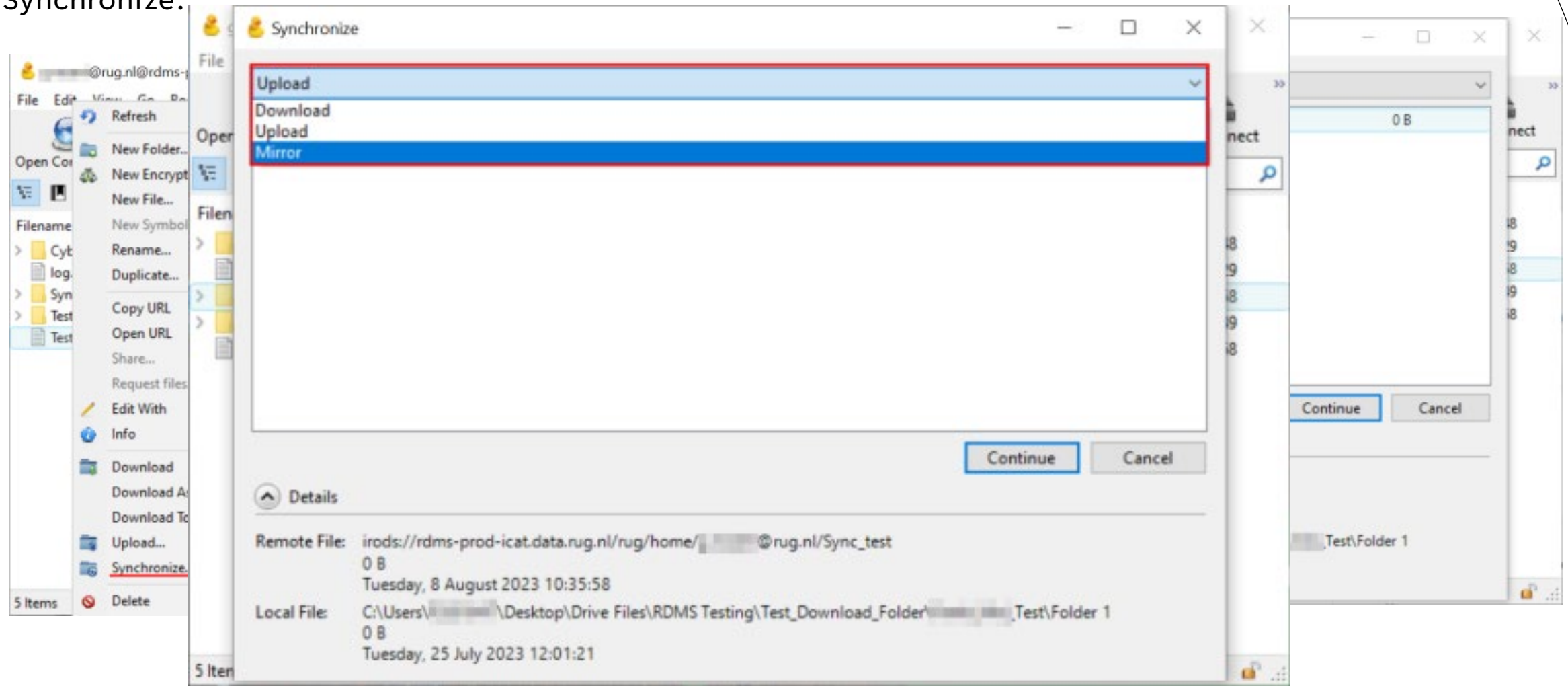
HOW TO TRANSFER DATA TO THE RDMS?

Synchronize:



HOW TO TRANSFER DATA TO THE RDMS?

Synchronize:



BEST PRACTICE

- If you have large amount data, such as many folders and sub folders with each of them hundreds to thousands of files, always zip or tar or tar.gz or tar.bz the folders before you transfer them to the RDMS.

- Do not use special characters such as `~!@#$%^&*()?.,” In your file of folder names including, characters Such as Chinese, Greek etc.

- Avoid giving space in between words when naming folders.

- Check the Best Practices and Cyberduck/iCommands chapters for updated tips and best practices.

DIFFERENT ENVIRONMENT IN THE RDMS

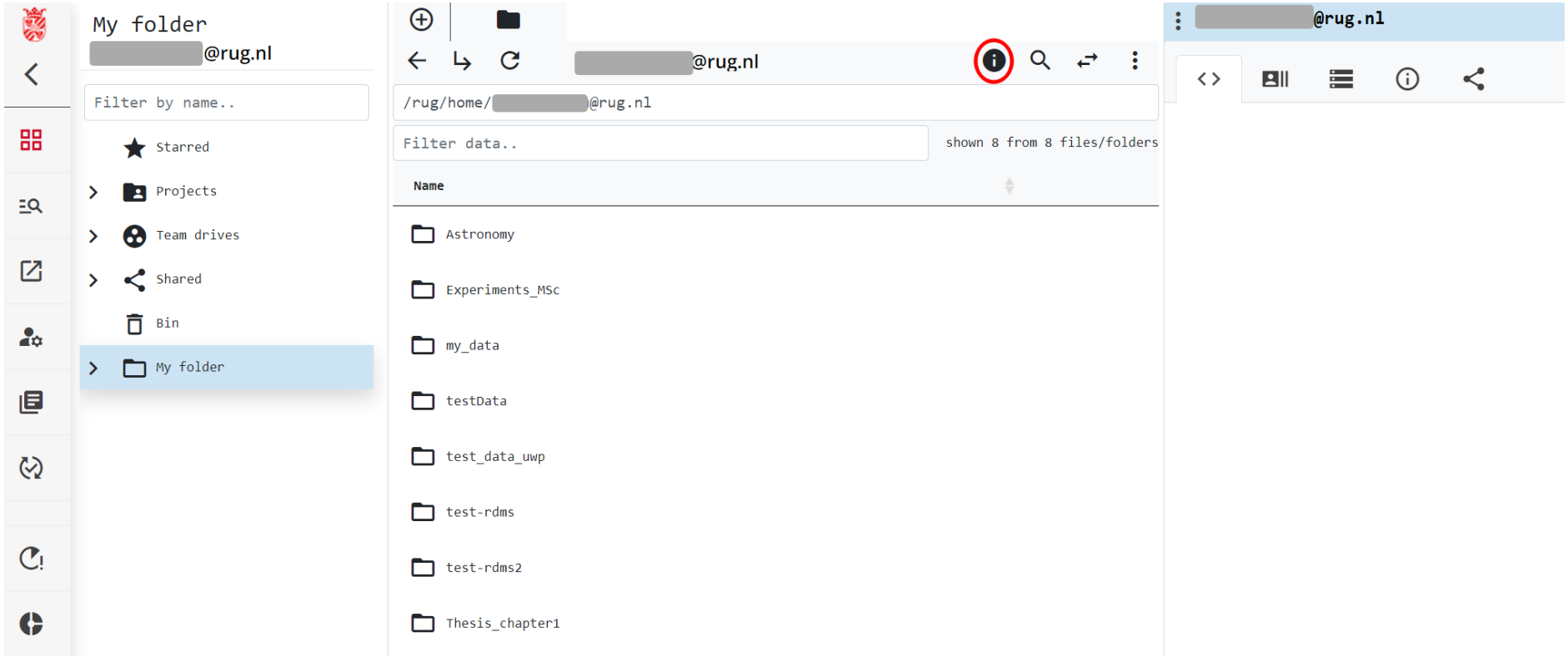
1) **Home Drive:** Personal environment which is by default only accessible to the user who owns the account. You can perform various data management functionalities in this environment: share, rename, add metadata, move, link, add user, etc.
   Home Drive always has the path: **/rug/home/<username>**

2) **Team Drive:** Group environment. In addition to all the data management functionalities you have Home Drive, you can implement policies, add groups and manage group permissions between Team Drives. In order to have a permission to create a Team Drive you need to contact rdms-support@rug.nl
   Team Drive always has the path**: /rug/home/<Team_Name>**

3) **Projects:** To store data sets that are intended to be published. Stricter permission management and user roles exist compared to Team Drives. The reason for this is to make sure that the datasets that are converted to a project comply with quality requirements of becoming a publication.
   Projects always has the path: **/rug/home/Projects/<Project_Name>**

DIFFERENT ENVIRONMENT IN THE RDMS

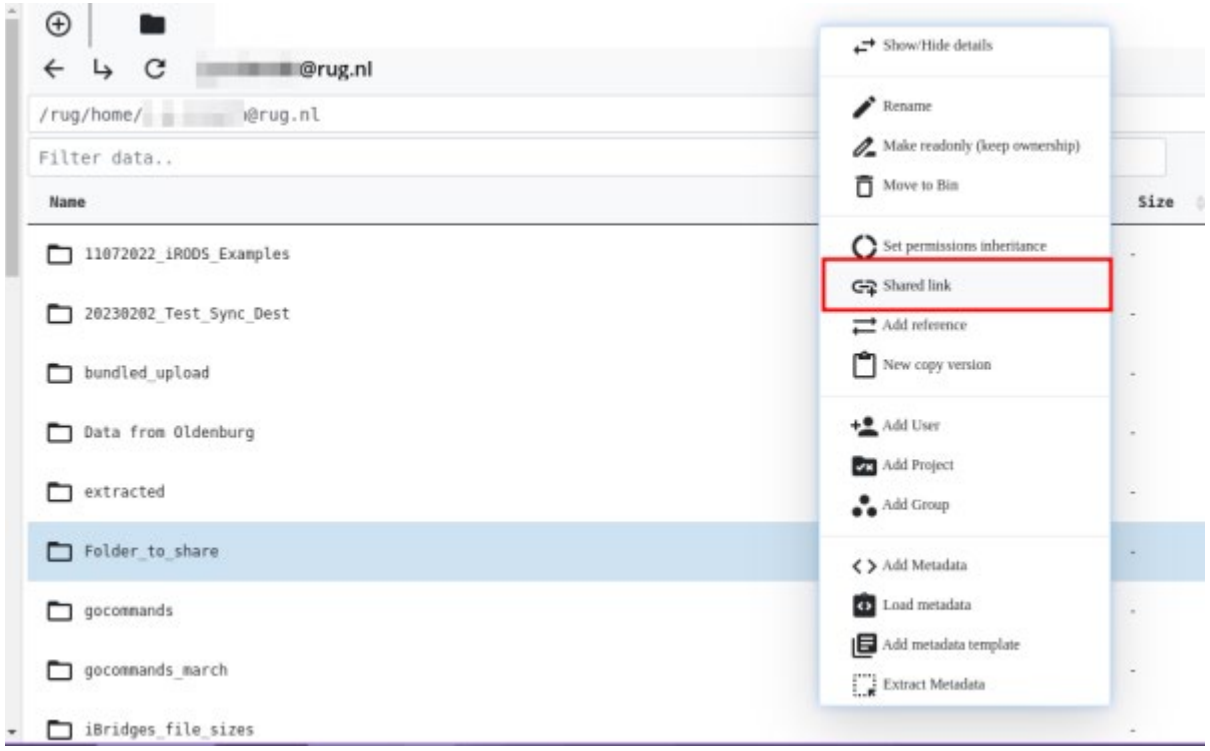
Home Drive:



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO SHARE DATA WITHIN THE RDMS?

Step 1

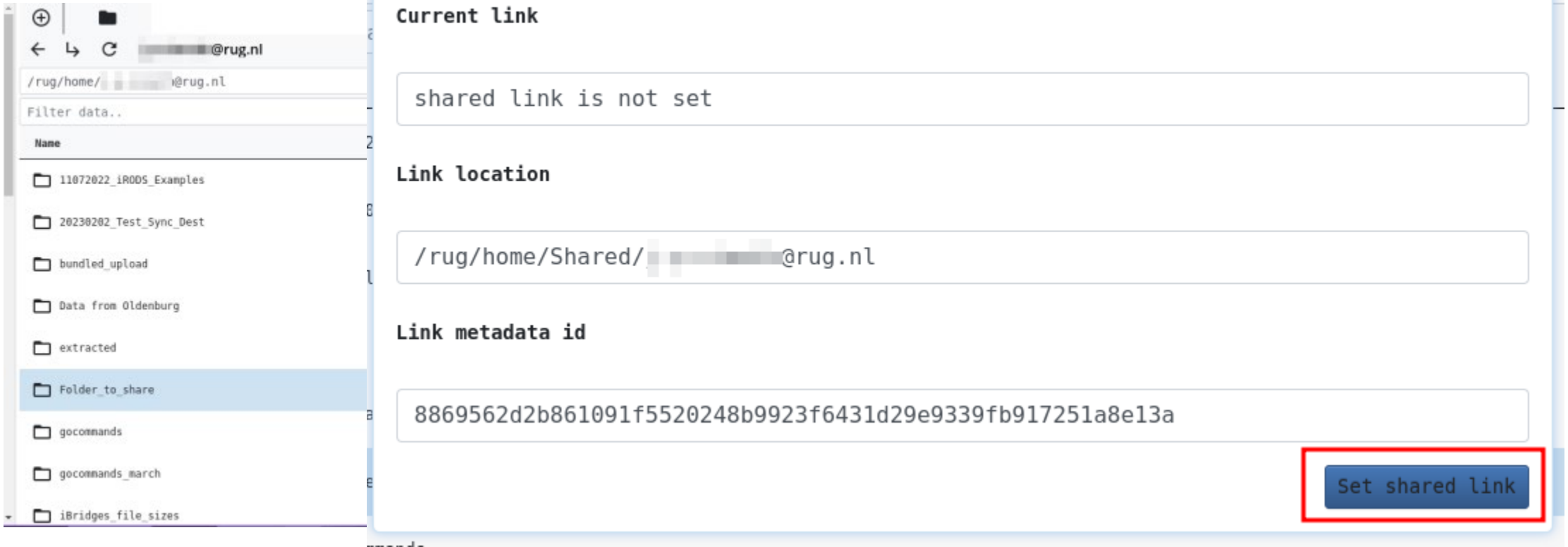


Set a shared link for the folder or file you want to share.

DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO SHARE DATA WITHIN THE RDMS?

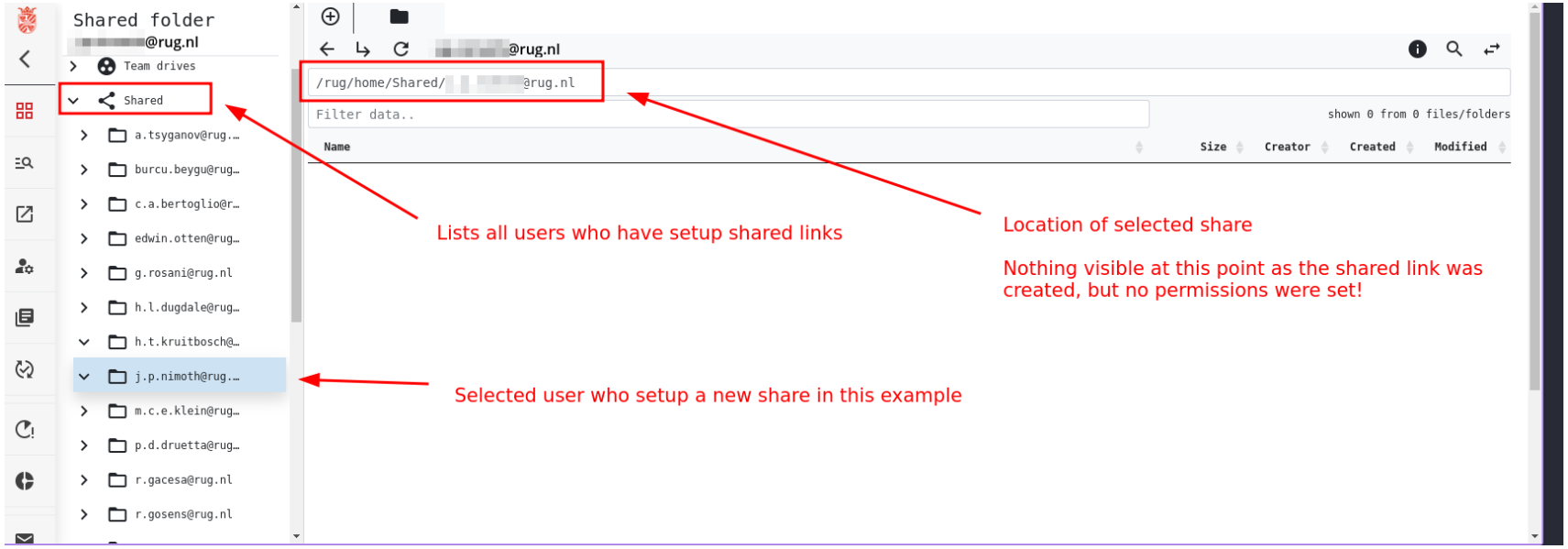
Step 1



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO SHARE DATA WITHIN THE RDMS?

Step 1

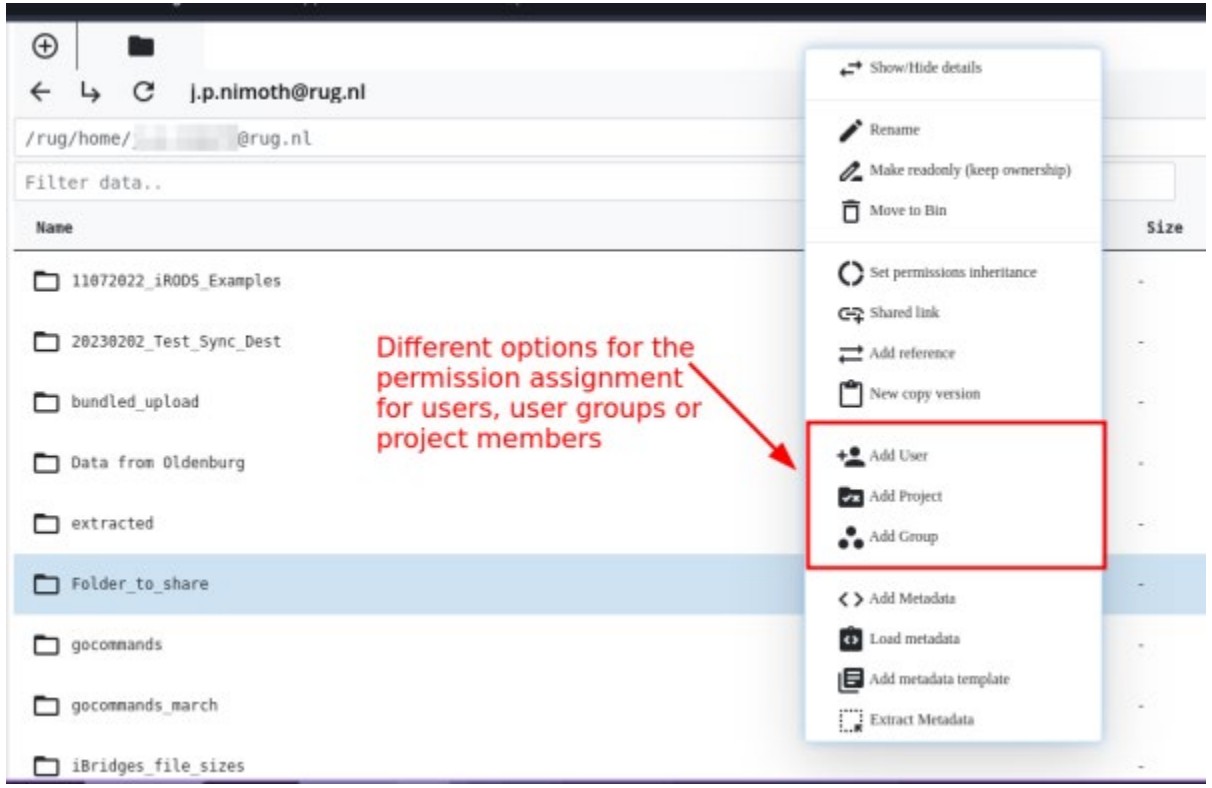


At this point, the name of the user that just setup the share will become visible within the Shared category of the web interface, but no content will be displayed as there were no permissions assigned for the shared object yet!

DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO SHARE DATA WITHIN THE RDMS?

Step 2

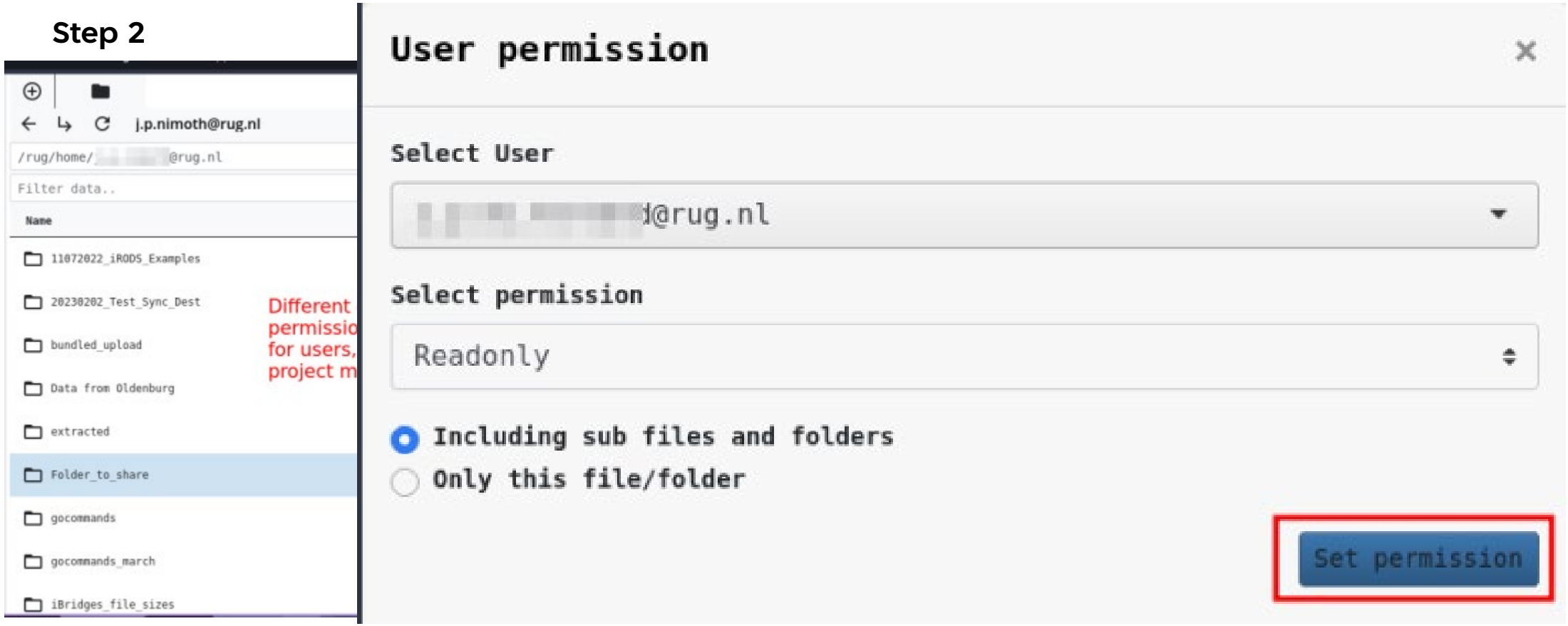


Assign permissions

DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO SHARE DATA WITHIN THE RDMS?

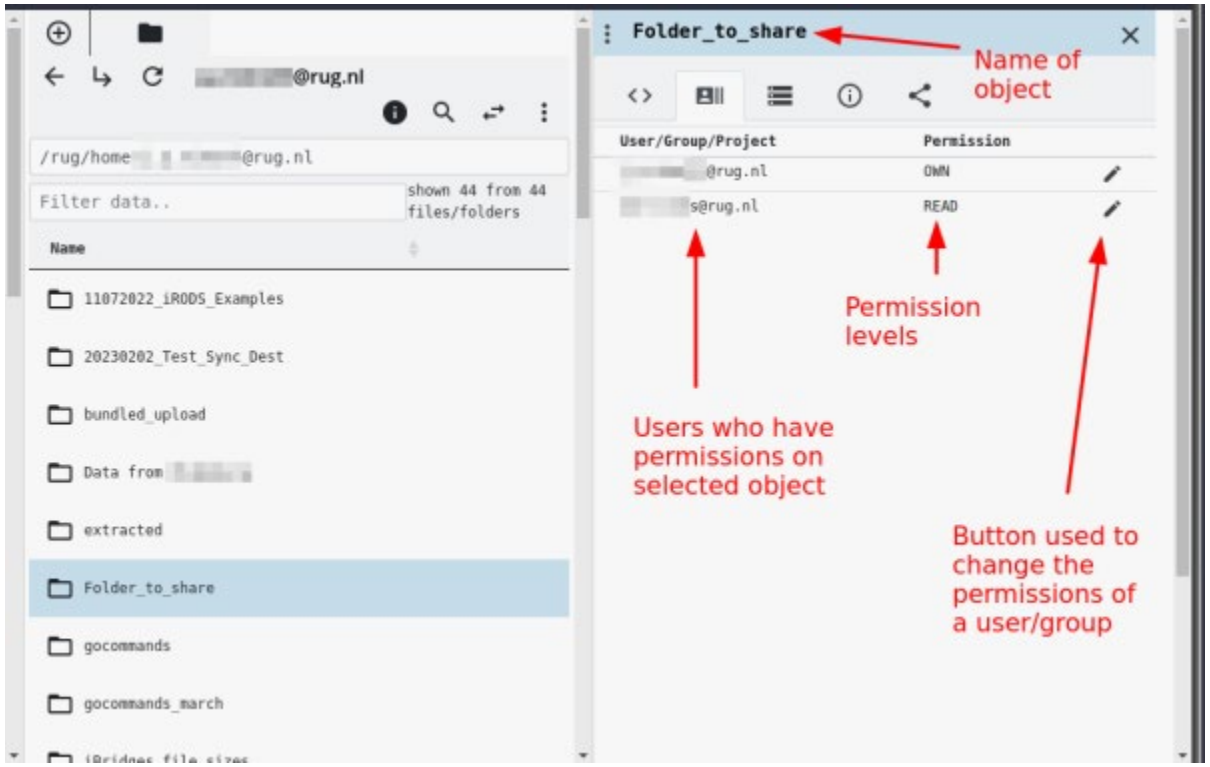
Step 2

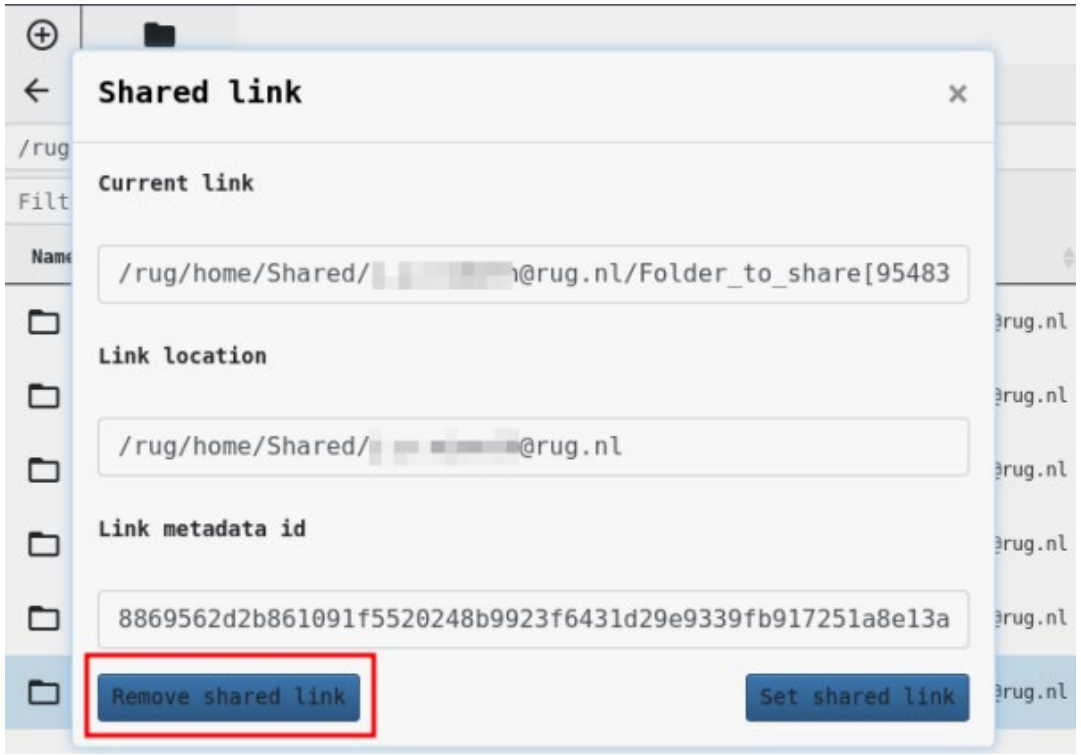


DIFFERENT ENVIRONMENT IN THE RDMS

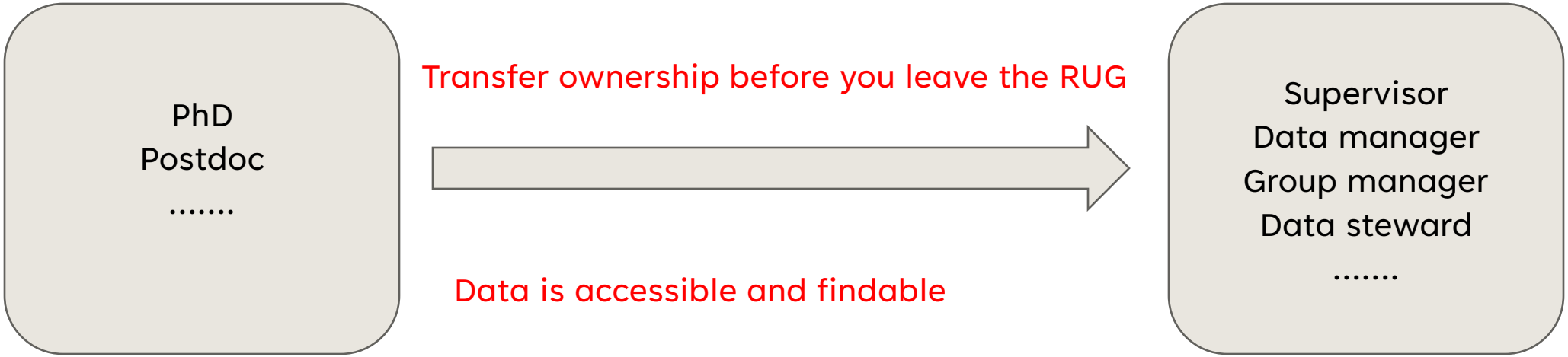
HOW TO SHARE DATA WITHIN THE RDMS?

It is always possible to revoke and change permissions and remove the shared link!





BEST PRACTICE



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

- Metadata makes your data searchable and findable

Scenarios:

- You have hundreds or thousands of data file you frequently generate,
- As a supervisor or data manager, you oversee students and projects each of them regularly generate many data files/output
- In the lab lots of data is generated daily with different machines/different set-ups/parameters
- .....

How do you find the one you need two years later? You search using key words which comes from metadata

DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

In the RDMS there are two types of metadata:

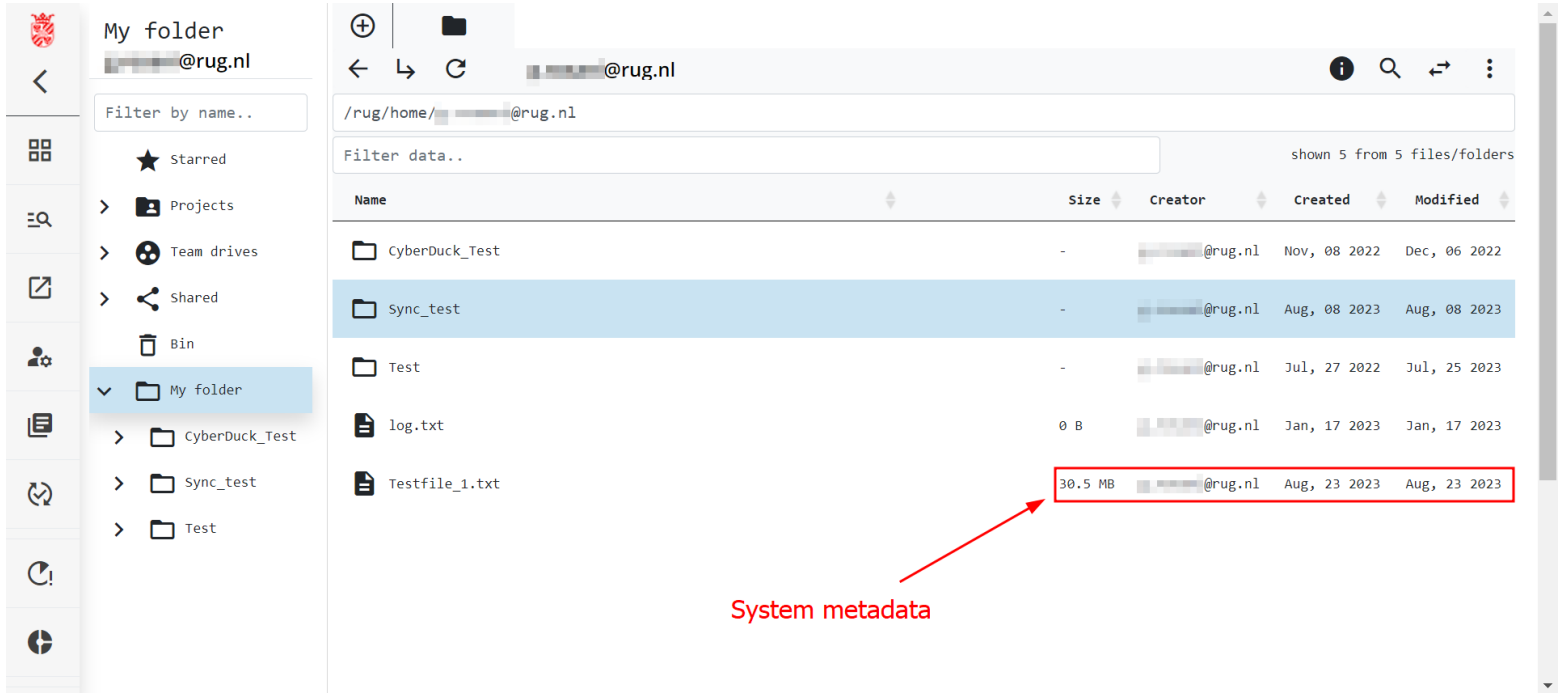
1) System metadata
2) User-defined metadata

DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

1) System metadata

This is the minimum level of metadata that is attached to data (file/folder) is the one generated by the system itself: The date it was uploaded to RDMS, the owner etc.



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

2) User-defined metadata

This type of metadata is added by a user. Based on your data, how you organize the, how you want to search for them, you have many options to create metadata.

You can:

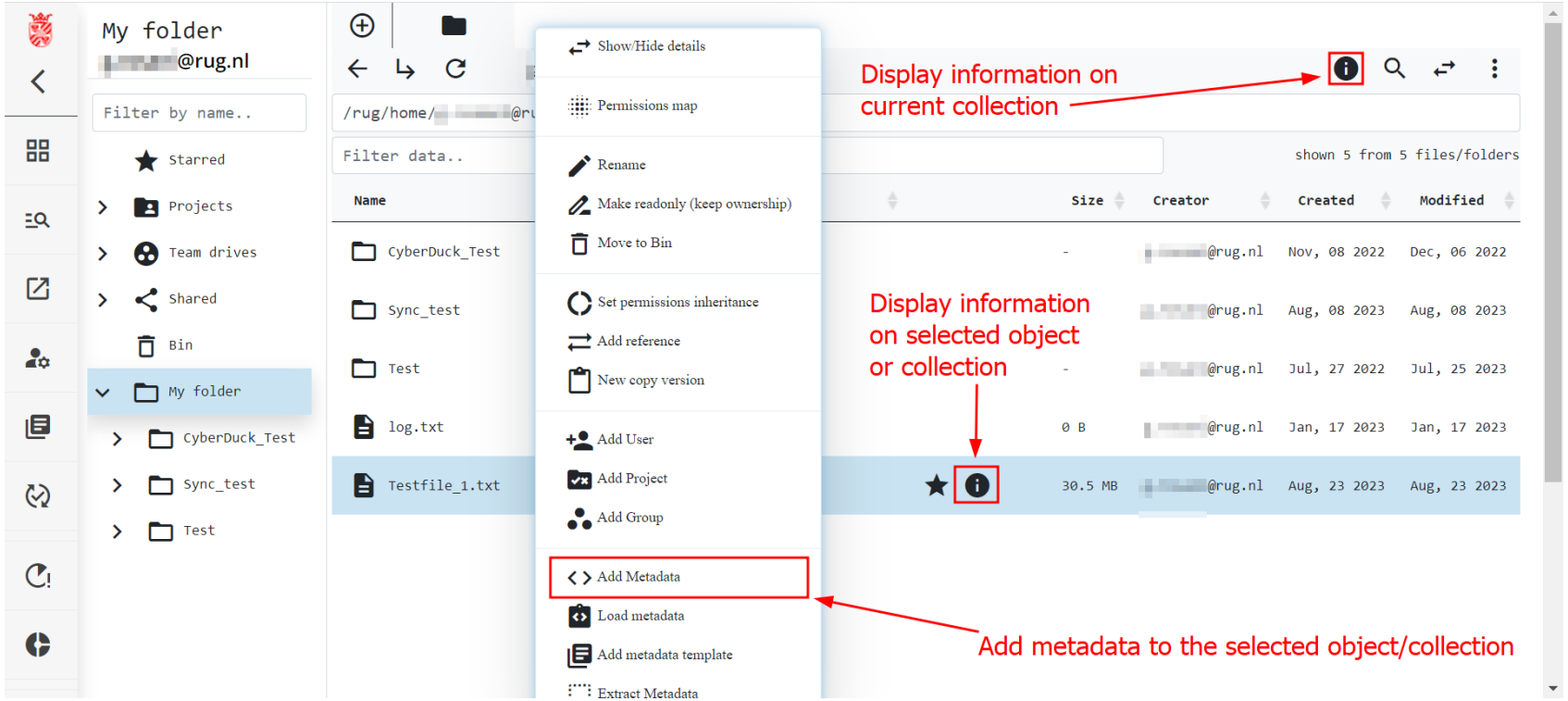
- Add, edit and delete
- Upload
- Extract
- Use templates

Metadata management is possible both via the Web interface and the iCommands

DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

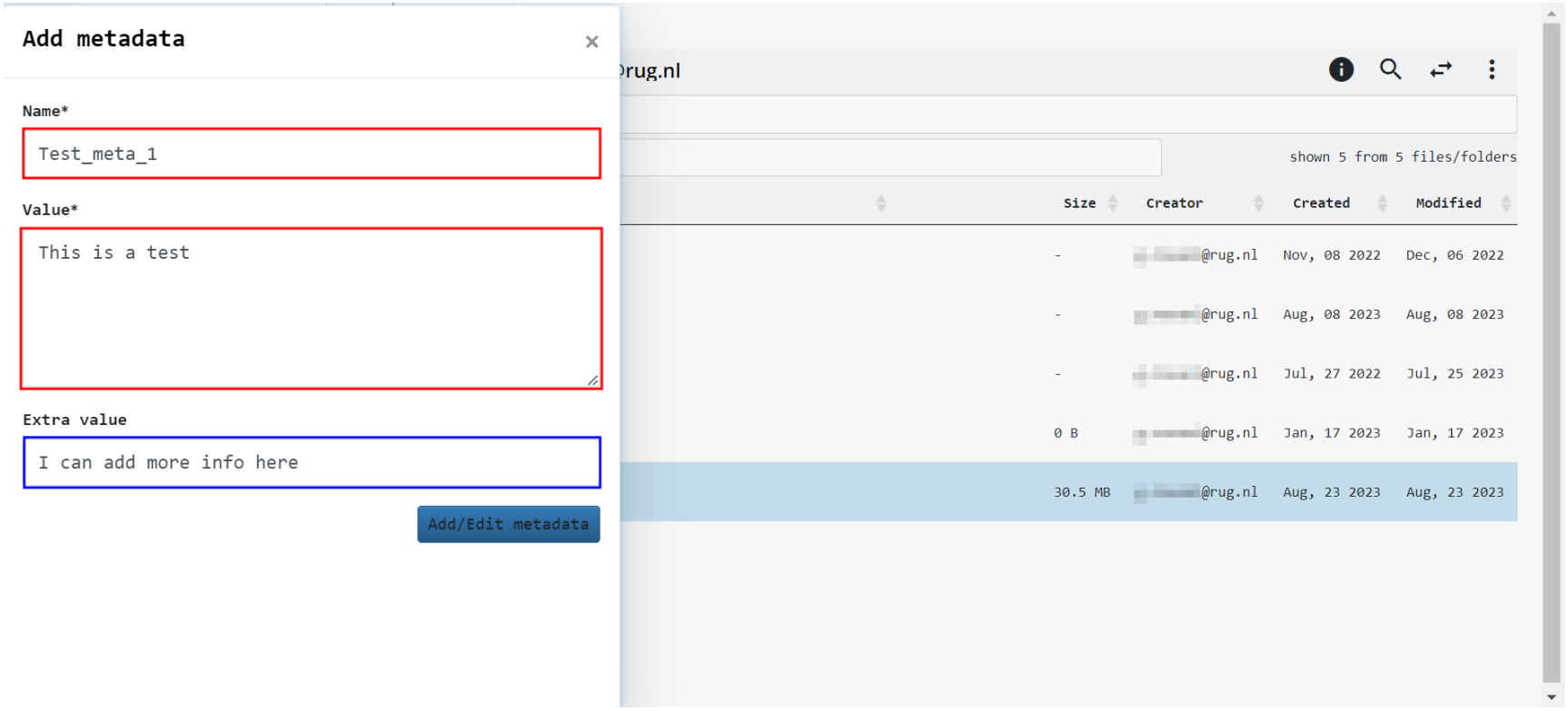
Add, update and delete metadata



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

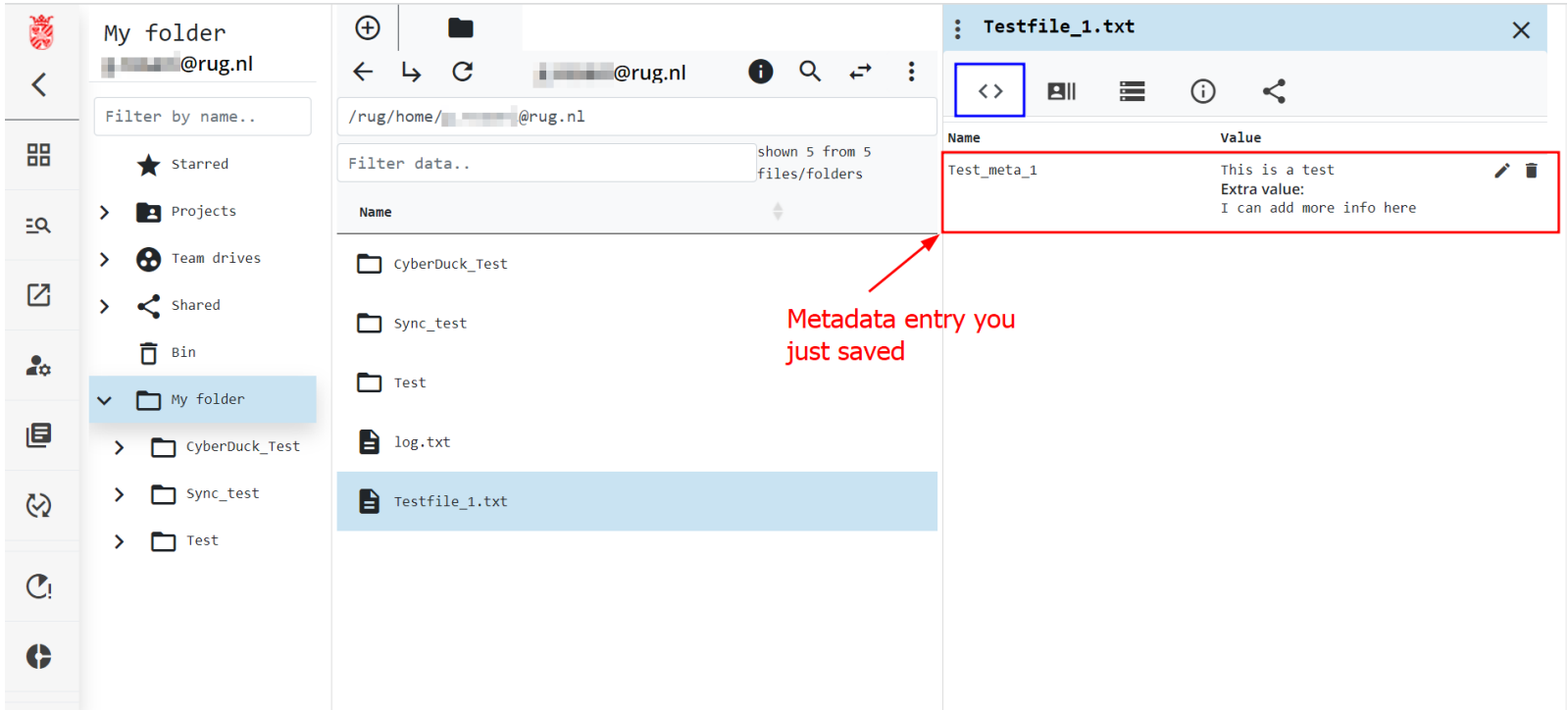
Add, update and delete metadata



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

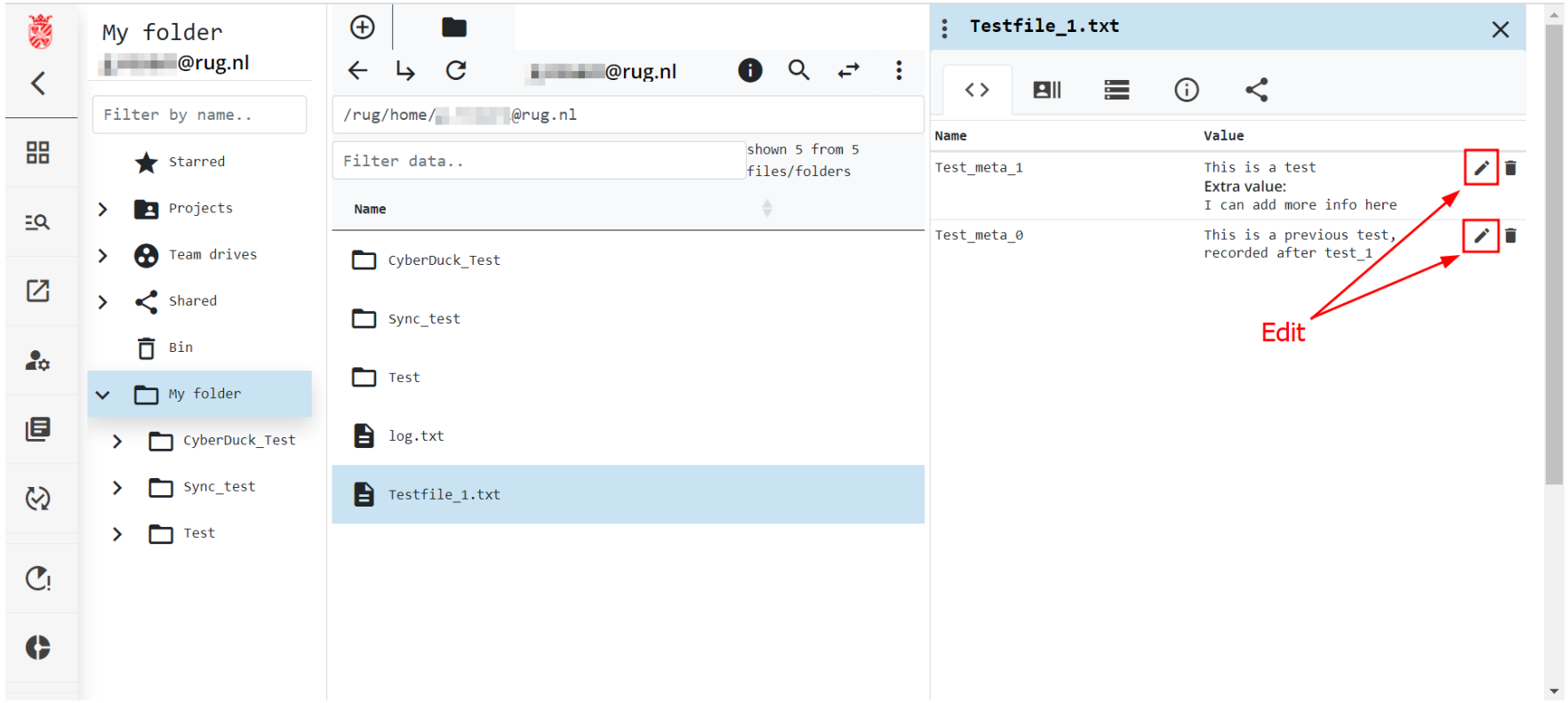
Add, update and delete metadata



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

Add, update and delete metadata



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

Upload metadata

Instead of adding each key-value pair one-by-one, you can create one metadata file and upload it. You can use it as you template and each time a similar data needs metadata you can reuse it.

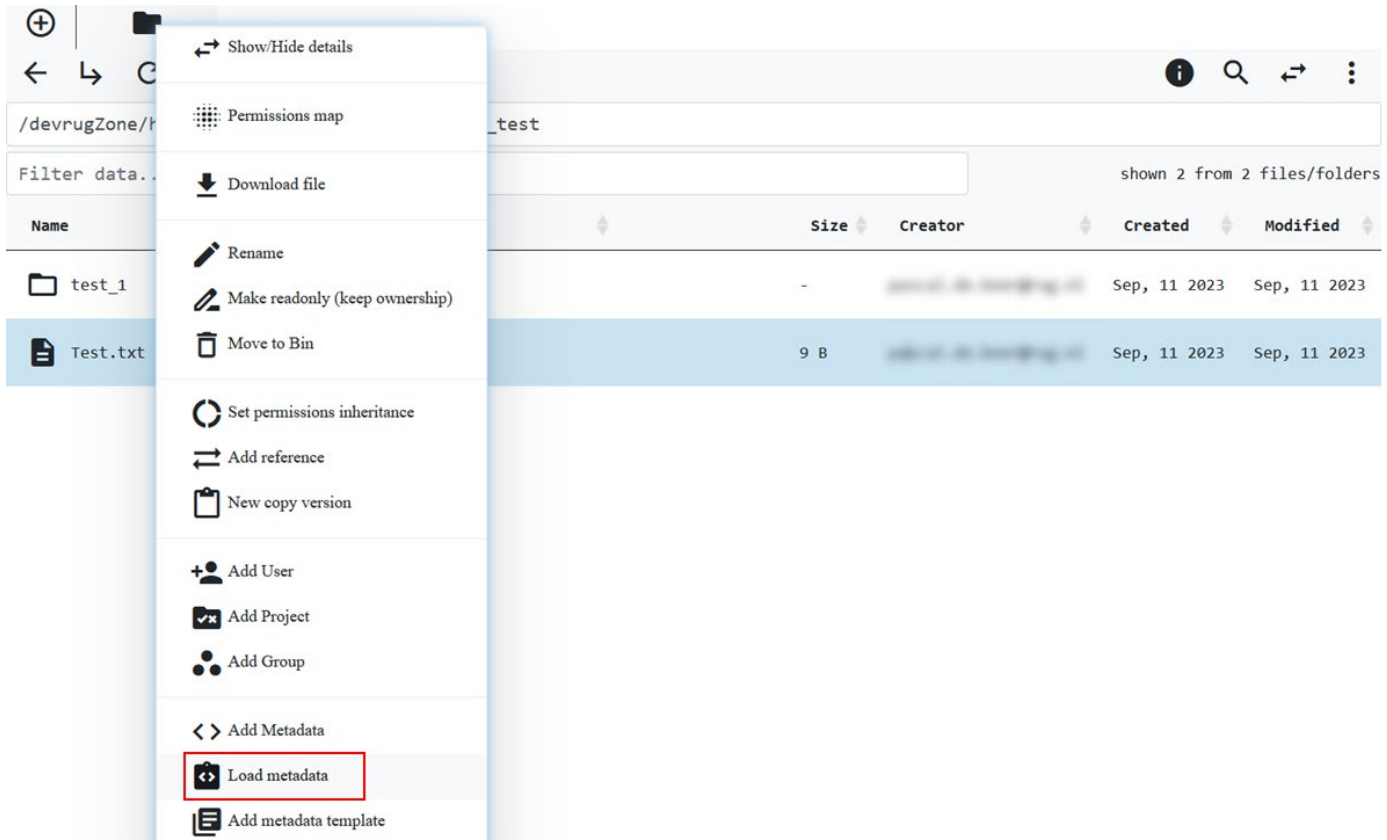
```
[
{"Attribute" : "Author",
"Value": "Tester_1",
"Unit": ""
},
{"Attribute" : "Distance",
"Value": "100",
"Unit": "km"
},
{"Attribute" : "Location",
"Value": "Groningen",
"Unit": ""
}
]
```

Create a file with extension **.txt** or **.json** exactly with this format and upload it to the RDMS.

DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

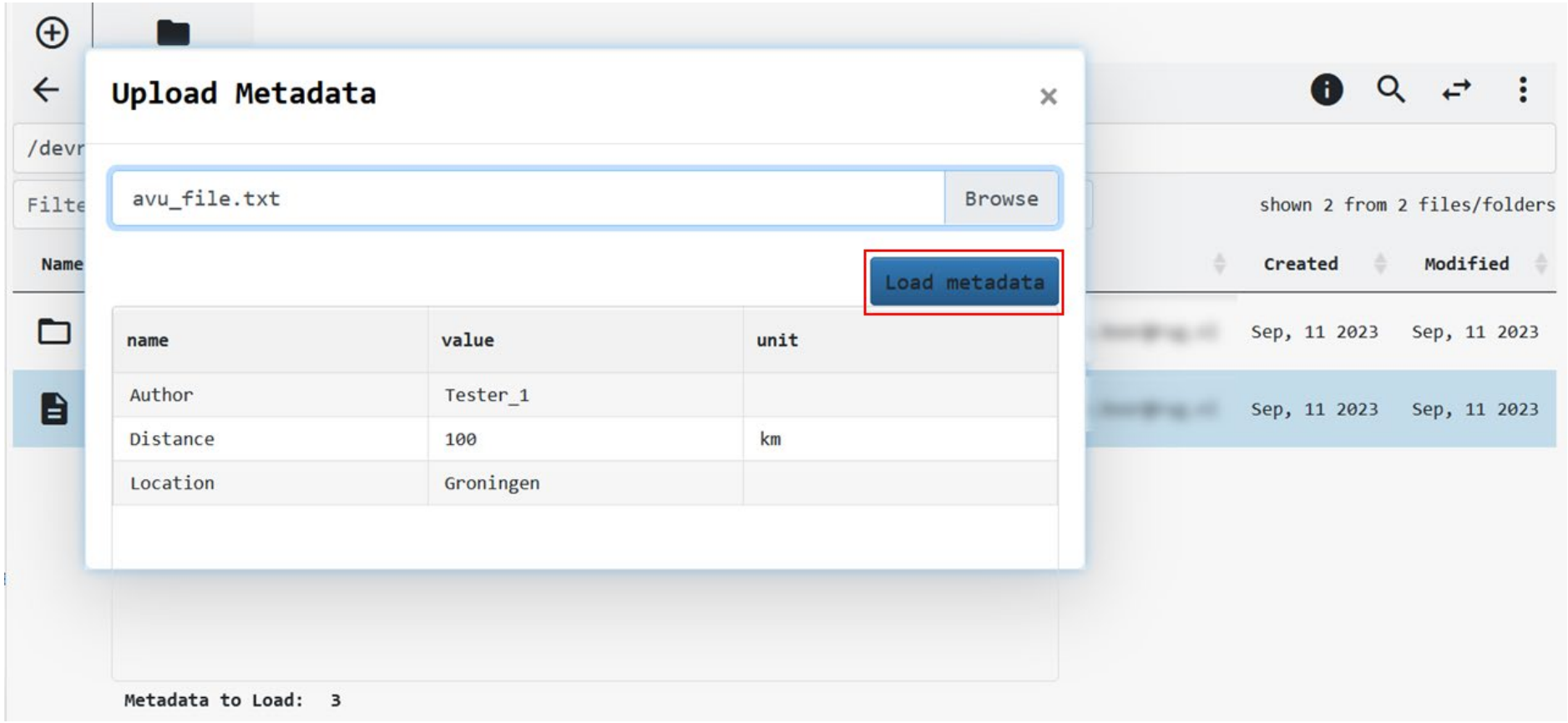
Upload metadata



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

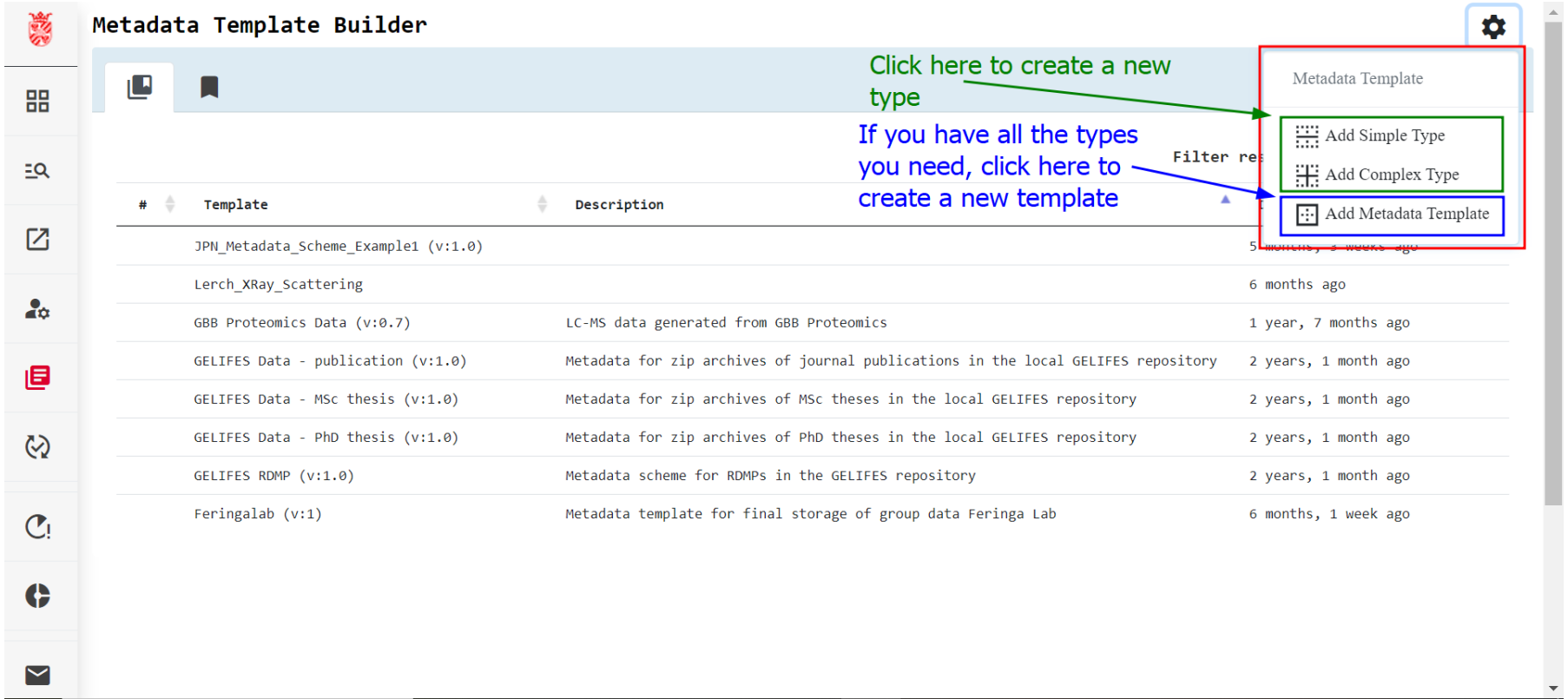
Upload metadata



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

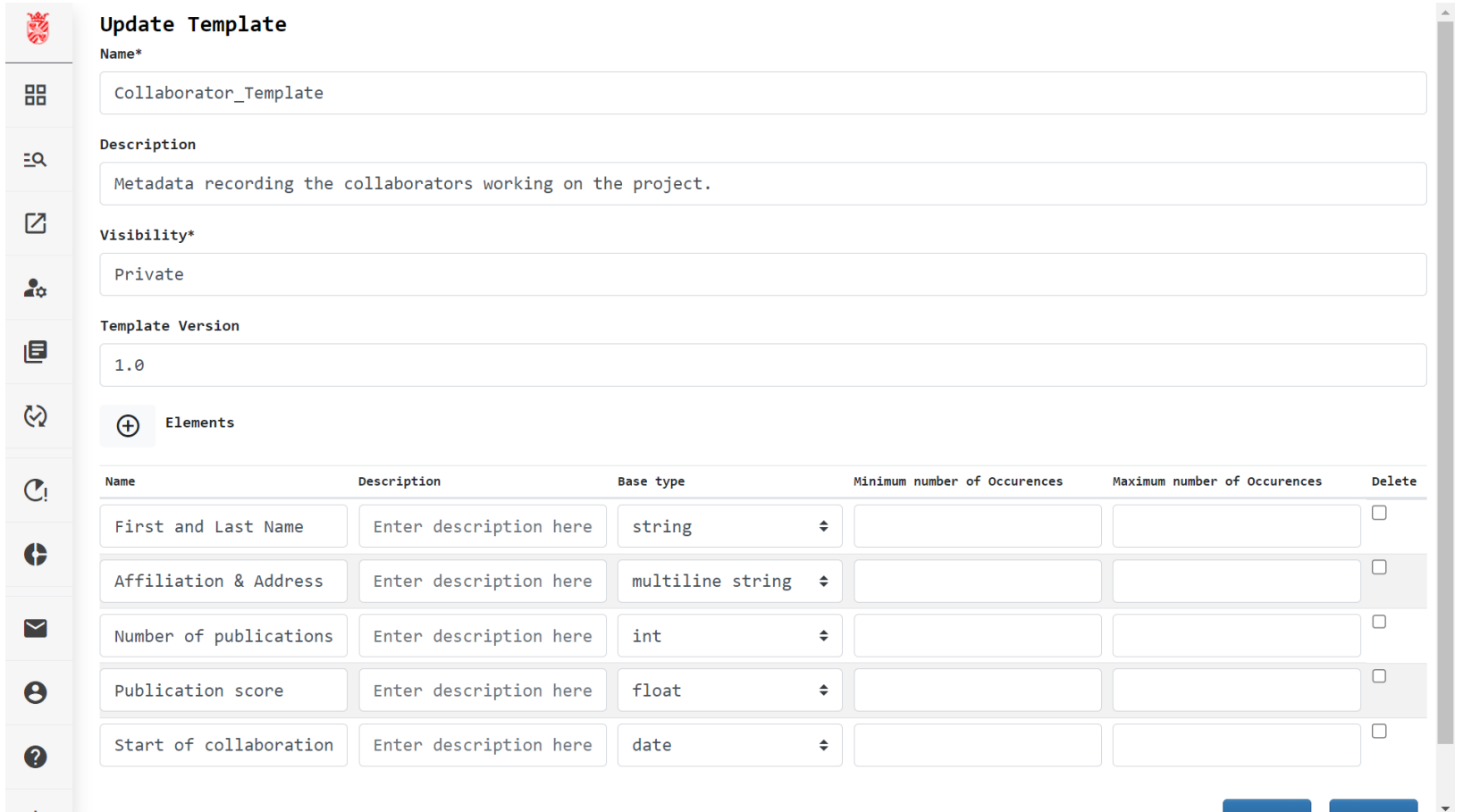
Create metadata template



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO WORK WITH METADATA

Create metadata template



BEST PRACTICE

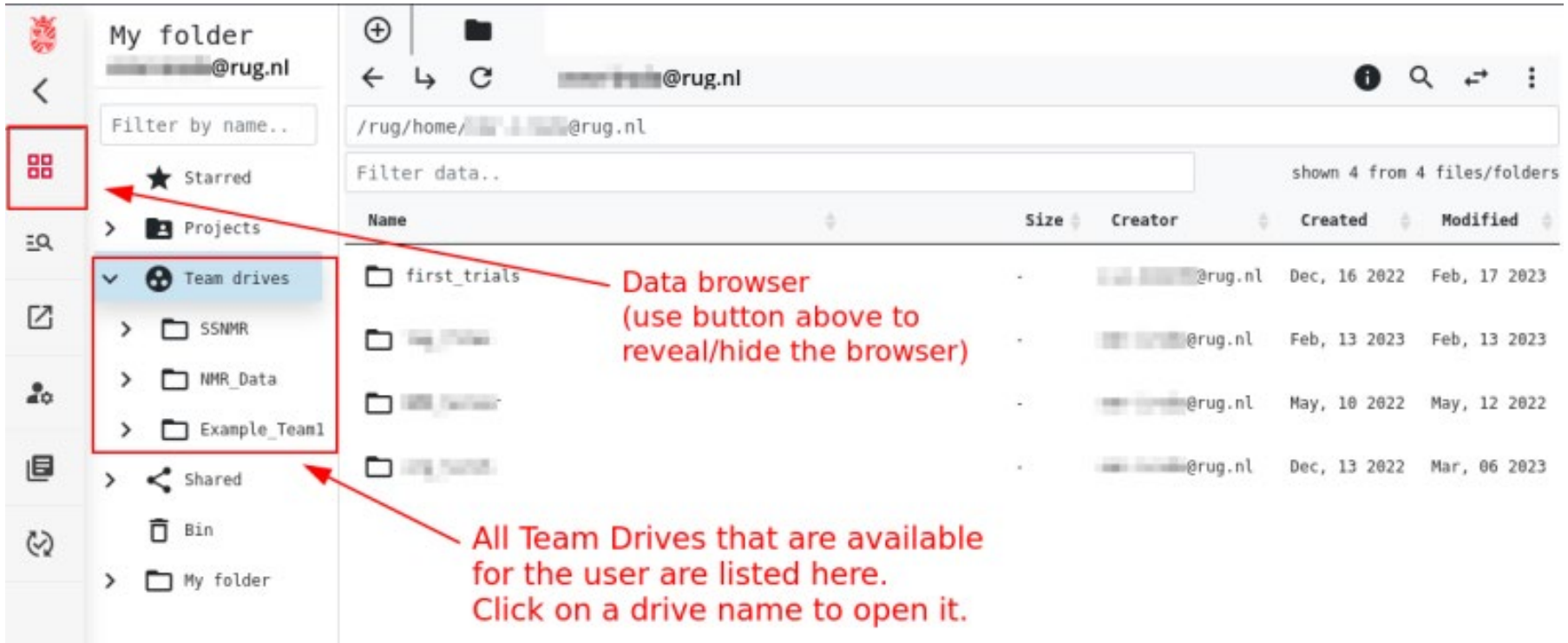
- Invest in creating metadata templates or metadata files. You will create them only once or twice but will use them frequently.

- Invest in useful metadata key/value pairs and types.

DIFFERENT ENVIRONMENT IN THE RDMS

Team Drive:

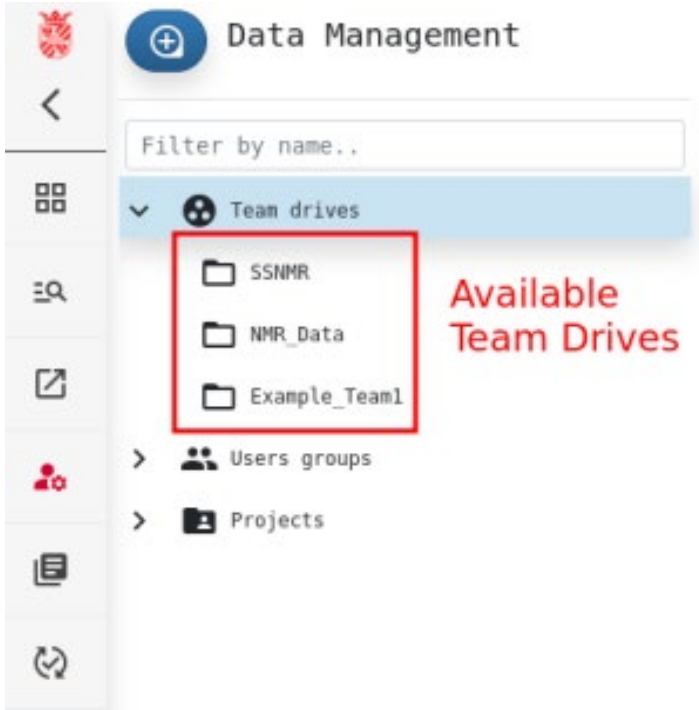
From a normal (not a group admin or owner) user point of view



DIFFERENT ENVIRONMENT IN THE RDMS

Team Drive:

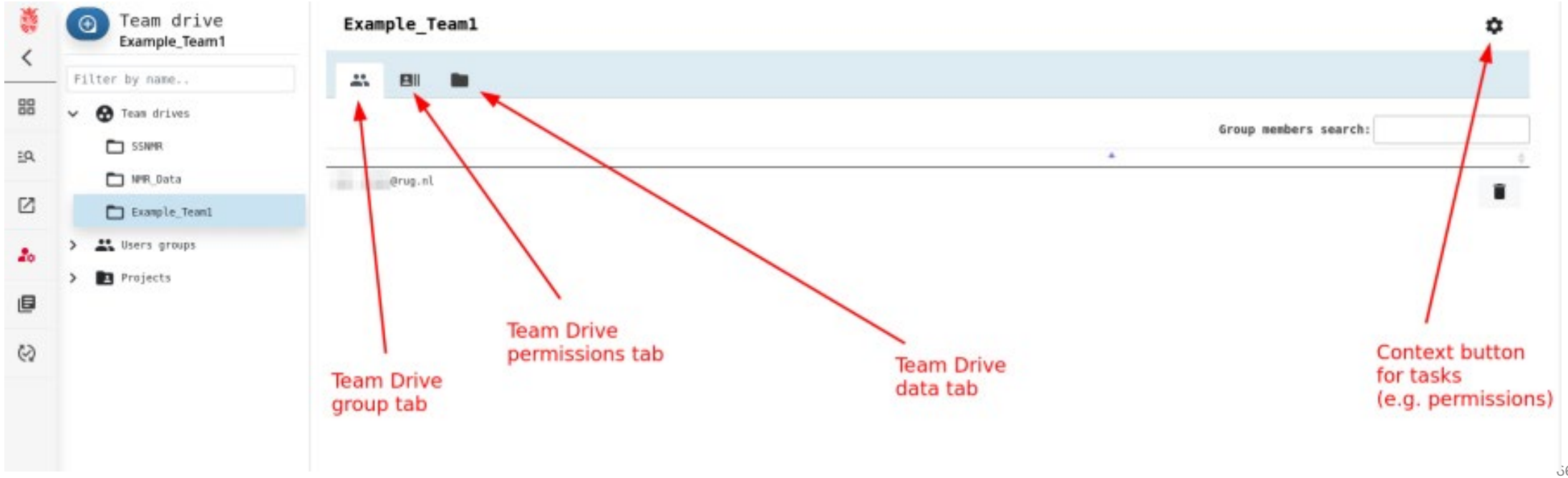
From a group admin or owner point of view



DIFFERENT ENVIRONMENT IN THE RDMS

Team Drive:

From a group admin or owner point of view

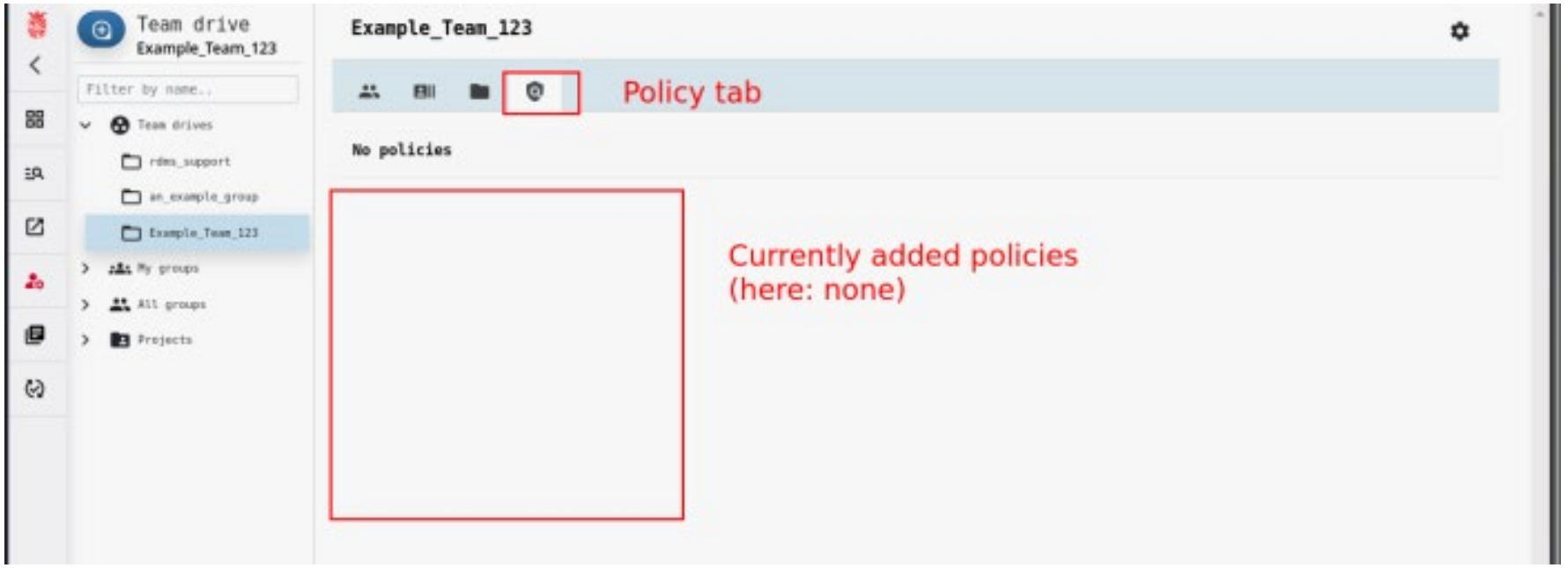


DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO IMPLEMENT AUTOMATIZED POLICIES?

Step 1

Select the Team Drive or folder with a Team Drive for which you want to implement policy. And go to the Policy tab.

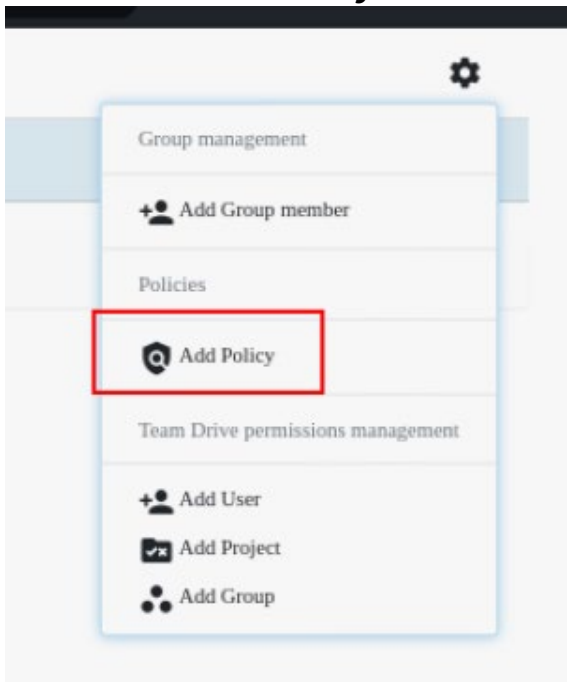


DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO IMPLEMENT AUTOMATIZED POLICIES?

Step 2

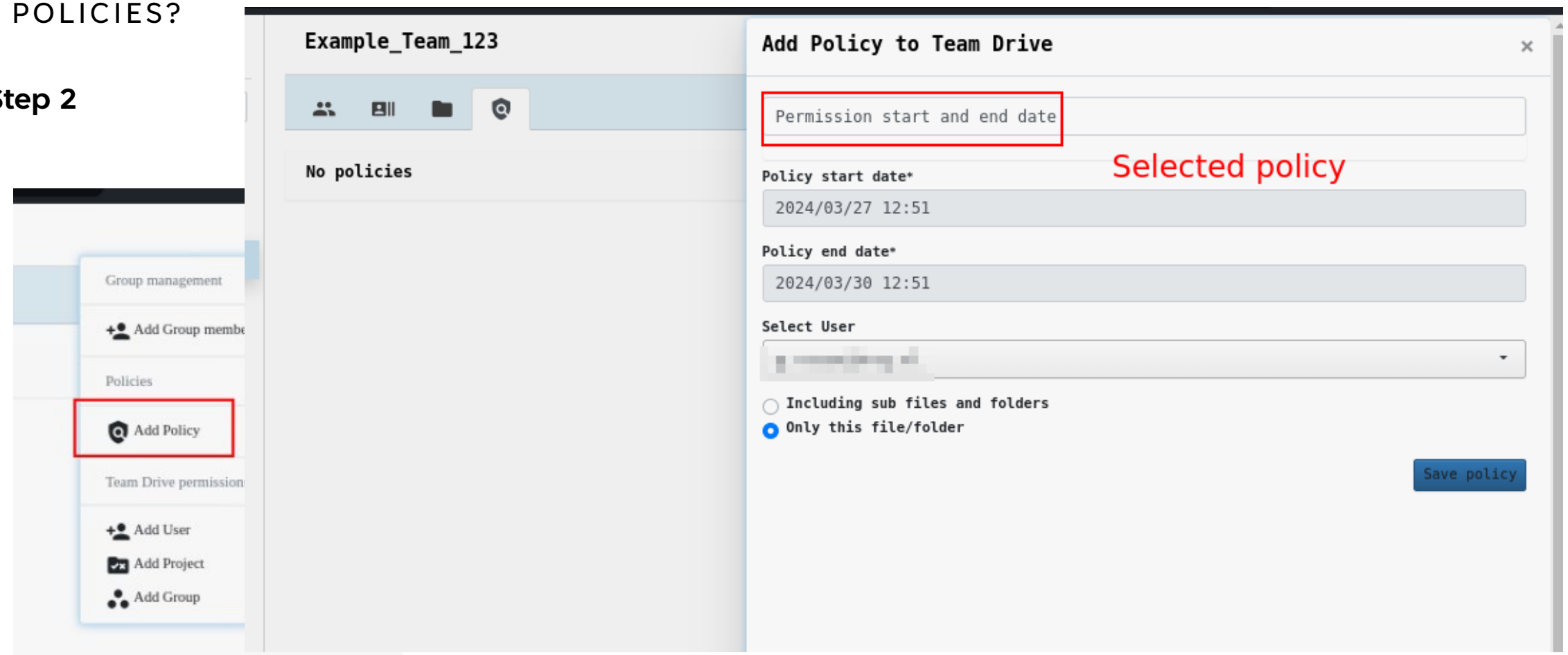
Add and create Policy for users by assigning them permissions.



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO IMPLEMENT AUTOMATIZED POLICIES?

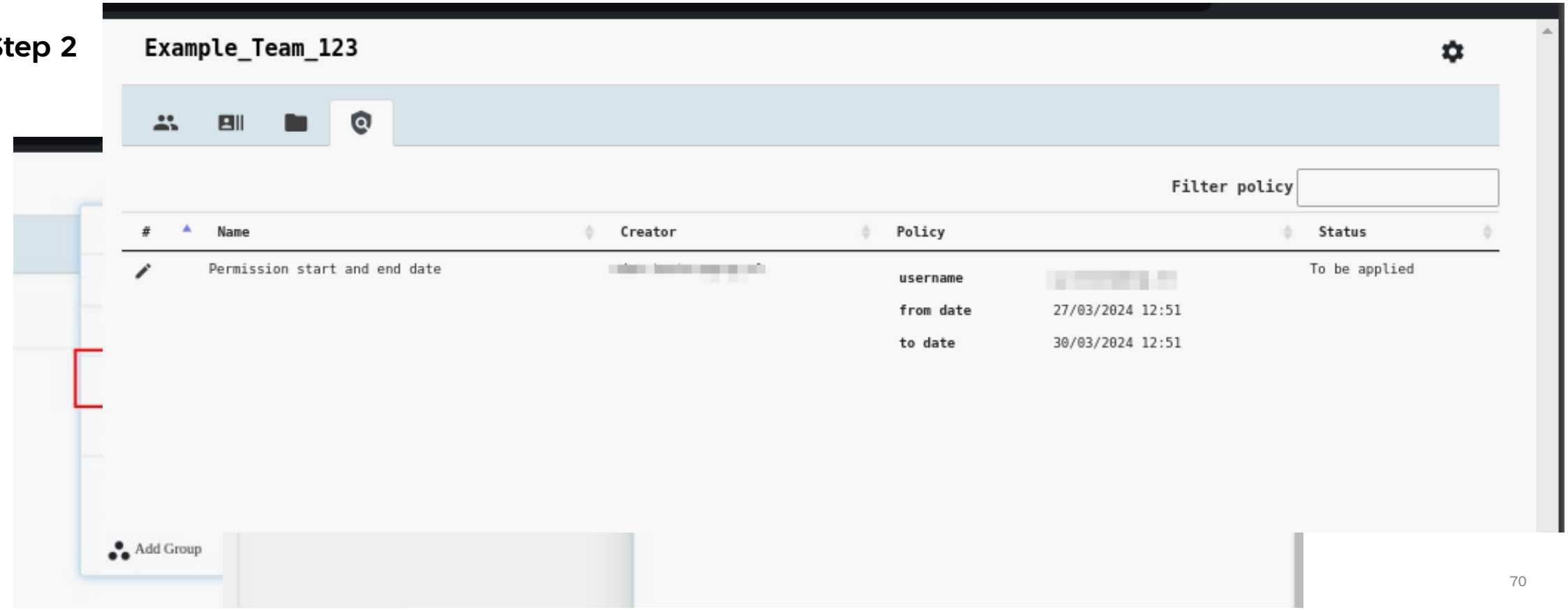
Step 2



DIFFERENT ENVIRONMENT IN THE RDMS

HOW TO IMPLEMENT AUTOMATIZED POLICIES?

Step 2



BEST PRACTICE

- Depending on the size of a research group and amount of research output:
  - Create multiple groups for different folders in a Team Drive

    And/Or
  - Create couple of Team Drives

- Example groups: PhD students
  Special projects
  Sub-groups in a research group
  Machines/detectors in labs
  .....
- Example folder structures: PhD Thesis
  MSc Thesis
  Machines/detectors in labs
  .....
- Assign permissions to different users for them to transfer their data to the designated folder(s) in a Team Drive.
- Make a research group policy to force members to transfer their data
- Work with metadata

ROLES AND PERMISSIONS

Permissions in RDMS

Permission Level	Read	Modify	Create New	Delete	Share
Null	●	●	●	●	●
Read	✓	●	●	●	●
Write	✓	✓	✓	●	●
Own	✓	✓	✓	✓	✓

Own: The user owns the data object (file) or the collection (folder) and has the full permission on reading, modifying (including deletion), and sharing.

Write: The user has read and write access to the object.

Read: The user can only read the object or its content. This also allows to make a (editable) copy of the file/folder.

Null: The user does not have any permission on the object. One can use 'none' when removing the previously assigned permissions to a user.

ROLES AND PERMISSIONS

Roles in RDMS

- **Group admin:** A group admin can create new team drives, groups, and projects and to manage them.
- **Data manager:** Data Manager is responsible for verifying that the data sent to the archive is complete and uncorrupted, and giving the final approval of the archive. This role is relevant during the Archiving workflow.
- **Metadata manager:** Metadata Manager is responsible for verifying and completing the metadata information related to the archive. This role is relevant during the Archiving workflow.
- **Project admin:** Project admin has the owner role for a project. This role is relevant during the Archiving workflow.
- **Project manager:** Project Manager is responsible for assigning the data manager and metadata manager roles, starting the archiving process. This role is relevant during the Archiving workflow.

FUTURE

- **Enable audit**
- **Enable publishing**
- **Enable non-RUG access**

Plan:

- **Integration with VRW, UWP and eLab journals**

THANK YOU

rdms-support@rug.nl