Data storage concepts

Optimally secured in every phase of the data lifecycle

RDM in HPC
22. September

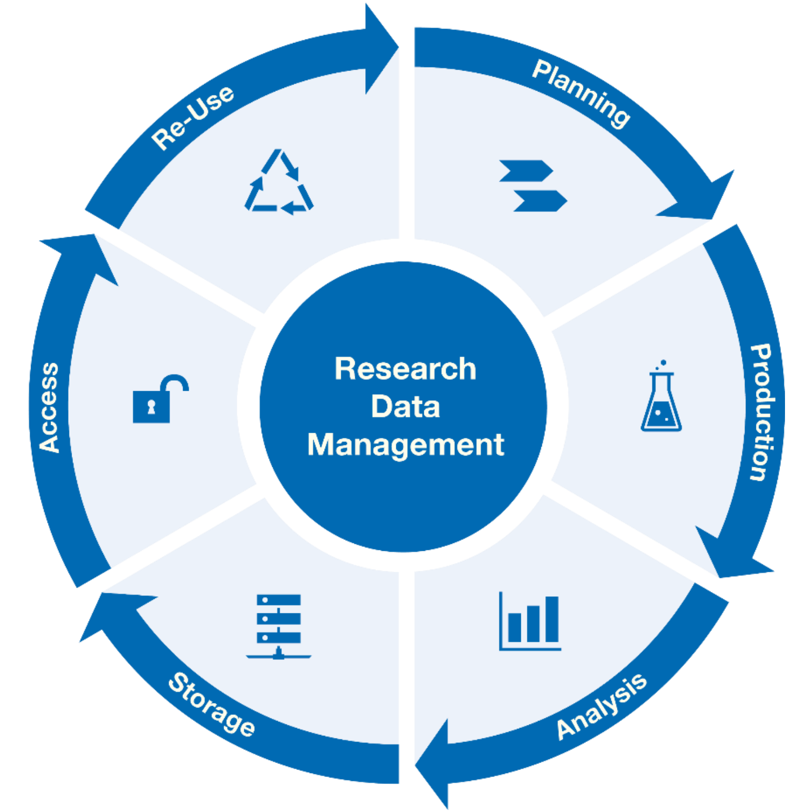
Marcel Nellesen (RWTH Aachen University)
nellesen@itc.rwth-aachen.de

Katja Jansen (RWTH Aachen University)
k.jansen@itc.rwth-aachen.de

Research data life cycle

- Research projects are a long term commitment

- Data is constantly created
  - First draft / proposal
  - Setup
  - Measurement
  - Analysis
  - Reports
  - Reusage



Storage types

RDS Web

- **Technology:**
  - RDS = Research Data Storage (object storage)
  - S3

- **Features:**
  - Enforces entering of metadata before data can be stored
  - Interaction only through Coscine (Web-Interface)

- **Envisioned use:**
  - Mostly useable for smaller files
    for files > 2 GB REST-API or S3 should be used (reason: browser limitations)

- **Default:**
  - 25 GB for each project (can be adjusted up to 100 GB) → for RDS-qualified universities/UAS

Storage types

RDS S3

- **Technology:**
  - RDS = Research Data Storage (object storage)
  - S3

- **Features:**
  - Easy transfer of data
  - Interaction through Coscine and the S3

- **Envisioned use:**
  - Large files (> 2- 3 GB)
  - Automization processes possible via REST-API

- **Default:**
  - nothing → application needed via Jards

Storage types

RDS Worm

- **Technology:**
  - RDS = Research Data Storage (object storage)
  - S3
- **Features:**
  - WORM (**W**rite **O**nce **R**ead **M**any)
  - Interaction through Coscine and the S3
- **Envisioned use:**
  - used for data with very high secure standards
- **Default:**
  - nothing → application needed via Jards

Storage types

Linked Data

- **Technology:**
  - RDF
- **Features:**
  - Referencing of external storage systems

  → Safety depends on the external storage system
  - Providing of metadata
- **Envisioned use:**
  - Data which is stored in an external system, just adding metadata in Coscine

Storage types

GitLab

- **Technology**:
  - GitLab
- **Features**:
  - Versioning of code
  - Interaction through Coscine or GitLab
  - Providing of metadata
- **Envisioned use:**
  - Data which is (already) stored in GitLab, just adding metadata in Coscine

Storage types

Cluster - Home

- **Technology:**
  - NFS/CIFS

- **Features:**
  - Regular backups
  - Regular snapshots

- **Envisioned use:**
  - Source code
  - Configuration files

- **Default:**
  - 150 Gb (easily extendable to 200Gb)

Storage types

Cluster - Work

- **Technology:**
  - NFS/CIFS
- **Features:**
  - Regular snapshots
- **Envisioned use:**
  - Output files
  - Working data
- **Default:**
  - 250 Gb (easily extendable to 350Gb)

Storage types

Cluster - HPC Work

- **Technology:**
  - Lustre

- **Features:**
  - Suitable for ~ 50.000 files
  - Neither backups nor snapshots

- **Envisioned use:**
  - IO intensive jobs
  - Large files

- **Default:**
  - 1 Tb (easily extendable to 30 Tb)

Storage types

RWTH Publications

- **Technology:**
  - S3

- **Features:**
  - Enforces entering of metadata before data can be stored
  - Interaction only through Coscine

- **Envisioned use:**
  - For data and metadata which can not be published in a specific repository
  - Mostly for RWTH members

Storage types

Archive

- **Technology:**
  - S3

- **Features:**
  - Long term archiving of data (10 years)
  - → Good scientific practice

- **Envisioned use:**
  - Enable reuse of data
  - FAIR data

Storage types

Long Term Archive

- **Technology:**
  - – varies
- **Features:**
  - – Long term archiving of data (more than years)
  - → Good scientific practice
- **Envisioned use:**
  - – Long term storage of important data
- **Remarks:**
  - – Currently not available at RWTH

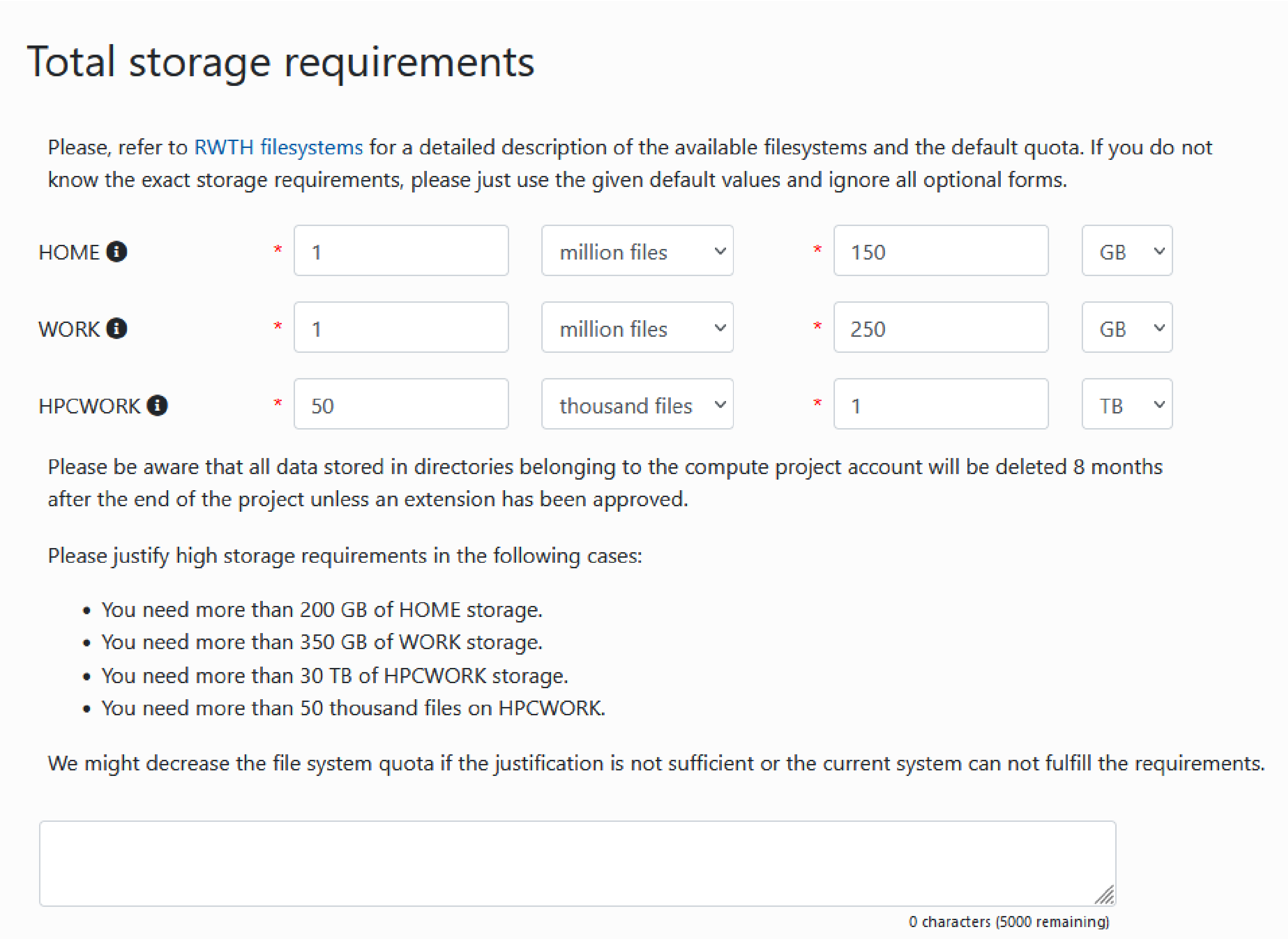
Many different storage types, but…

how to apply and manage the storage?

How to apply for storage

HPC Systems

- Applications for computation time
  - Scientific led review process (Jards)
- Availability: usually project end + 8 months
- Reviewed by the HPC-team and domain experts



Total storage requirements

Please, refer to RWTH filesystems for a detailed description of the available filesystems and the default quota. If you do not know the exact storage requirements, please just use the given default values and ignore all optional forms.

HOME ℹ	*	1	million files	* 150	GB
WORK ℹ	*	1	million files	* 250	GB
HPCWORK ℹ	*	50	thousand files	* 1	TB

Please be aware that all data stored in directories belonging to the compute project account will be deleted 8 months after the end of the project unless an extension has been approved.

Please justify high storage requirements in the following cases:

- You need more than 200 GB of HOME storage.
- You need more than 350 GB of WORK storage.
- You need more than 30 TB of HPCWORK storage.
- You need more than 50 thousand files on HPCWORK.

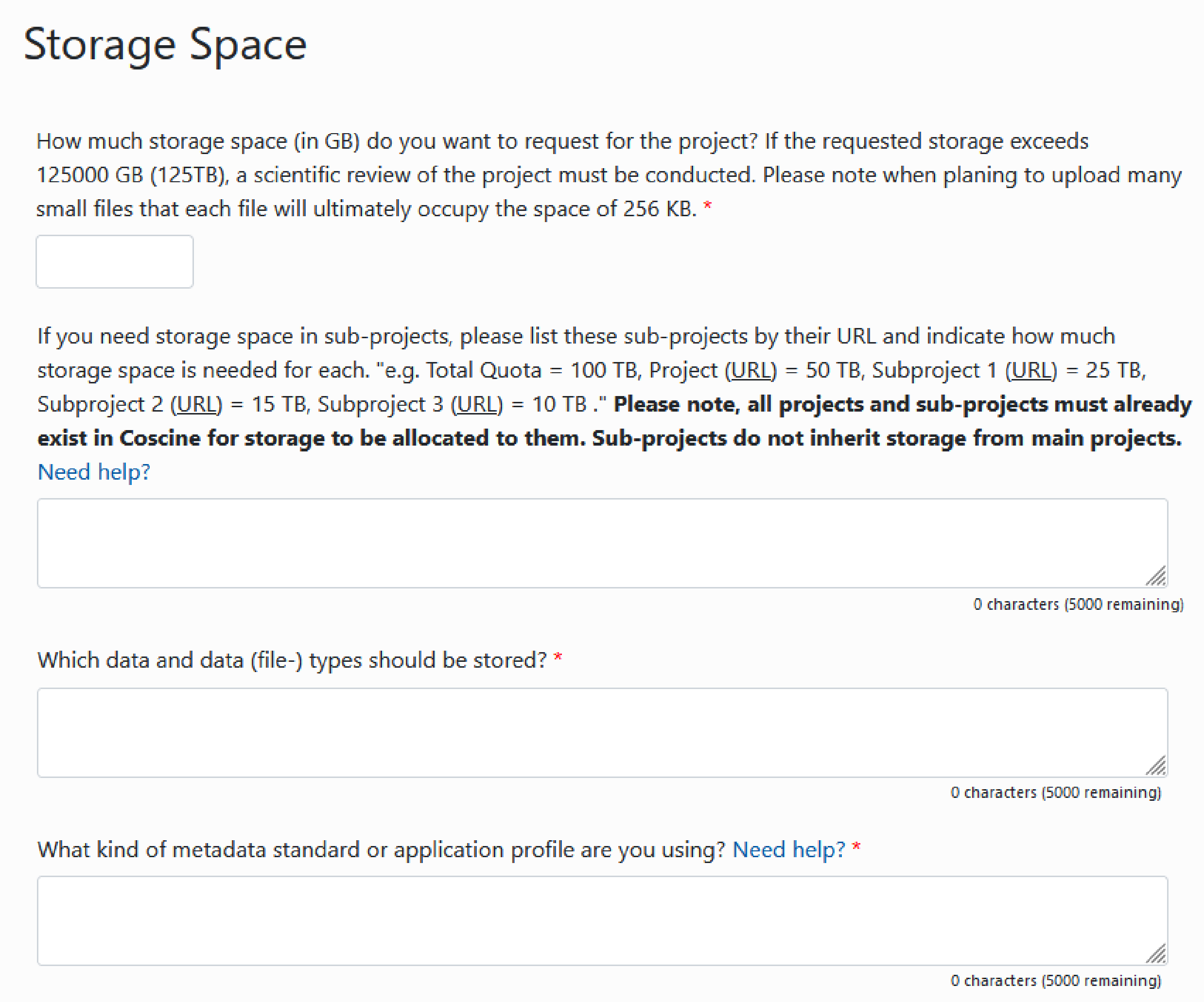
We might decrease the file system quota if the justification is not sufficient or the current system can not fulfill the requirements.

0 characters (5000 remaining)

How to apply for storage

RDS S3 Storage

- Applications for storage space
  - Scientific led review process (Jards)
- Availability: usually project end + 10 years
- Reviewed by the RDM-team and domain experts
- Better Scaleability



Storage Space

How much storage space (in GB) do you want to request for the project? If the requested storage exceeds 125000 GB (125TB), a scientific review of the project must be conducted. Please note when planing to upload many small files that each file will ultimately occupy the space of 256 KB. *

If you need storage space in sub-projects, please list these sub-projects by their URL and indicate how much storage space is needed for each. "e.g. Total Quota = 100 TB, Project (URL) = 50 TB, Subproject 1 (URL) = 25 TB, Subproject 2 (URL) = 15 TB, Subproject 3 (URL) = 10 TB ." **Please note, all projects and sub-projects must already exist in Coscine for storage to be allocated to them. Sub-projects do not inherit storage from main projects.** Need help?

0 characters (5000 remaining)

Which data and data (file-) types should be stored? *

0 characters (5000 remaining)

What kind of metadata standard or application profile are you using? Need help? *

0 characters (5000 remaining)

Sounds nice, but…

Encountered problems regarding storage

- Users often don‘t know what they need
→ Which resources? RDS, LinkedData, GitLab etc.?

- Users tend to overprovision
→ Applications for too much storage space (see figure 1)

- What data is worth keeping? Everything?
→ Try to calculate before as accurate as possible

- Users often don‘t know how to transfer data
→ Automization processes useful/needed?
→ Transfer data via web interface, REST-API or S3?

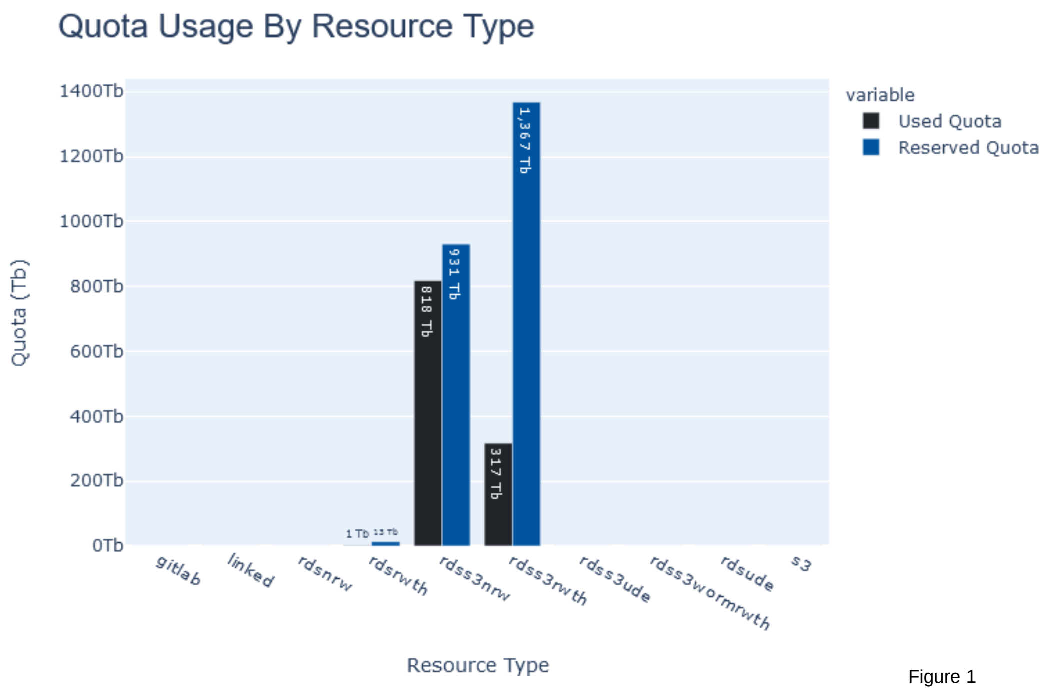


Figure 1

Sounds nice, but…

Encountered problems regarding metadata

- Metadata handling

→ is easily promised but difficult to enforce

→ AIMS platform for application profiles

- What metadata is available, what metadata is useful?

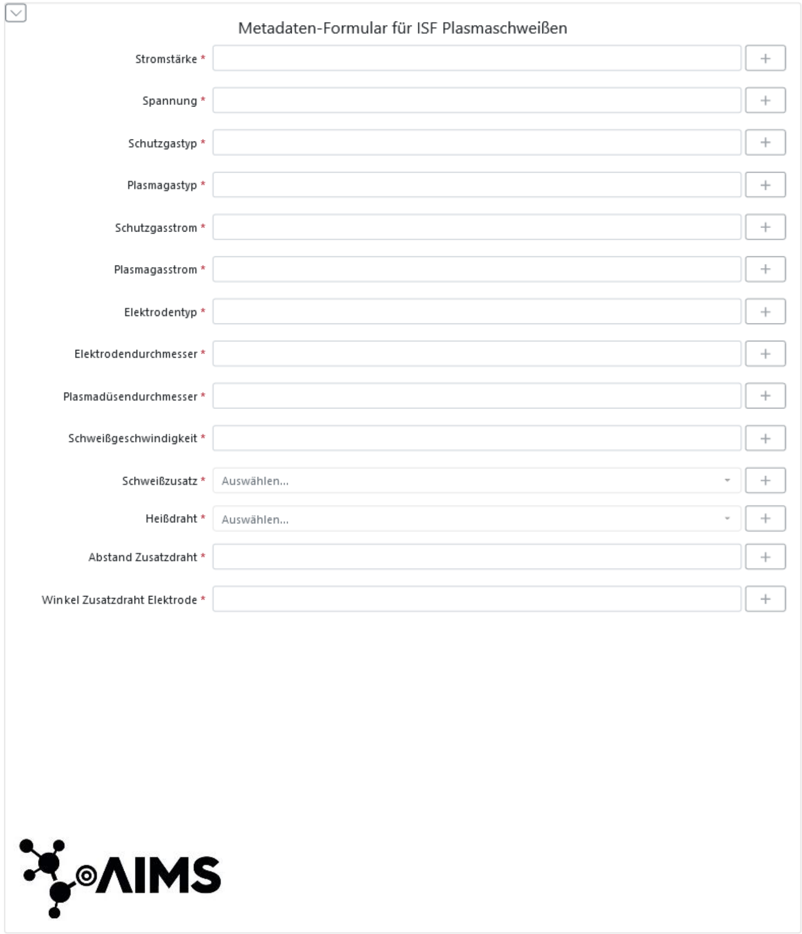
→ save as much as needed and as little as possible

- Can metadata be directly transported from instruments?

→ save as much time/effort as possible

- Should metadata be publicity be available?

→ Enables other researches to find the project (≠ open data!)



Metadaten-Formular für ISF Plasmaschweißen

Stromstärke *

Spannung *

Schutzgastyp *

Plasmagastyp *

Schutzgasstrom *

Plasmagasstrom *

Elektrodentyp *

Elektrodendurchmesser *

Plasmadüsendurchmesser *

Schweißgeschwindigkeit *

Schweißzusatz * Auswählen...

Heißdraht * Auswählen...

Abstand Zusatzdraht *

Winkel Zusatzdraht Elektrode *

Thank you
For your attention!