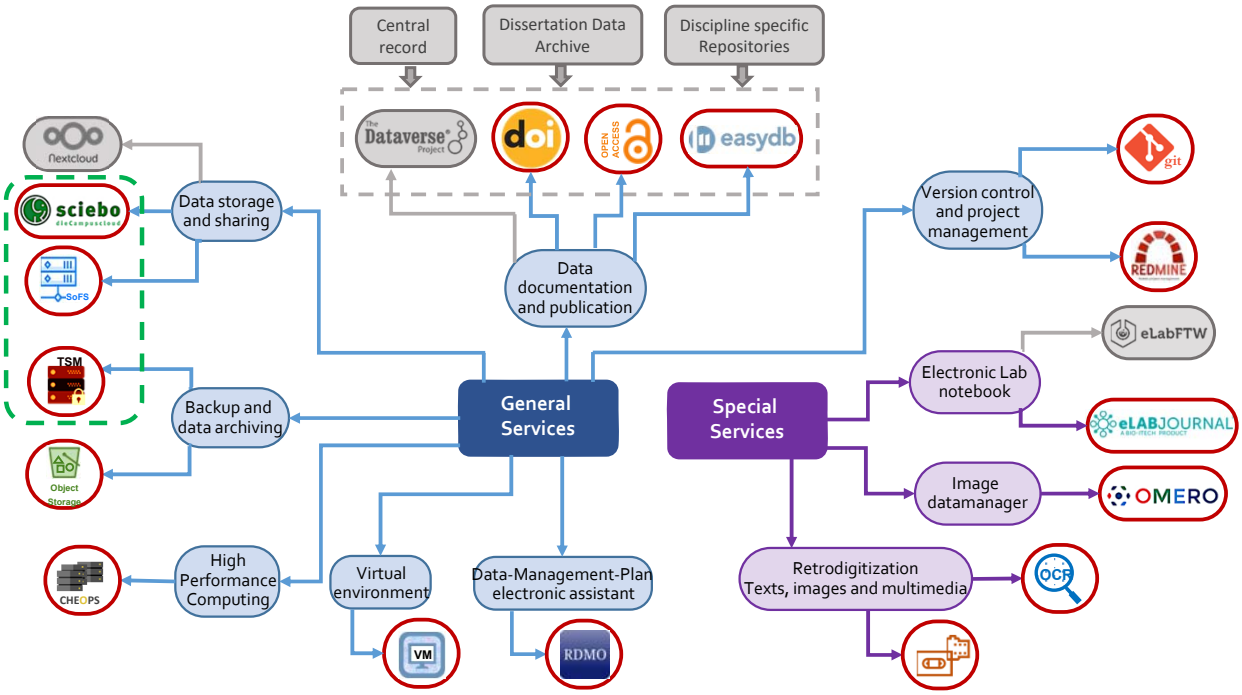
Monica Valencia-S. (RRZK)
Andreas Mühlichen (USB)

C³RDM-Services

CEPLAS | 3rd August 2021

RRZK = Regional Computing Center
USB = University- and City Library


Central record
Dissertation Data Archive
Discipline specific Repositories
Dataverse
doi
OPEN ACCESS
easydb
Nextcloud
sciebo
SoFS
TSM
Object Storage
Data storage and sharing
Backup and data archiving
Data documentation and publication
General Services
Special Services
Version control and project management
git
REDMINE
eLabFTW
Electronic Lab notebook
eLABJOURNAL
Image datamanager
OMERO
Retrodigitization Texts, images and multimedia
OCR
High Performance Computing
CHEOPS
Virtual environment
VM
Data-Management-Plan electronic assistant
RDMO

Guiding Questions on Storage

- How much storage space?
- How long and for what purpose?

- Who needs access? Is synchronisation necessary?
- Which storage solution is suitable for (personal) data?

→ What are necessary precautions to protect against data loss?

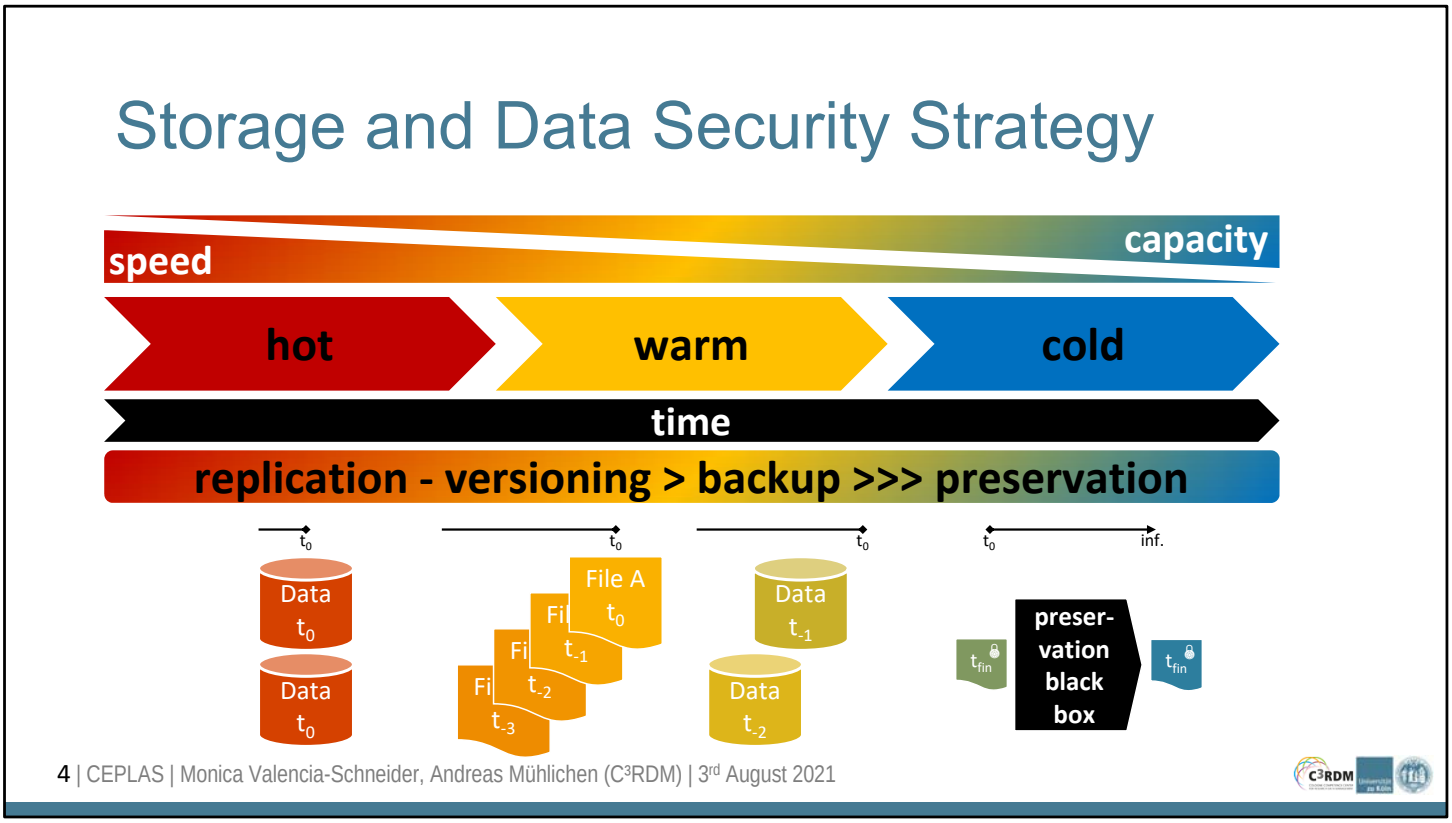
Depending on which phase a project is in and whether hotter or colder data or storage is adequate in the phase, different things have to be considered when planning storage and backup.

Typical questions to be considered and can change over time and with the transitions from hot to warm to cold data:

- **How much** storage space do I need?
- **How long** should the data set be stored (for a specific use - e.g. hot, warm, cold)?

- Are we talking about data sets that multiple researchers from multiple institutions should be able to work on?
- This becomes more complicated when dealing with **personal data**:
    - **Who** needs access?
    - Can this be set up in a **legally secure manner**? Are data and transmission **encrypted**?
    - Does data have to be **deleted** at some point? Is that even possible with the medium?

These considerations must then be incorporated into the planning of which precautionary measures are necessary to protect against data loss in the respective project phase. This needs careful planning, especially because these measures may vary depending on project phase and data usage scenario (i.e. the “temperature” of the data). Hence a data security strategy is needed for all phases of the project and all data usage scenarios.

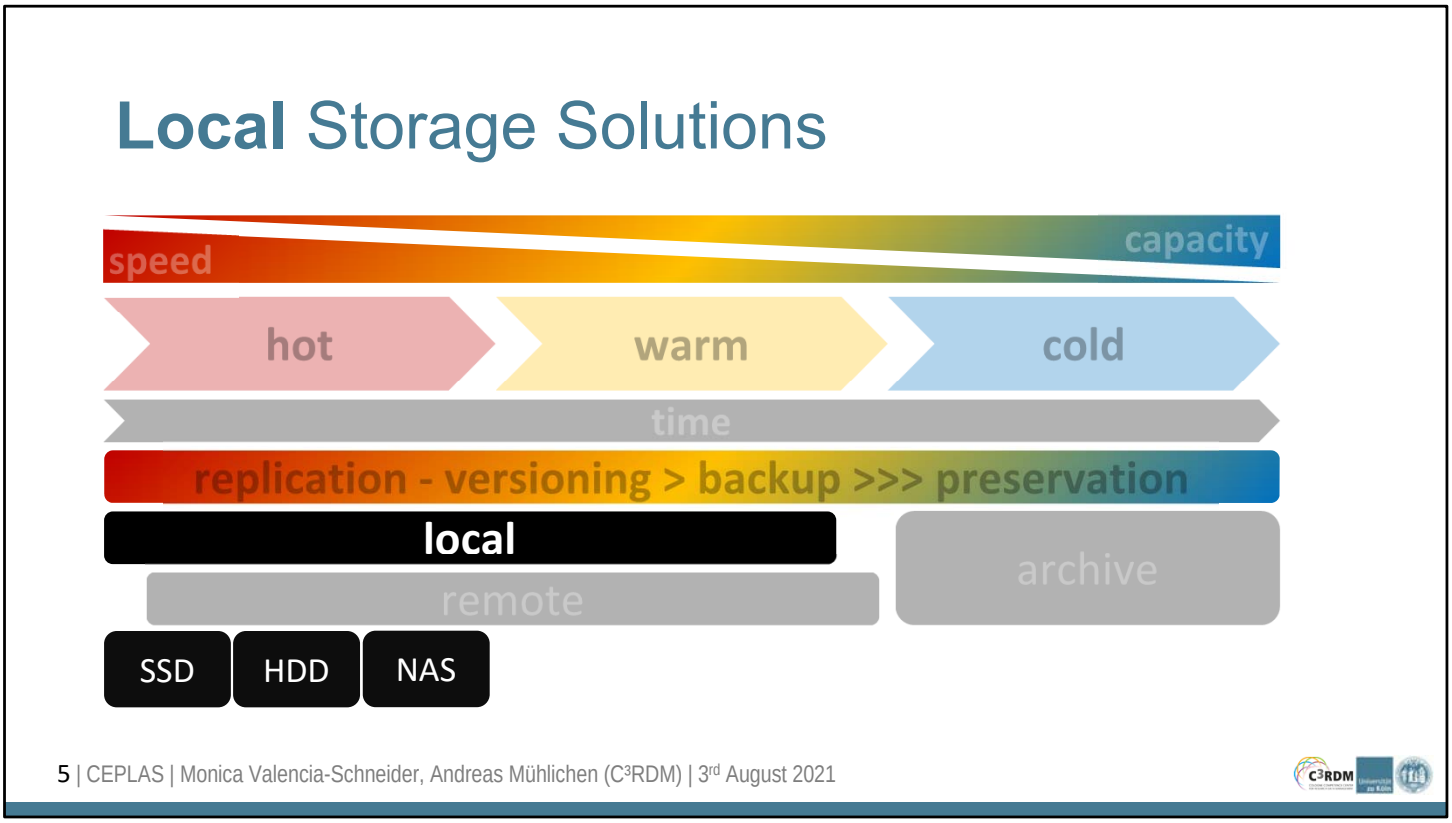


Because your research data can be tremendously valuable (e.g., because of cost or uniqueness), it is vital to ensure the security of your data. In a data security strategy several steps are involved. Somewhat simplified:

- **replication:** creates a duplicate of your data, ideally in real time
  - A typical example would be a RAID 1 (basically meaning a RAID controller writes all data not on one, but two physical drives – creating redundancy by duplication) or a perpetually synchronized cloud storage
  - Advantage: if a drive fails, all data can be recovered – even the most recent changes
  - Disadvantage: an error or malicious software (e.g. ransomware) immediately affects all replicas
- **versioning:** creates copies of different states of changed files, often on-the-fly while working on them
  - Typical examples are git or MS Word AutoSave.
  - Explanation: Saving a changed file overwrites the old one – which is usually irrevocably lost in the process. Discovering an error in the newest file with no way to get back to the old version is a typical cause for data loss.
  - Advantage: Proper versioning provides a history of your work, allowing you to compare your current work with older versions to identify problems or errors.
  - Disadvantage: Badly implemented versioning can lead to clutter and confusion
- **backup:** creates a duplicate of your data at specific points in time, often incorporates at least one decentralized (off site) duplicate
  - Typically, complete copies of your data on external and/or off-site storage media e.g., daily, weekly and monthly. Backup schemes usually start with the complete copy and then incrementally save only changes for the next n backups to save time and storage space – until another cycle begins with a complete copy.
  - Advantage: Almost complete data recovery is possible even after loss of all live systems (e.g., due to electrical surges, fire or malicious software), especially if decentralized. This is also true if the data loss is not immediately discovered.
  - Disadvantage: New data and recent changes that are not yet covered in a recent backup are lost

- **preservation:** saving curated data in specialized storage solutions repositories and data centers that ensure the long-term accessibility and preservation.
    - In the context of research data this is typically provided by repositories. There are also especially secure data storage solutions in data centers, e.g. to ensure data records is kept tamper proof on non-erasable and/or non-modifiable systems.
    - Advantage: Allows for long term archiving of important data
    - Disadvantage: Usually the data needs to be especially prepared and curated to allow for proper storage which is a labor and cost intensive process.

Replication and versioning are usually done live – so they cover even the most recent changes. Backups are done in certain time intervals which are set according to the required data security and the resources available. Replication, versioning and backups are ideally mixed and are important for hot and cold data. Archiving is a process that is usually only employed at the end of a project on a select, curated and final state of the data.



Local Storage Solutions
What kind of storage media are now typically used for local data and what properties do they have that are relevant for data security?

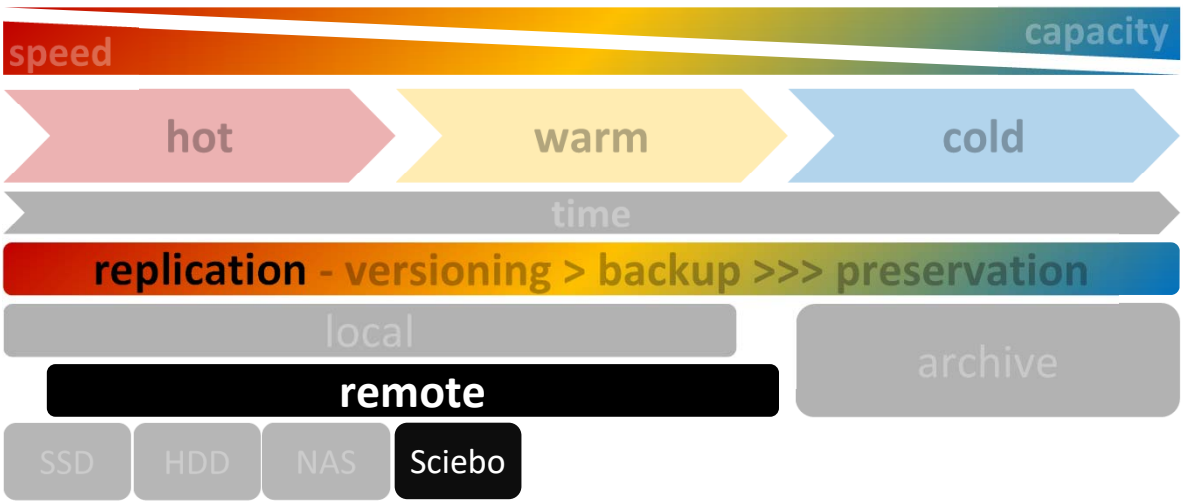
SSD:
- Solid State Disk
- Hard drive with no moving parts
- fast, small size possible
- "Relatively" expensive per TB (2 TB currently typical size for standard sizes)
- technical limitation of the number of read/write cycles, usually no longer relevant in practice
- especially in older models this led to problems with certain encryption methods
- especially old SSDs cannot be securely erased!
- especially old SSDs have to be connected to the power supply from time to time!

HDD:
- Hard disk drive
- conventional hard disk with moving parts (magnetic storage medium)
- significantly slower than SSD, typical designs currently 2.5 (laptop) and 3.5 (desktop) inches
- relatively cheap per TB (12 TB currently approx. 300 €)
- different types for different usage scenarios (e.g. special server disks)

NAS:
- Network attached storage
- **Ideal as your own backup solution**
- File server, i.e. a simplified computer, is specially designed to provide storage space
- HDDs are usually used as storage media in a NAS
- **Attention**: NAS with multiple hard drives can be configured for different scenarios. The rule of thumb is: the safer, the slower or the less storage space. If you configure a NAS as either a high-capacity (a lot of storage space) or a high-performance (in terms of speed) system, you usually have MASSIVE losses in terms of data security (no redundancy, partial loss of all

memory if a medium is defective, etc.) .

Remote Storage Solution: **Sciebo**



First remote storage solution: Sciebo

Remote Storage Solution: **Sciebo**

- NRW-cloud for research and teaching
- replication
- students 30 GB, employees 500 GB, projects 2 TB
- synchronisation of multiple devices
- sharing and collaborative real-time editing

BUT

- **UN**encrypted (no personal data allowed)
- **NOT** for backup
- **NOT** for permanent data storage

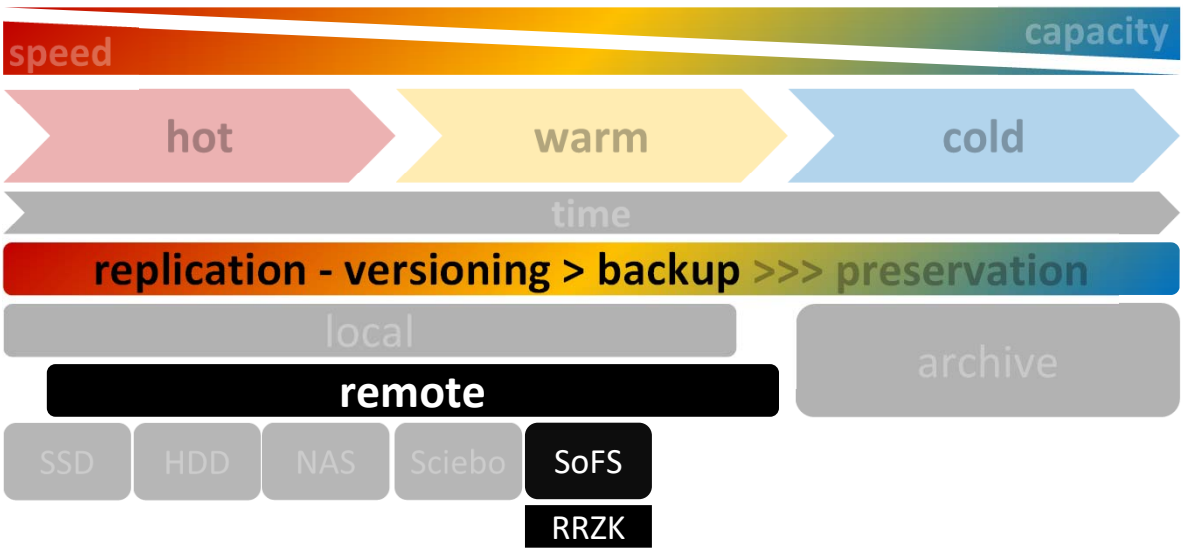
Sciebo
- non-commercial cloud for science
- 22 Unis, NRW-Supported, main storage centre is in process of relocation to Münster, data currently stored in Bonn
- can be synchronized on different devices via desktop client or can be used via browser
- with Sciebo, folders and files can be made available to other employees of the University of Cologne as well as to external parties.
- Students: 30 GB
- Employees: 30 GB, can be extended to 500 GB, can request project box (up to 2 TB)
- [Originally, employees with storage requirements of over 500 GB could apply for a project box intended for work or institute groups. This is apparently currently suspended because the system is reaching its capacity limits and an expansion is pending.]

Advantages:
- a good alternative to commercial providers, as the data storage is located in Germany
- more storage capacity than SoFS
- data can be synchronized
- data can also be easily shared with external parties and can be edited together in the browser

Disadvantages:
- data storage is not located in the RRZK (storage location is Münster)
- no automatic backup
- no clearly defined redundancy (i.e., it is not clear how securely the data is replicated internally on redundant systems - it cannot be guaranteed that a system failure at Sciebo will not lead to permanent data loss)
- it is therefore not suitable for permanent data storage
- saving **personal data** (data that is GDPR sensitive) it **explicitly prohibited**!

General info on sciebo: https://rrzk.uni-koeln.de/en/data-storage-and-share/sciebo
Register (how-to): https://rrzk.uni-koeln.de/en/data-storage-and-share/sciebo/registration
Upgrade to 500 GB (how-to, employees only): https://rrzk.uni-koeln.de/en/data-storage-and-share/sciebo/new-functions
Project box (how-to, employees only): https://rrzk.uni-koeln.de/en/data-storage-and-share/sciebo/project-boxes
Sciebo register (students and employees): https://sns-sp.sciebo.de/secure/de.php

Remote Storage Solution: **SoFS**



Second remote storage solution: SoFS

Remote Storage Solution: **SoFS**

Scale out FileServer

- RRZK, online NAS system
- main with replication, versioning, backup (via TSM)
- personal use max. 10 GB, facilities 1TB+
- versioning of older file versions (h+/d/w)
- sharing options (guest accounts possible)
- UKLAN/VPN/WebDAV

BUT

- **UN**encrypted storage
- versioning only directly accessible via Windows

SoFS RRZK

Capacity
- personal: 10 GB
- institution: 1TB free per cost center. More is possible but additional storage has to be acquired by the institution. The storage per cost center can be split up using permissions management.

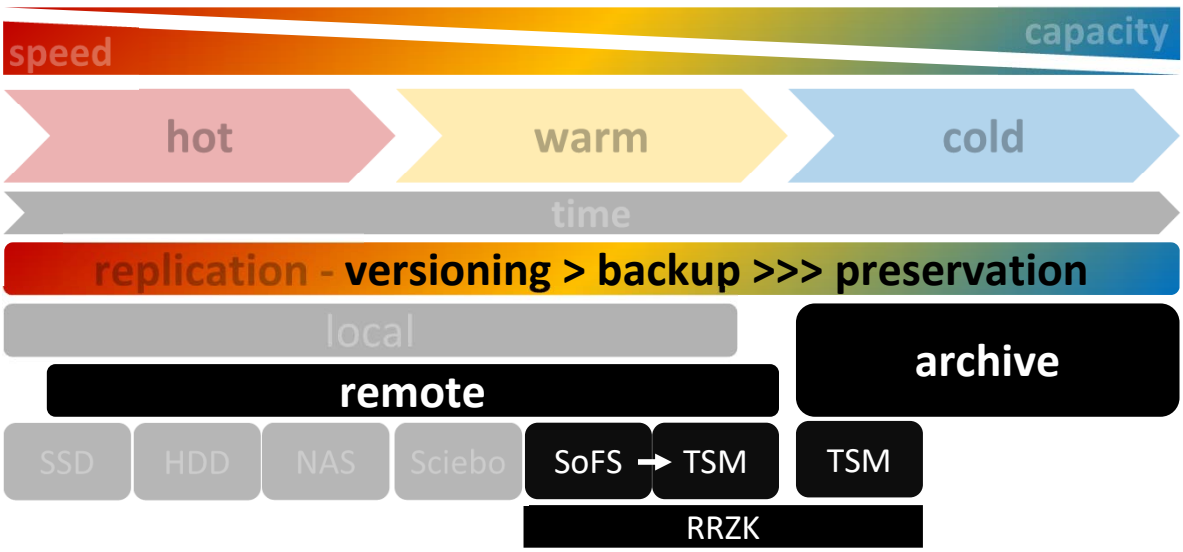
Versioning (and redundancy)
- snapshots at 8:00, 12:00, 16:00 and 20:00 automatically, overwritten the next day of changed files
- daily (at midnight, kept for a week)
- weekly (backup at midnight Sunday to Monday - of which a total of 30 backups are kept)
- In addition: Internal replication (redundancy). Redundancy means: There is not just one copy in the RRZK but several (at least 2) versions that are stored on different systems.
- SoFS is automatically backed up in the TSM (i.e., there is an automatic backup) - details on the TSM will follow soon).
- Versioning is directly accessible with windows. Access to the versioning outside of windows is only possible via the helpdesk.

Sharing options
- possible on the personal as well as on the institutional level
- guest accounts have to be applied for, renewal is required after one year

Encryption
- data is stored unencrypted in the RRZK, manual encryption of personal or sensitive data necessary (e.g. VeraCrypt https://www.veracrypt.fr/en/Home.html)
- data transmission is encrypted with VPN and WebDAV

SoFS overview: https://rrzk.uni-koeln.de/en/data-storage-and-share/online-storage-sofs
Info on SoFS versioning: https://rrzk.uni-koeln.de/en/data-storage-and-share/online-storage-sofs/features
Access instructions (Windows): https://rrzk.uni-koeln.de/en/daten-speichern-und-teilen/online-speicher-sofs/anleitungen/windows
Access to versioning (how-to): https://rrzk.uni-koeln.de/en/data-storage-and-share/online-storage-sofs/instructions/access-to-previous-versions
SoFS info for institutions: https://rrzk.uni-koeln.de/en/data-storage-and-share/online-storage-sofs/institutions

Remote & Archive Solution: **TSM**



TSM provides two distinct functionalities: a backup and an archive function.

Remote & Archive Solution: **TSM**

- RRZK TSM (Tivoli Storage Manager)
- **backup:** SoFS, some institute servers, computer (2 versions and replication)
- **archive:** permanent (with replication)
- UKLAN/VPN/WebDAV
- not encrypted but secured

BUT

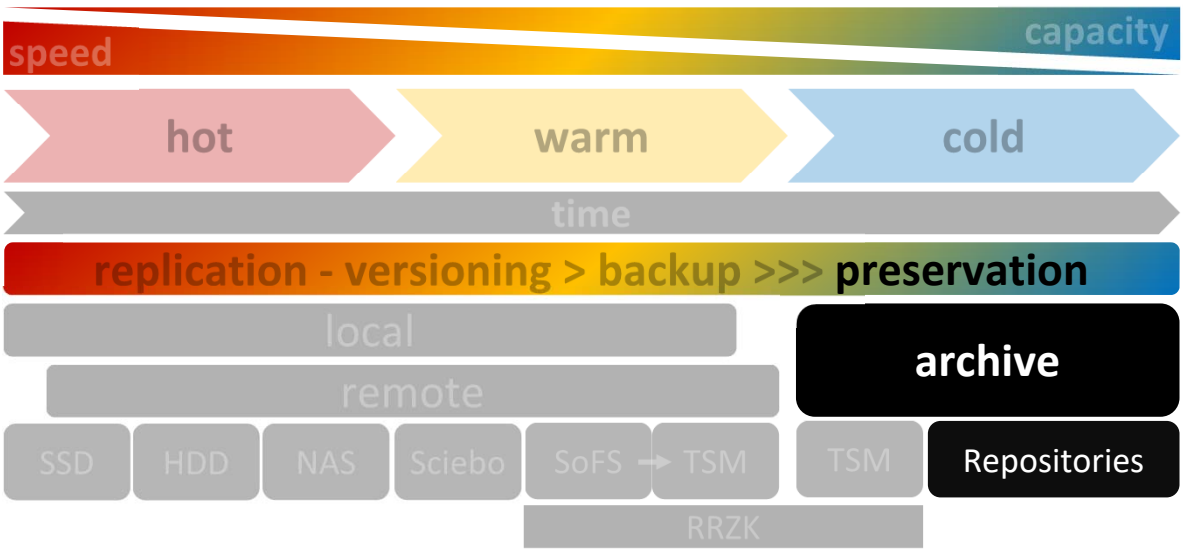
- application, client and configuration necessary
- archive non-erasable (by user) (actually: it's a feature)

TSM

- is a backup and archiving solution provided by the RRZK
- SoFS is automatically backed up in TSM (note: active data only - no versioning). Cycle is approx. once/week, 2 states are saved.
- Institute servers are usually backed up automatically in the TSM (ask your local IT support if in doubt)
- Data stored on local/personal devices can be stored in TSM if generated as part of teaching/research tasks. A login is required for this purpose. Additionally, client software must be installed and configured.
- TSM as a backup: There is a minimal versioning (2 backups are kept, backup cycle is dependent on service and user requests, usually data is dropped after 60 days in SoFS)
- TSM as an archive: The data is stored indefinitely, regardless of what happens to the local original files.
- There are no general capacity limits (but storage requirements exceeding several TB may incur costs).
- access to the backup via help desk
- tape storage
- permanent storage: theoretically infinite

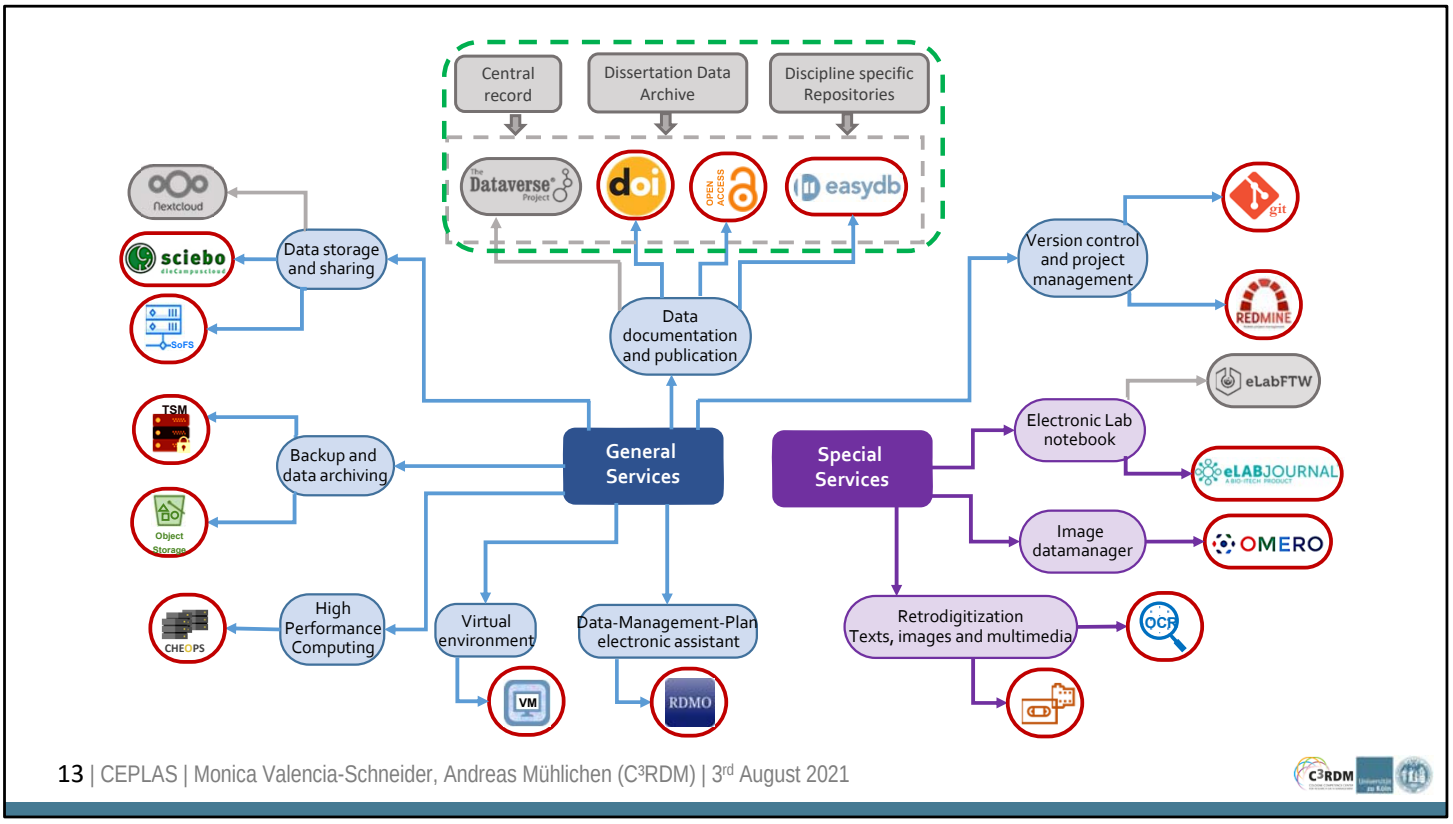
TSM general info: https://rrzk.uni-koeln.de/en/data-storage-and-share/backup-system-tsm
TSM account application (DE only): https://rrzk.uni-koeln.de/daten-speichern-teilen/backup-system-tsm/tsm-registrierung
TSM instructions (DE only): https://rrzk.uni-koeln.de/en/daten-speichern-teilen/backup-system-tsm/tsm-anleitungen
TSM client software: https://rrzk.uni-koeln.de/en/daten-speichern-teilen/backup-system-tsm/tsm-client-software

Archive Solution: **Repository**



In a nutshell:

- A repository is a storage solution provided by a trusted institution (e.g. university, publisher) to store research data for medium (max. 10 years) or long term (10 years +).
- They are commonly either generic, discipline specific or instutitional
- In addition, data is stored and offered through the repository for use by third parties.
- If data may not be used by third parties or must be deleted after a certain time (e.g., due to DSGVO), access to the data itself may be restricted or the data may be deleted, while a record (i.e. the description of the data using metadata) that the data exists or existed is retained permanently.
- A Research Data Storage (RDS) is in the process of being installed, basis is object store, start of regular operation is expected mid. 2022.
- implementation of repository solutions is in process



- A Research Data Storage (RDS) is in the process of being installed, basis is object store, start of regular operation is expected mid. 2022.
- implementation of repository solutions is in process

Preservation of Data – Requirements

- DFG Guidelines for Safeguarding Good Research Practice
- UoC Guideline on the Handling of Research Data 2018
- Doctoral Degree Regulations of Math.-Nat. Faculty (2020) [DE only]

  §7 (9): [1]**If** the dissertation involves the acquisition of primary **data** or the analysis of such data, or if the reproducibility of the results presented in the dissertation requires the availability of data analyses, experimental protocols, or sample materials, the dissertation shall describe **how such data and materials are secured and accessible**. […] [own translation]

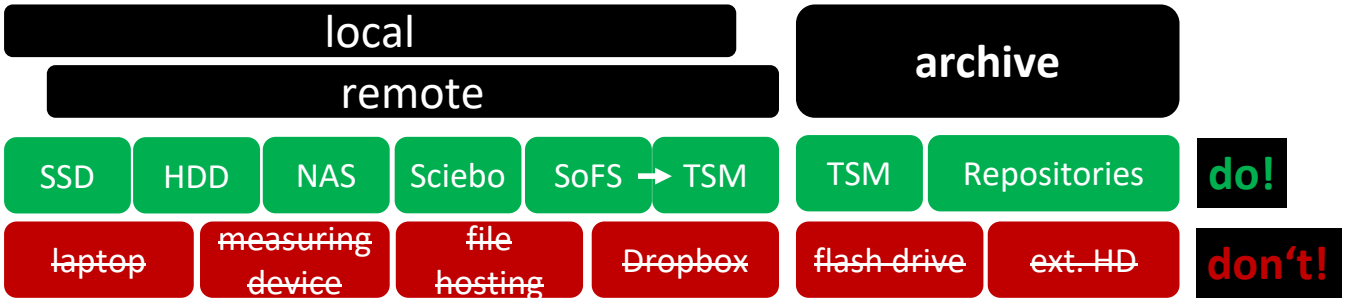
- DFG: Advocates to hold available data for usually at least 10 years
- UoC Guidelines: Advocate for RDM and DMP, supports by supplying infrastructure and advice for RDM
- PhD Regulations Natural Sciences: Implements measures to ensure compliance with Good Research Practice policy

---

LINKS

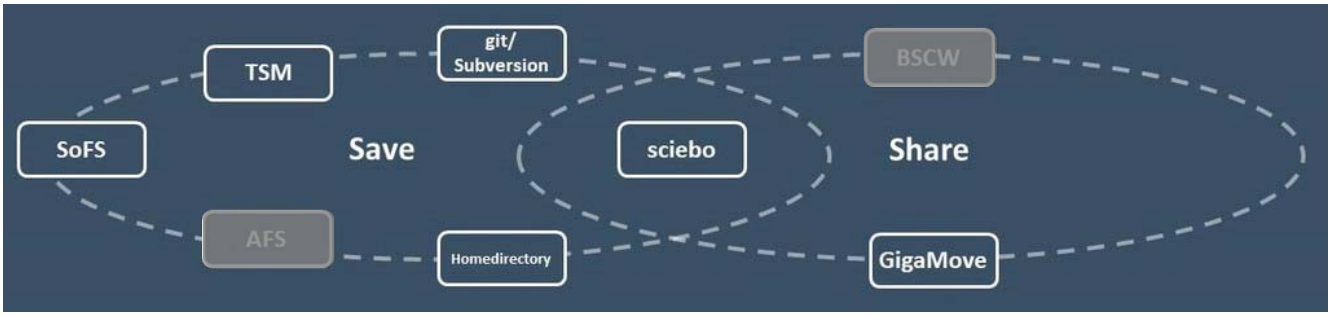
- UoC Guideline: https://fdm.uni-koeln.de/en/rdm-guideline
- Doctoral Degree Regulations: https://mathnat.uni-koeln.de/sites/dekanat/official/Ordnungen/Promotionsordnung_2020.pdf
- DFG Guidelines: https://zenodo.org/record/3923602

C³RDM Storage **Dos** and **Don‘ts**



Sharehosters (or One-Click-Hosters) are filehosting services without login like MEGA (files are usually uploaded and are available for download under a more or less arbitrary link - which is at best providing a completely inadequate security solution called security by obscurity, therefore please DO NOT use!)

Overview Storage Solutions RRZK



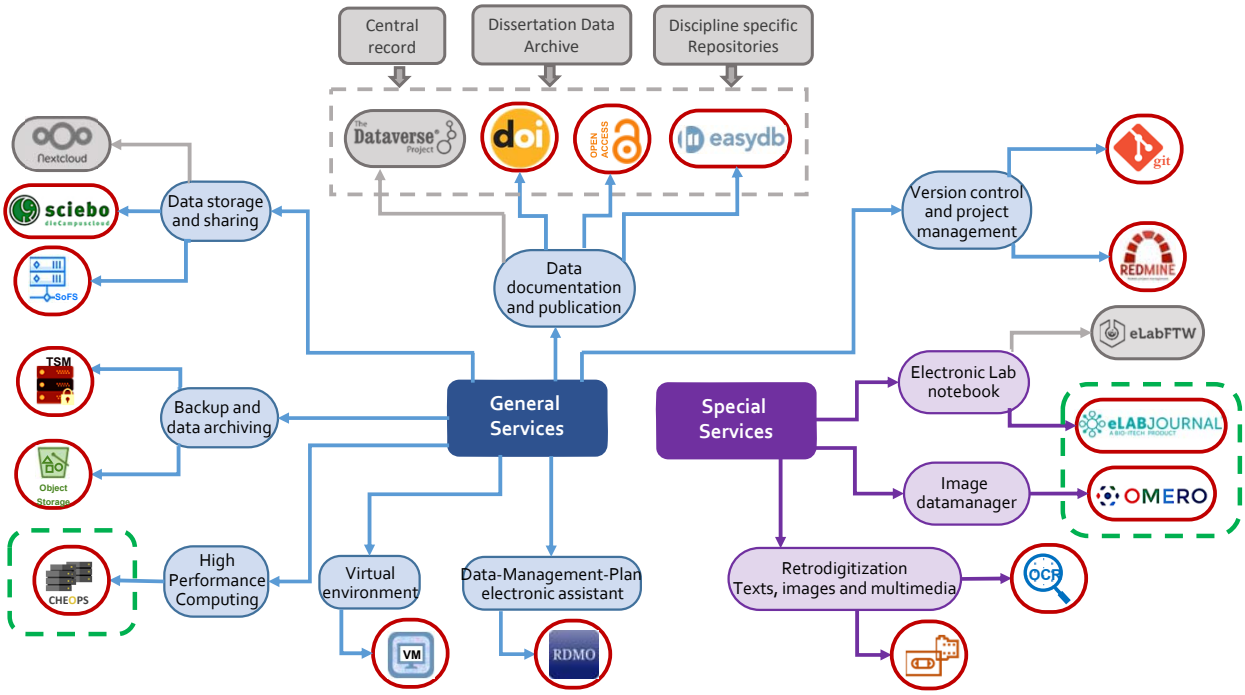
Useful links

- Overview: https://rrzk.uni-koeln.de/en/data-storage-and-share
- sciebo Cloud
- SoFS Speicher
- Application for use of the Tivoli Storage Manager (TSM) (DE only)

See the RRZK website for an overview of software solutions and services for storing and sharing data.

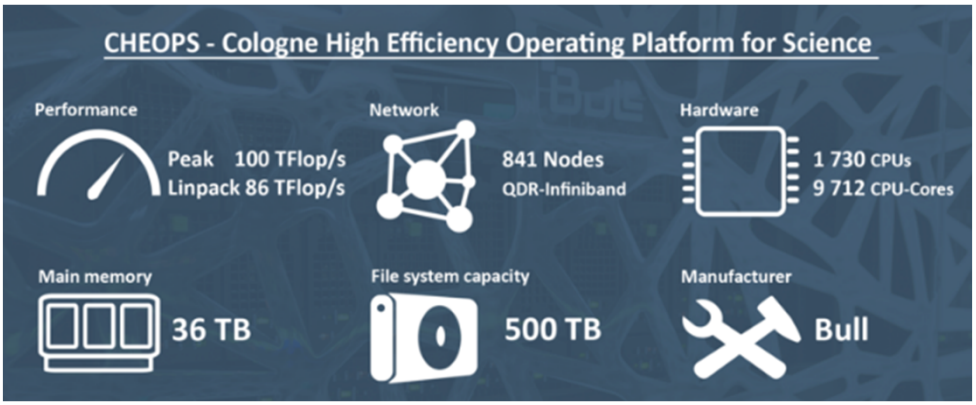
AFS (Andrew File System) und **BSCW** (Basic Support for Cooperative Work) **are being phased out.**

- **Overview:** https://rrzk.uni-koeln.de/en/data-storage-and-share
- **sciebo Cloud:** https://rrzk.uni-koeln.de/en/data-storage-and-share/sciebo
- **SoFS Speicher:** https://rrzk.uni-koeln.de/en/data-storage-and-share/online-storage-sofs
- **Application for the use of Tivoli Storage Manager (TSM):** https://rrzk.uni-koeln.de/daten-speichern-teilen/backup-system-tsm/tsm-registrierung (DE only)


Central record
Dissertation Data Archive
Discipline specific Repositories
Dataverse
doi
OPEN ACCESS
easydb
Nextcloud
sciebo
SoFS
Data storage and sharing
Data documentation and publication
Version control and project management
git
REDMINE
eLabFTW
TSM
Object Storage
Backup and data archiving
General Services
Special Services
Electronic Lab notebook
eLABJOURNAL
Image datamanager
OMERO
CHEOPS
High Performance Computing
Virtual environment
VM
Data-Management-Plan electronic assistant
RDMO
Retrodigitization Texts, images and multimedia
OCR

High Performance Computing

CHEOPS



High Performance Computing - CHEOPS

Structure of the file system

File system/ location	Total Capacity	Quota per user	Speed	Backup/ Archive
/home	24 TB	100 GB 100 000 files	Low	Daily Backup
/projects	1300 TB	As requested	Medium-low	Archive must be requested (elsewise none)
/scratch	404 TB	10 000 000 files	High	None (data is deleted after 30 days)

High Performance Computing – CHEOPS

How can I get access?

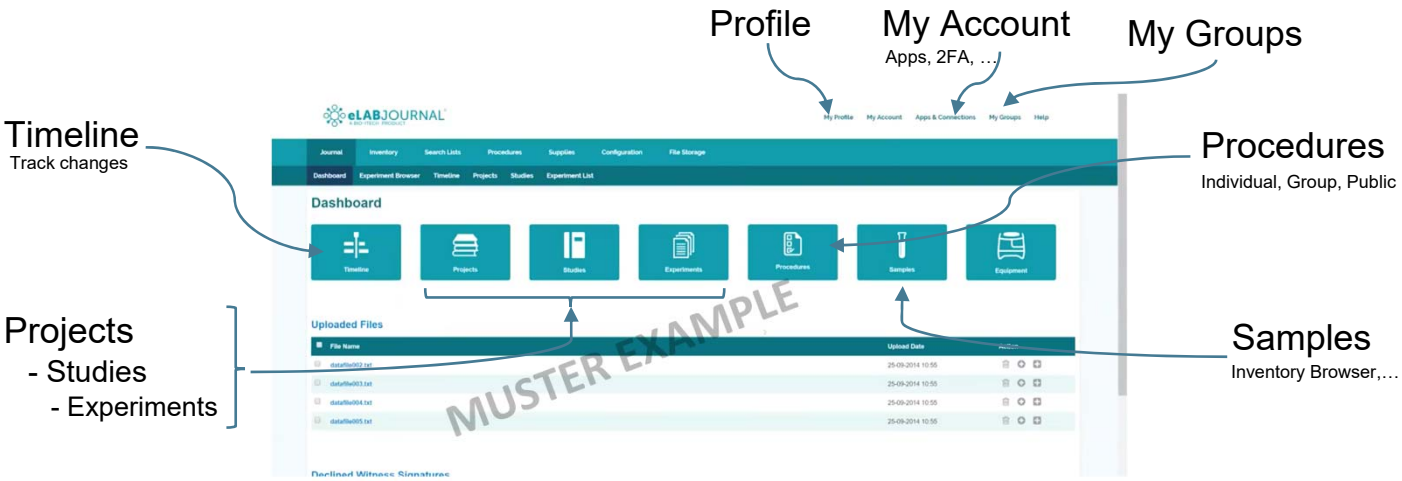
- UoC account (external users: guest account application form)
- and HPC authorization (everyone: application form)

Type of authorization set up

- Trial Account (max 1000 CPU hours)
- HPC-Authorization (provide separate description of the project)

CHEOPS
Information to get access
https://rrzk.uni-koeln.de/en/hpc-projects/hpc/access-and-use-instructions
Guest account form for external users:
https://rrzk.uni-koeln.de/fileadmin/sites/rrzk/Account_Kommunikation/Accounts/application-guest-account.pdf
Application form for HPC authorization:
https://rrzk.uni-koeln.de/sites/rrzk/Account_Kommunikation/Accounts/HPC-Berechtigungsantrag_Englisch.pdf

Electronic Lab Notebook – **eLAB**JOURNAL



Electronic Lab Notebook - **eLAB**JOURNAL

Is it for me/our team?

- Within UKLAN (VPN) watch explanatory video
- Test account (3 per group for 2-3 weeks)

How to get access?

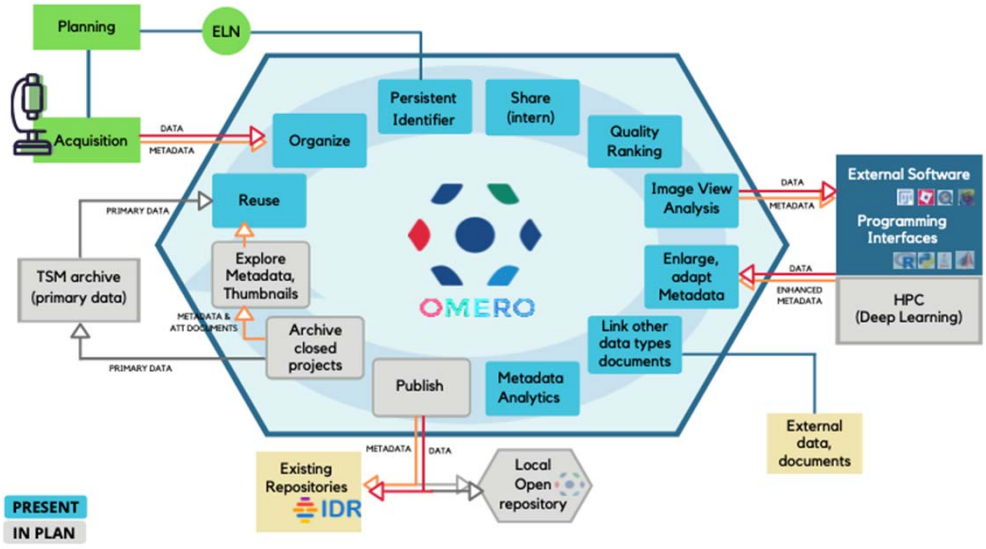
- Requisite: UoC account
- Email to ELN admins with CC to gral. Manager Biology Department indicating:
  Institute, AG (or Lab), group admin name and uni-account, number of needed licences
- In Lectures: please contact ELN admins eln-keyuser@uni-koeln.de

eLABJOURNAL
explanatory video:
https://elj.uni-koeln.de/videotutorial/Introduction%20to%20eLABJournal%20for%20end%20users%20@Uni-Koeln%20(I)-20190902%200802-1.mp4
Get access:
Email to eln-keyuser_at_uni-koeln.de with CC to stefanie.zeretzke_at_uni-koeln.de

Managing Imaging Data – OMERO

Savadel & Bass (2017) doi:10.1080/15592324.2017.1311437

Managing Imaging Data – OMERO



Thank you for listening!

Questions?
Please do not hesitate to contact us via Email or RocketChat:

Monica Valencia-Schneider (RRZK)
mvalenci@uni-koeln.de
RC: mvalenci

Andreas Mühlichen (USB)
muehlichen@ub.uni-koeln.de
RC: amuehli2

C³RDM

https://fdm.uni-koeln.de
fdm-support@uni-koeln.de