MODAMA

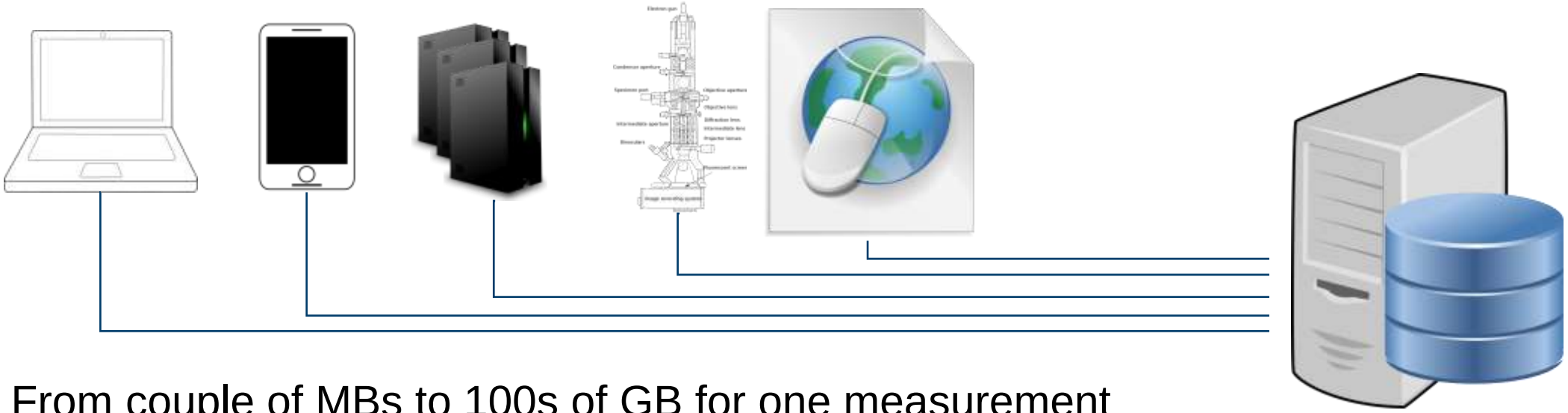
Modular Data Management System at ER-C

- 2025-08-26 I DIETER WEBER, ALEXANDER CLAUSEN, RITWIK SHANKER

DATA AT ER-C: COMPLEX USE CASE

Only partially solved in other set-ups

- People oversee their own data, different from other large facilities.
- External people participate, this is at the heart of a user center like ER-C!
- Combination of mobile devices, HPC, web applications



- From couple of MBs to 100s of GB for one measurement
- Very high throughput: GB/s processing speed even on ultrabook laptops!

https://openclipart.org/
https://commons.wikimedia.org/wiki/File:Scheme_TEM_en.svg

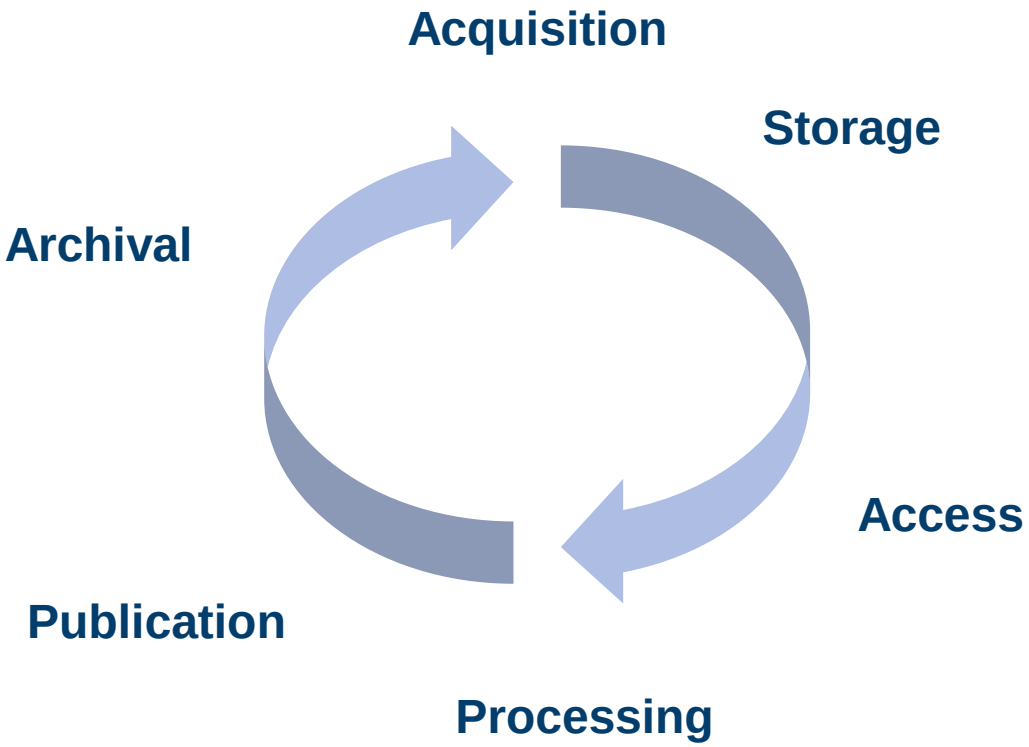
USER BASE AND APPLICATIONS

- Very diverse, incl short term visitors
  - Quick onboarding!
- From IT professionals to scientist without any IT background
- Full spectrum of EM methods and applications
- Method development: One-of-a-kind tools and projects
- Integrate with instruments and their peculiarities
  - TEMs have complex and individual IT systems, usually legacy systems
- Base line: external hard drives
- NAS and Nextcloud instances as well

Goals

- Seamless sharing and collaboration
- Support “big data” projects
- Data protection, Good Scientific Practice
- FAIR data (Findable, Accessible, Interoperable, Reusable)
- Support scientists, don’t get in their way
- Smooth transition

DATA LIFE CYCLE



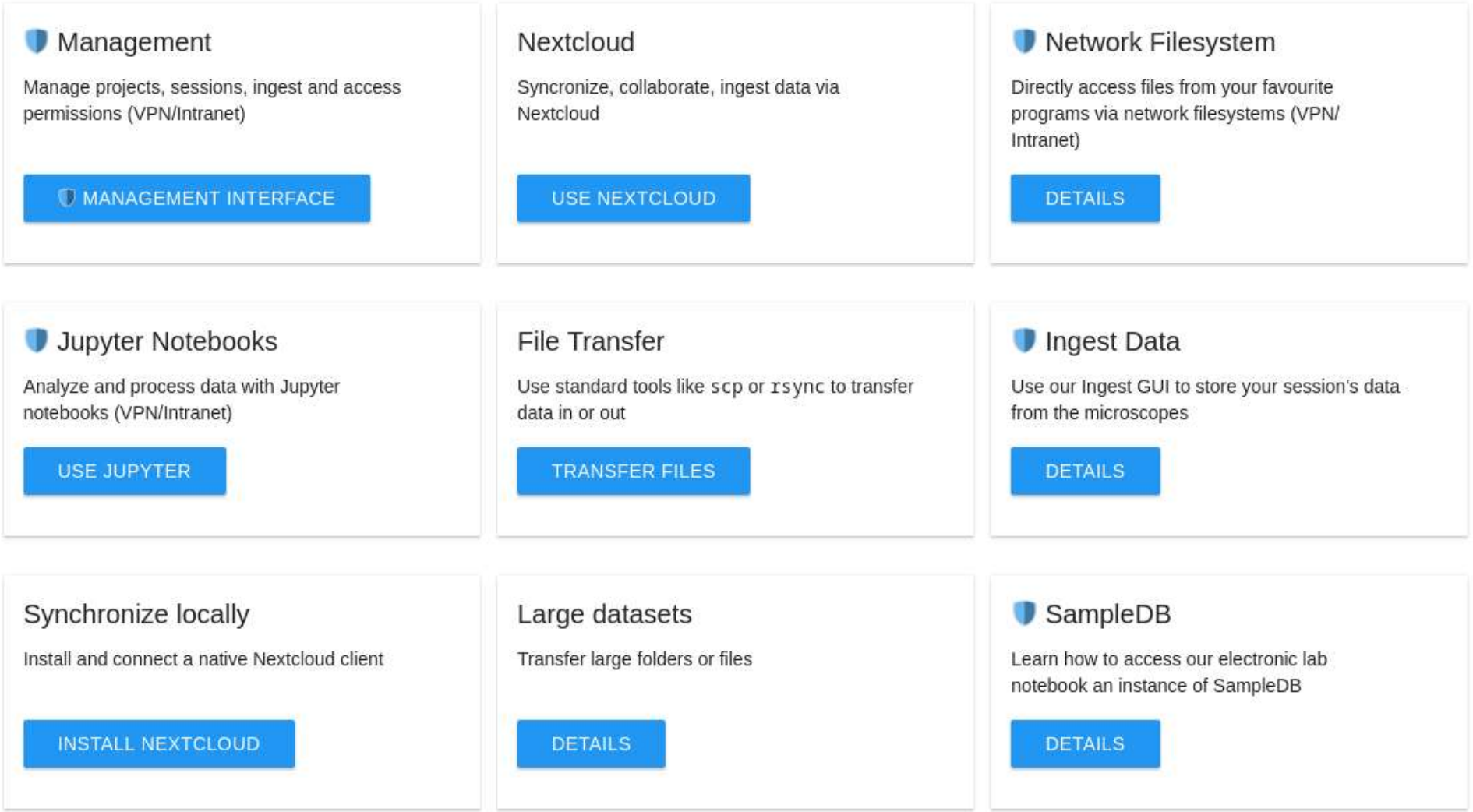
DESIGN OF MODAMA

Combine NAS with Nextcloud, additional services

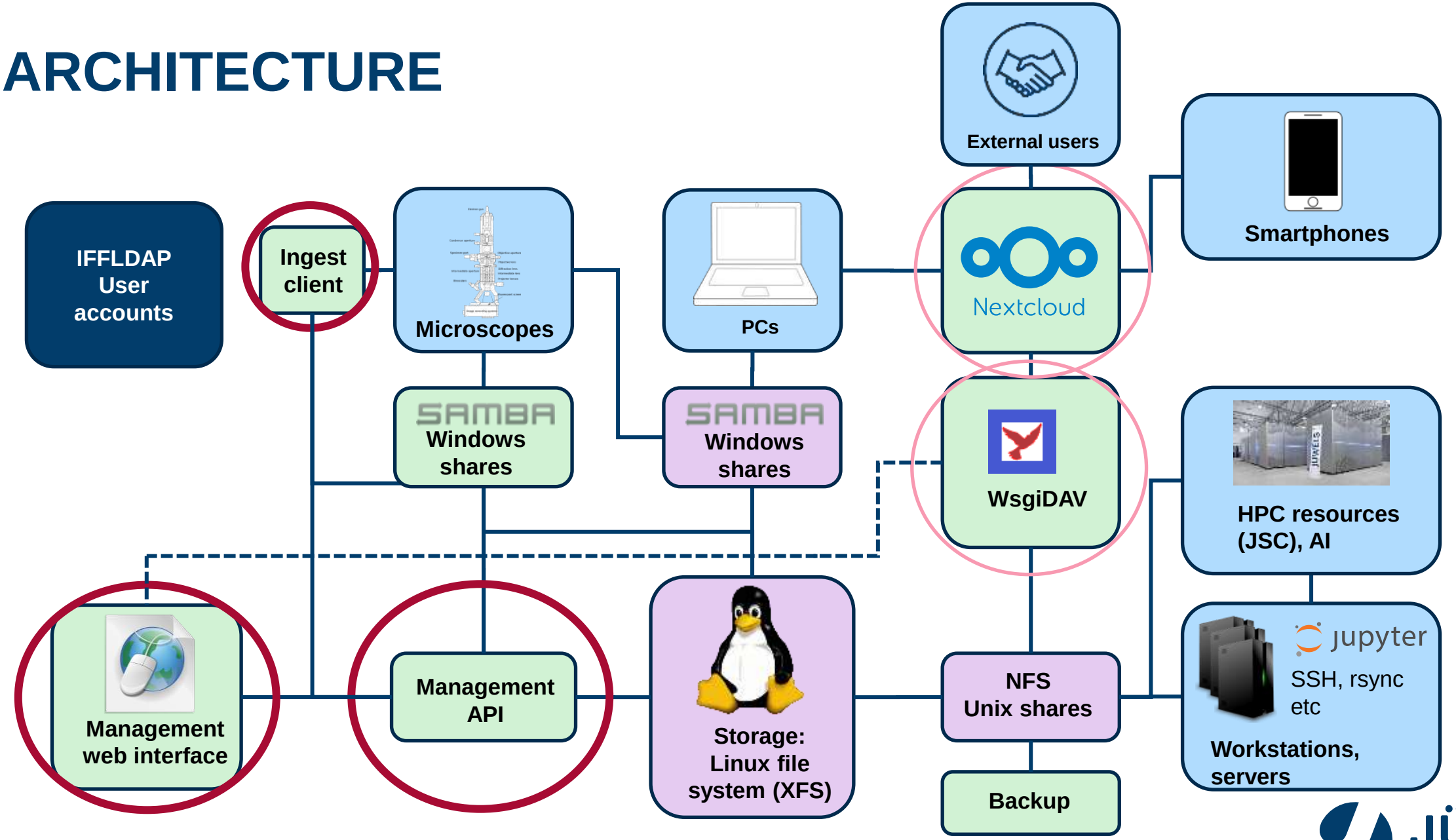
https://er-c-data.fz-juelich.de/

- Central file system
- Exported via CIFS (Samba), NFS and Nextcloud
- Available on shared workstations
- Self-service web interface to manage projects
- Web-based browse interface to link with other services
  - Electronic Lab Notebooks (ELNs), data viewers, Jupyter notebooks, …

LANDING PAGE



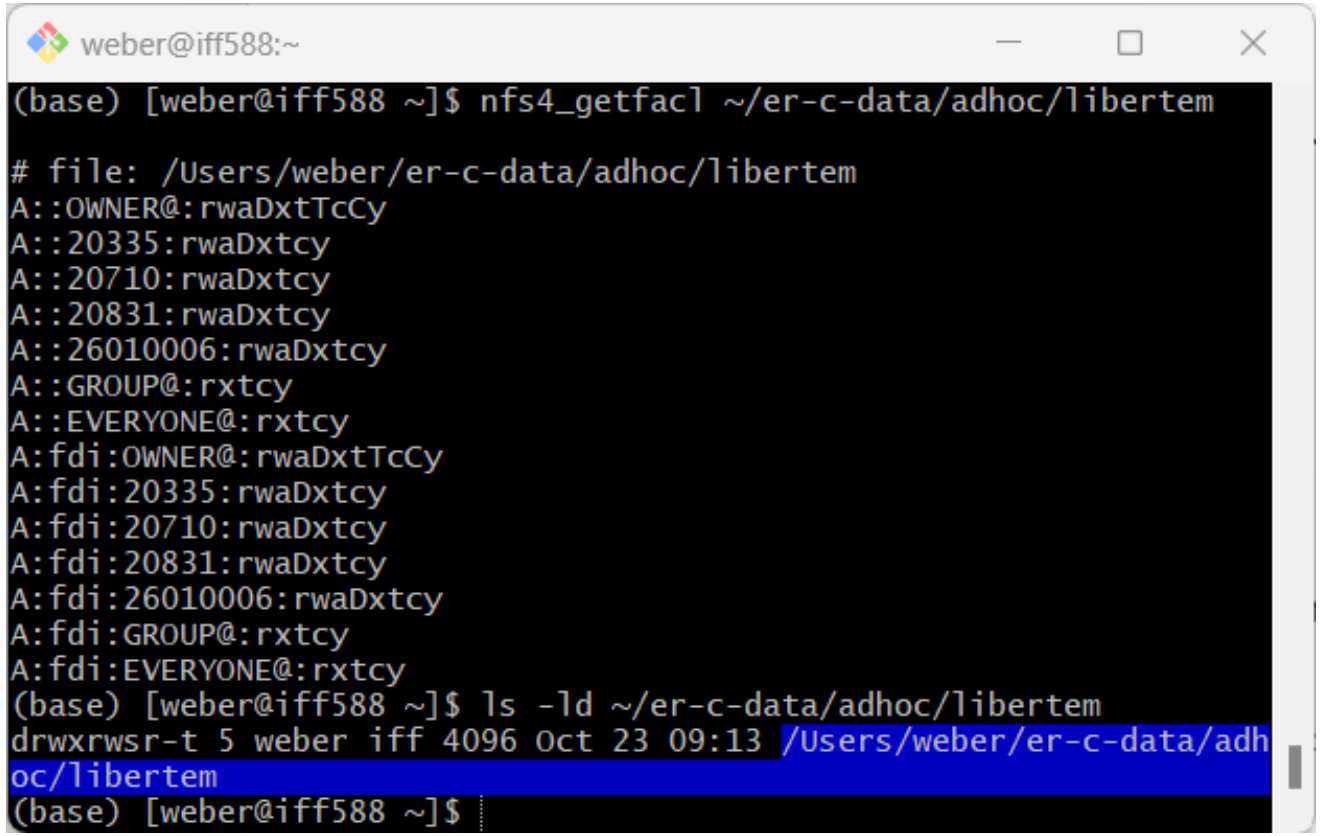
ARCHITECTURE



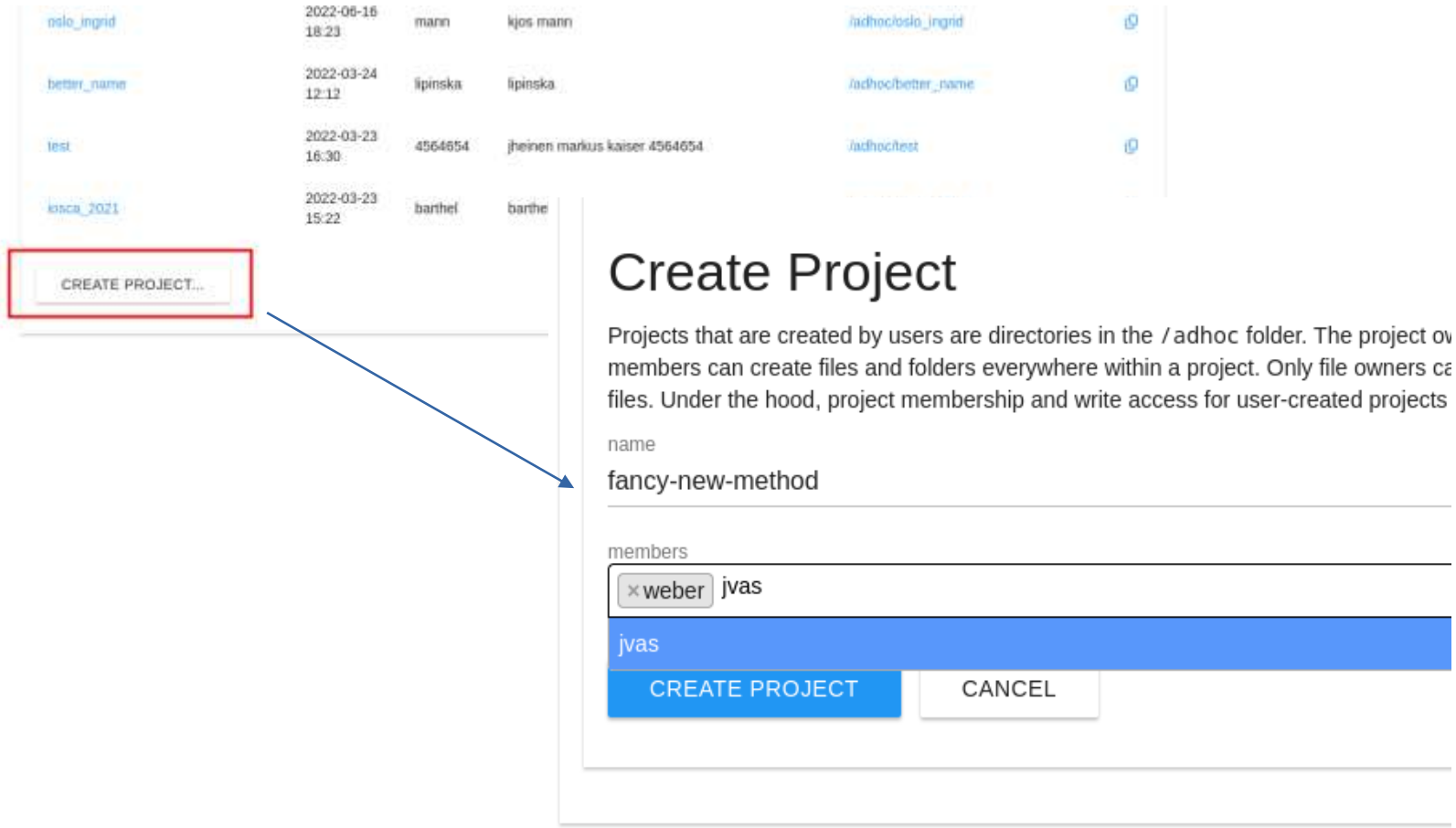
JÜLICH Forschungszentrum

CORE: LINUX FILE SYSTEM

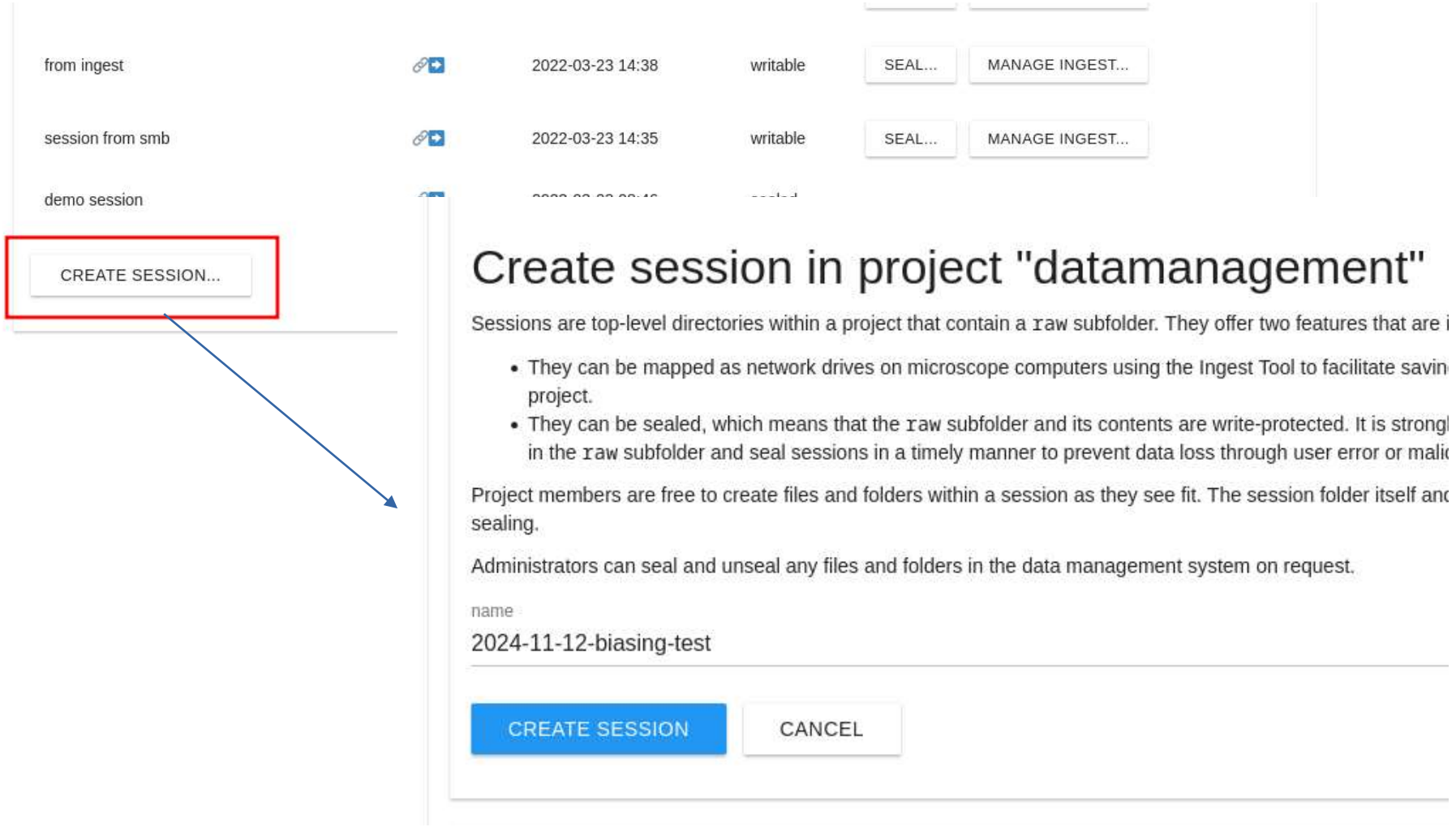
- Full range of FS features available
  - Direct I/O, random access, FS cache, memory mapping, …
- Easier to export FS content as blobs than vice-versa
- File system permissions, ACLs, immutable flag, inotify to create desired behavior
- “Truth is in the file system”



The management interface - create and update projects



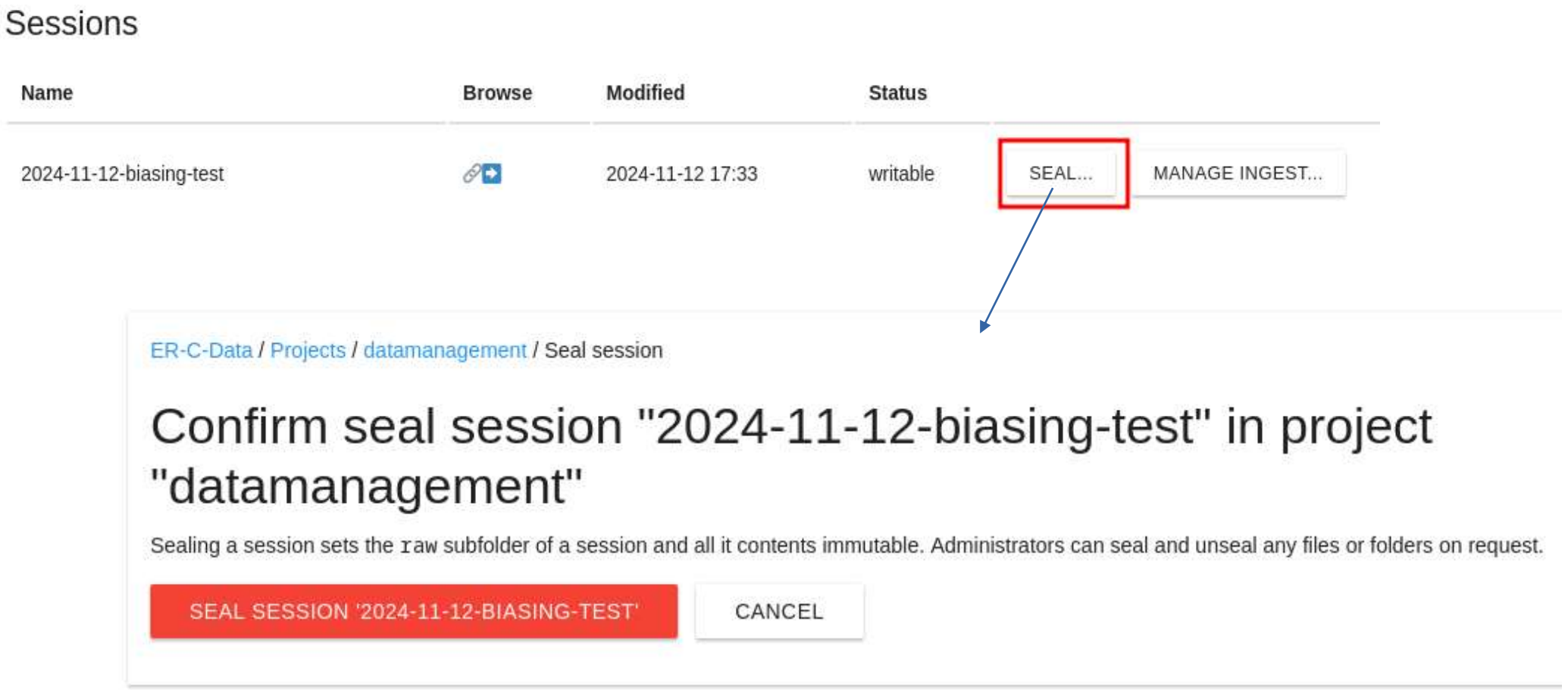
The management interface – create session



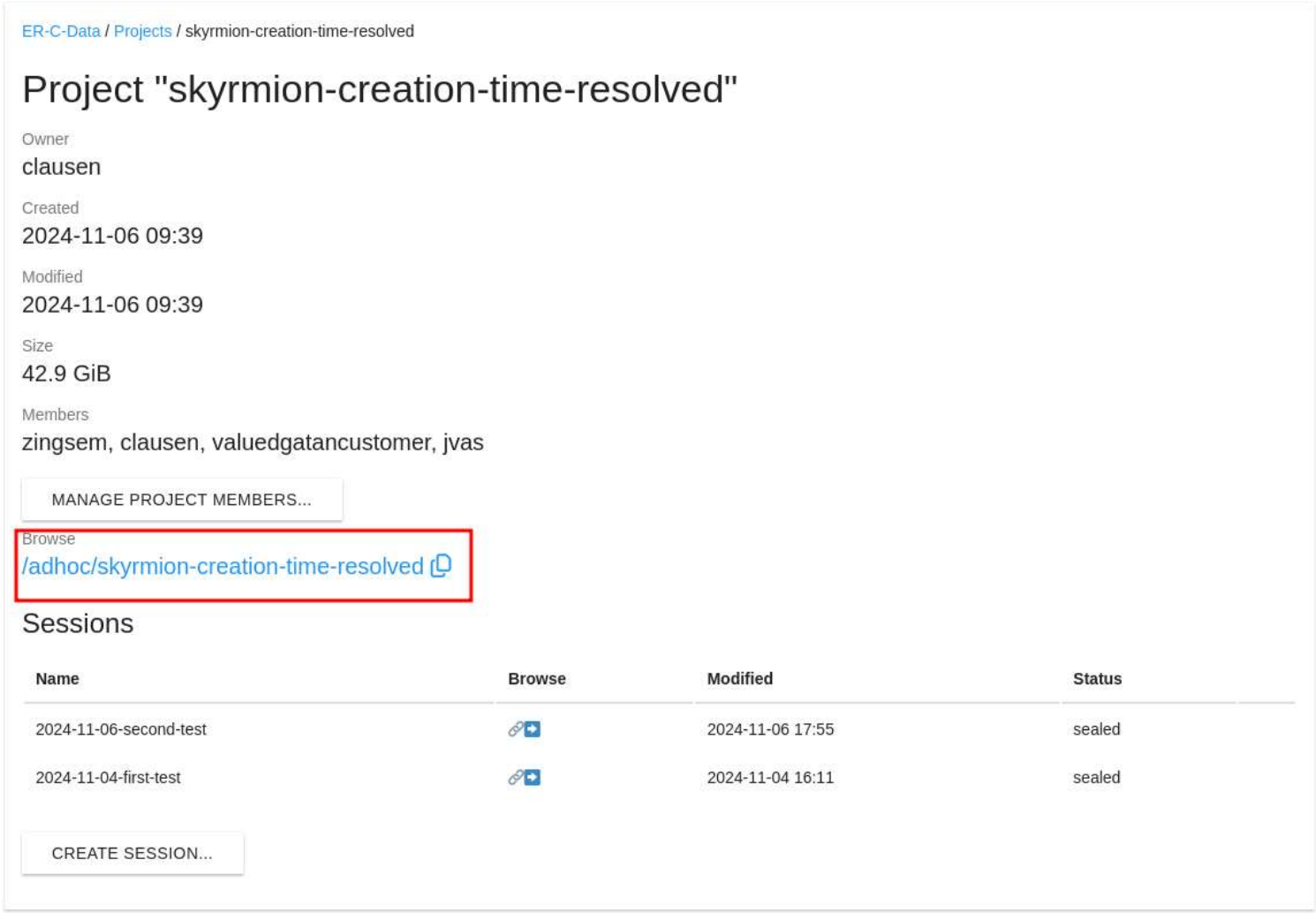
The management interface – overall folder structure

- /storage/er-c-data/adhoc/<project>/<session>/
- Session contains a “raw” subdirectory
  - This should contain the raw experimental/simulation data
  - It can be sealed to write-protect it
  - Good for integrity: can’t (accidentally) change data!

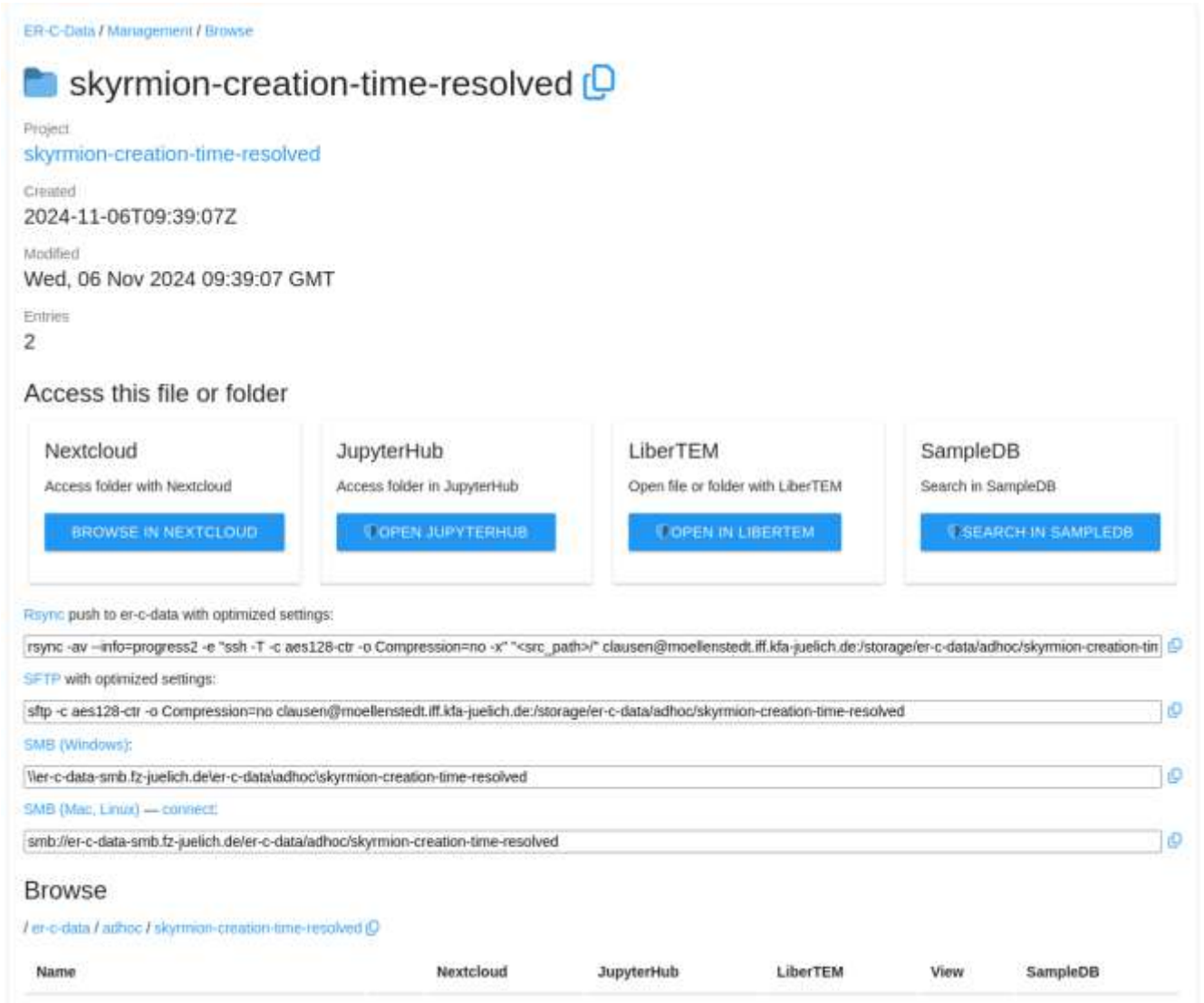
The management interface – seal session



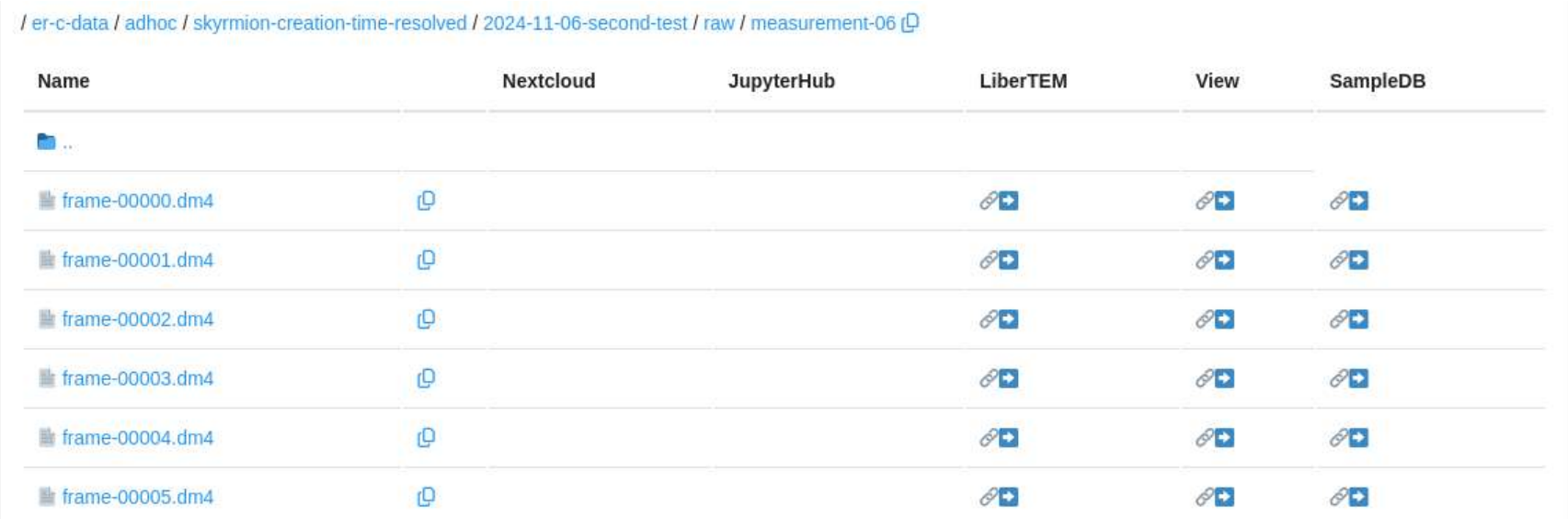
The management interface – browse and preview



The management interface – browse and preview



The management interface – browse and preview

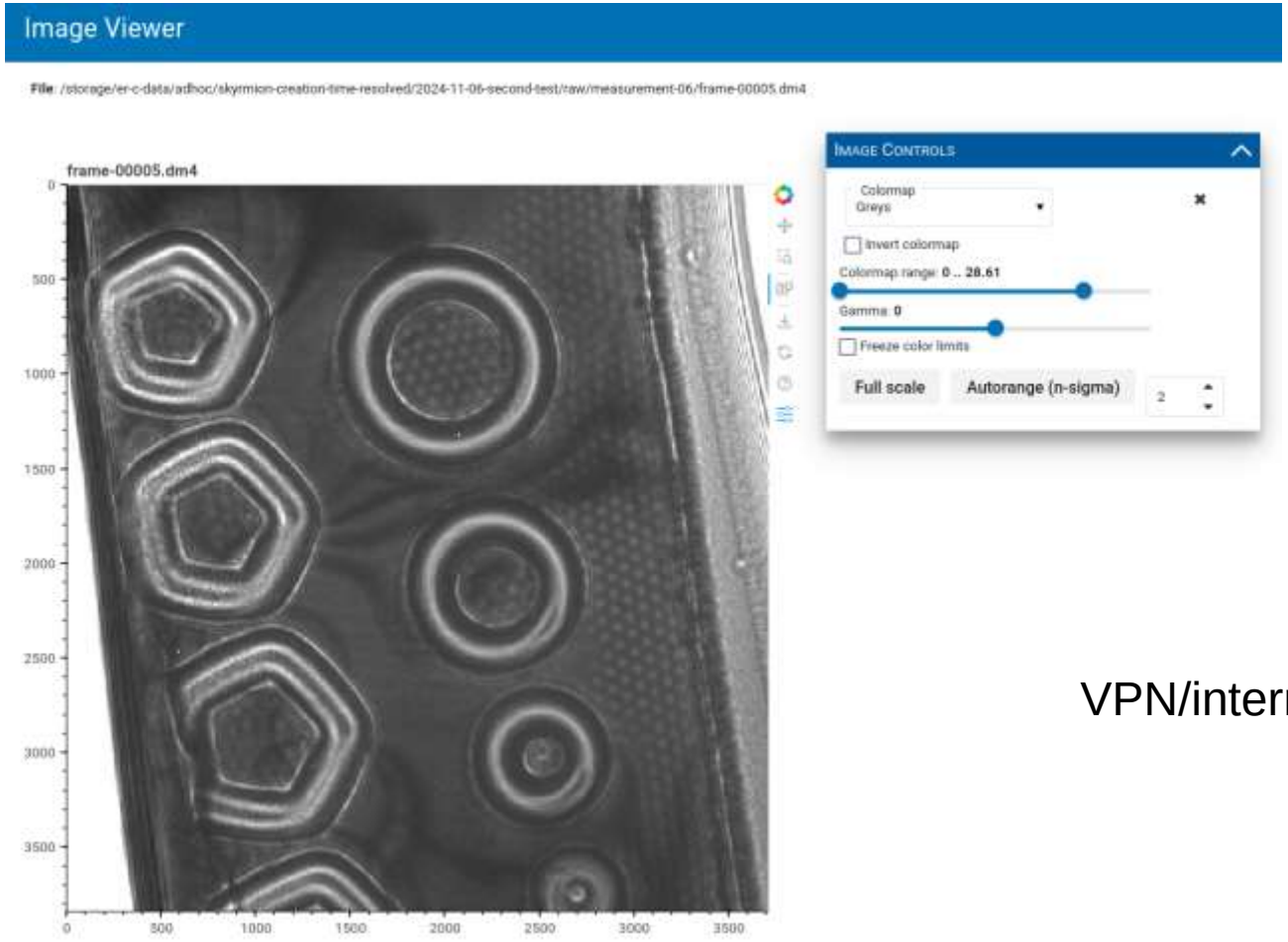


The management interface – browse and preview

/ er-c-data / adhoc / skyrmion-creation-time-resolved / 2024-11-06-second-test / raw / measurement-06

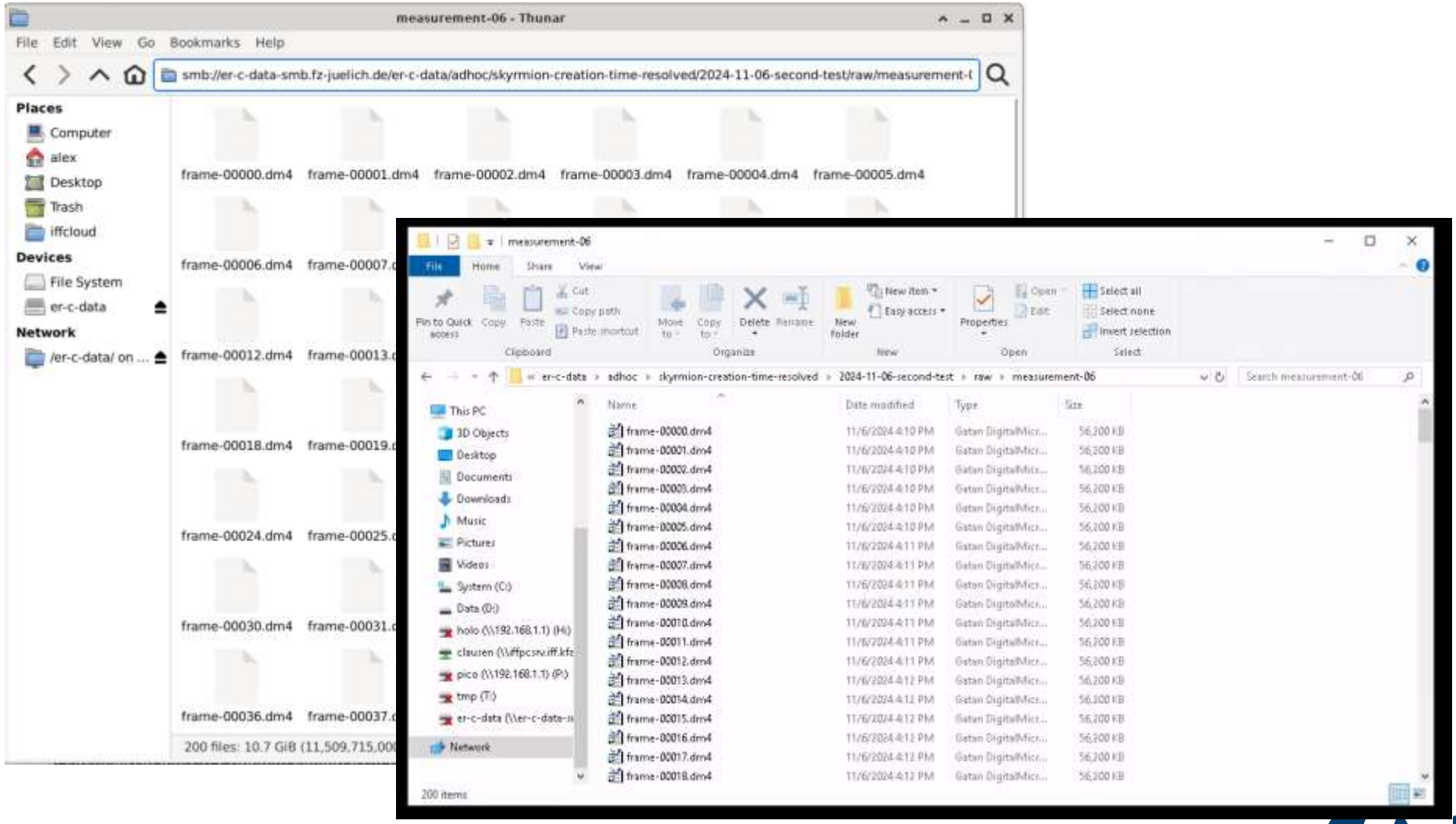
Name	Nextcloud	JupyterHub	LiberTEM	View	SampleDB
..					
frame-00000.dm4					
frame-00001.dm4					
frame-00002.dm4					
frame-00003.dm4					
frame-00004.dm4					
frame-00005.dm4					

The management interface – browse and preview



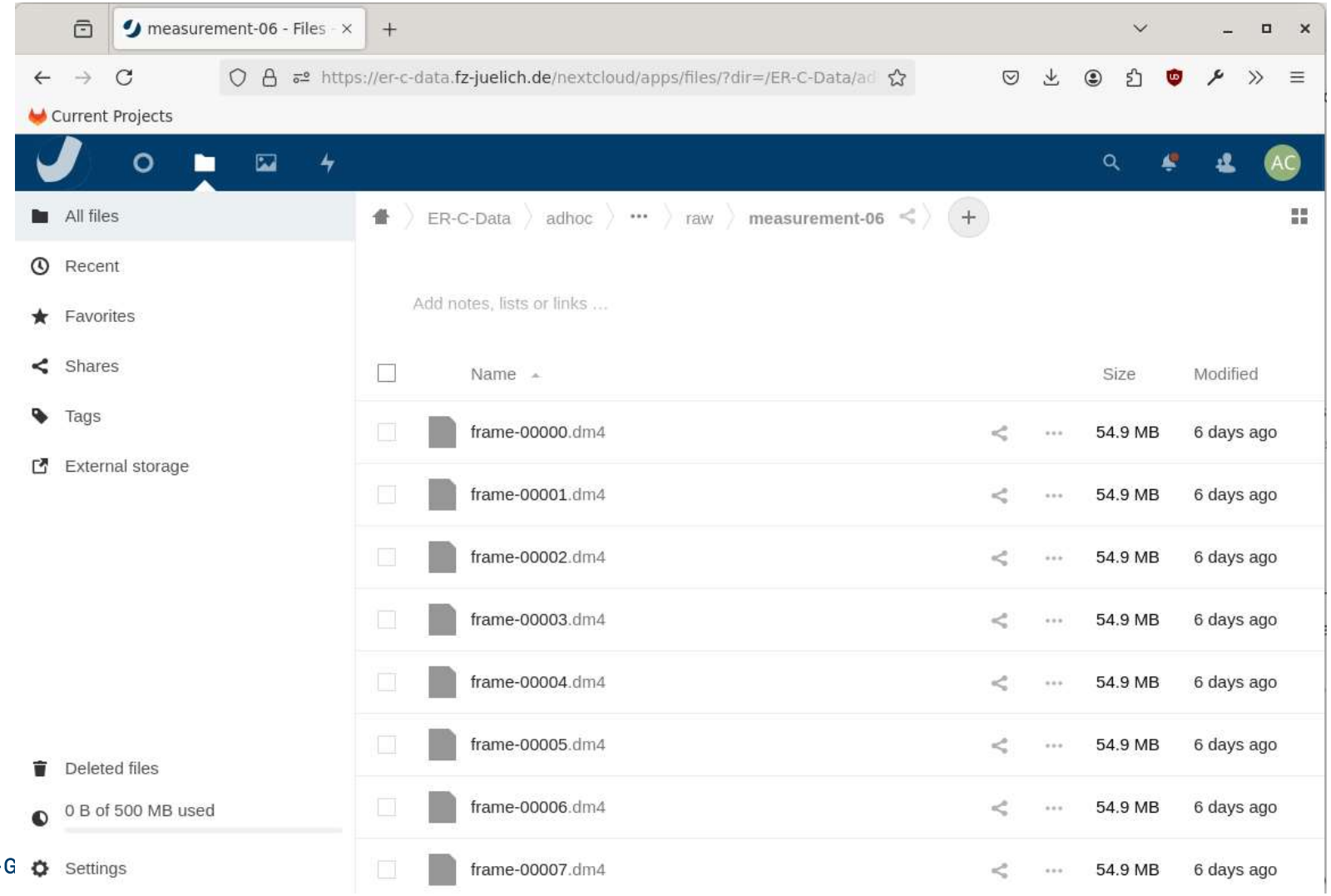
VPN/internal only!

Data access - Windows Network Drive (Win/Mac/Linux)

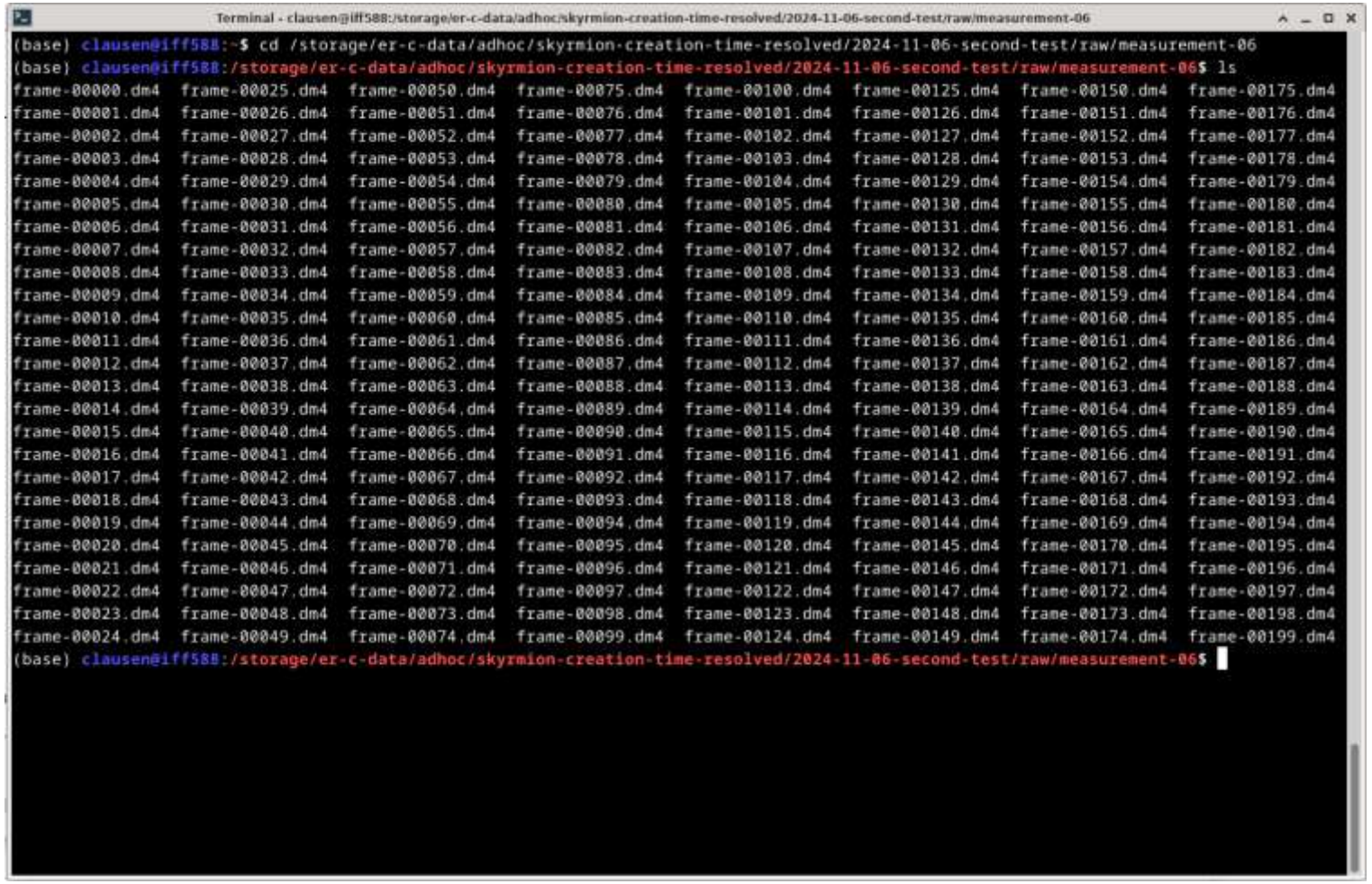


Data access – Nextcloud (sync and **share**; external read/**write** access)

https://er-c-data.fz-juelich.de/nextcloud/

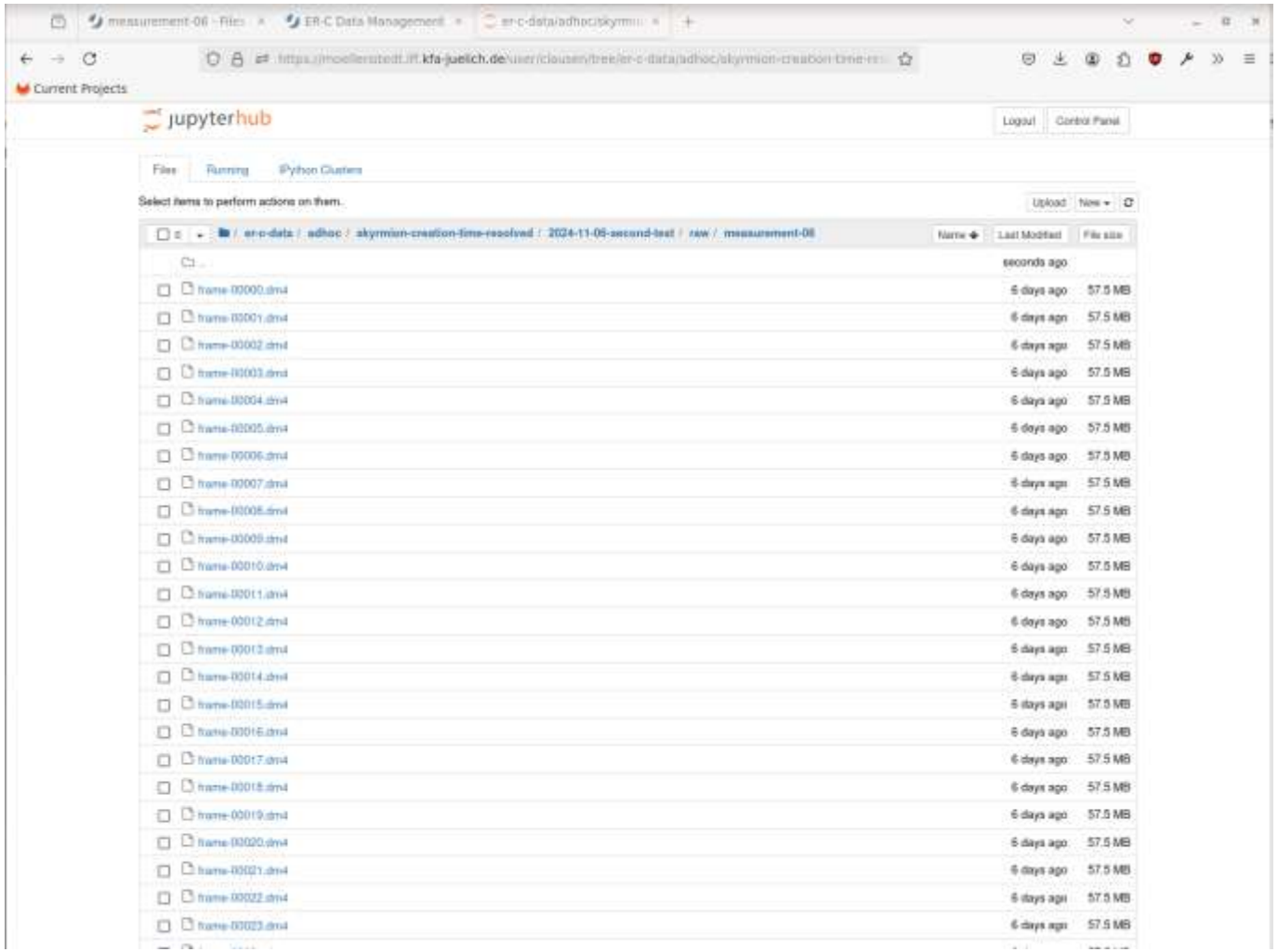


Data access – data processing workstations



Standard location: /storage/er-c-data/

Data access – jupyterhub



https://moellenstedt.iff.kfa-juelich.de/ (VPN/internal only)

DISCUSSION

- Simple and flexible – make a lot of data a bit more FAIR
- Familiar and mature components, standard interfaces
  - POSIX file system, NFS, CIFS, WebDAV, Nextcloud
  - Easy onboarding – most users familiar with network drives or Nextcloud
- Modular design
  - Robust – core keeps running even if components have issues
  - Extensible without central coordination
  - Flexible solutions via diverse interfaces

  → Don’t create a complete system, but a starting point and conditions for it to grow
- Positive user feedback, steady use

CURRENT DEVELOPMENTS

- Member of EU-wide EOSC Data Commons project
  - https://www.eosc-data-commons.eu/
- Ingest and integrations
- Reorganize storage
  - Likely Ceph cluster, remote block device
- Improve interface and services
  - Web interface overhaul
  - Private projects
- Integrate ER-C-3 (CryoEM)
- Continue growth

ACKNOWLEDGMENTS

- IT group of PGI/JCNS
  - Markus Consoir, Florian Rhiem, Josef Heinen
- People at ER-C
  - Alexander Müller, Juri Barthel, Daniel Mann, Thomas Heidler
- External partners
  - Steffen Brinckmann and team (IEK-2)
  - Matthew Bryan (CEA-LETI, AIDAS)

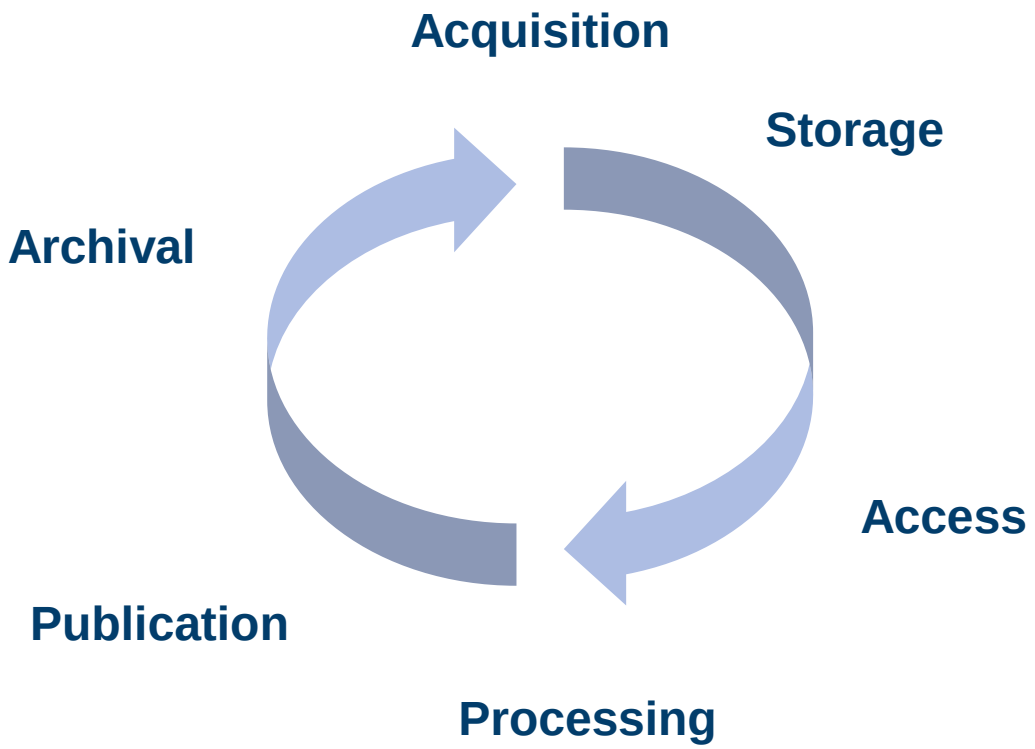
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823717 (ESTEEM3)

Cooperation within framework of AIDAS

This project has received funding from the European Union's Horizon Europe Research infrastructures programme under grant agreement No 101188179 (EOSC Data Commons)

REQUIREMENTS DATA MANAGEMENT

Acquisition – Storage – Access – Processing – Publication – Archival



JÜLICH Forschungszentrum

REQUIREMENTS DATA MANAGEMENT

Acquisition – Storage – Access – Processing – Publication – Archival

Absorb the data coming from the detector and store it somewhere

- Usually provided by detector vendors in some form
- Some form of data access with sufficient speed

Currently: Stored on instrument systems

WIP: Store directly in data management

REQUIREMENTS DATA MANAGEMENT

Acquisition – Storage – Access – Processing – Publication – Archival

Inspect data quickly during session

- Decide next steps of experiment
- Possibly live analysis

Currently: Install inspection software locally (→ LiberTEM project)

WIP: Use web services

REQUIREMENTS DATA MANAGEMENT

Keep notes, meta-information linked to raw data files

- Create a stable reference: File paths or other identifiers
- Relevant data scattered on multiple systems!
  - Microscope, spectrometer, camera, in-situ control, …

- Easily create new projects – avoid later data movement or catch-all megaprojects
- Move data into management system ASAP
- Make available on microscope systems (WIP)
- Link to raw data via landing page URL e.g. in ELNs, presentations, GitHub issues, …

REQUIREMENTS DATA MANAGEMENT

Transfer to permanent storage location

- During or at end of session
- Correct ownership, provenance tracking
  - Microscopes usually run under system accounts!
- Keep connection to notes, other meta-information

- Ingest: WIP (later discussion if interested)
  - Also web-based ingest with Nextcloud
- Fast network access: WIP
- Allows storing all kinds of files, incl. notes etc. – not limited to microscopy datasets

REQUIREMENTS DATA MANAGEMENT

Acquisition – **Storage** – Access – Processing – Publication – Archival

Prevent data loss

- Main risk are you and I – ask me how I know 
- Write protection of source data
- Keep associated mutable user data close
  - Analysis scripts, figures, paper drafts, …
- Replication, backup, snapshots

- Incremental daily backup of the central file system
- Allow making data immutable in the file system – not overridable on clients
- Restrict write access to own files

REQUIREMENTS DATA MANAGEMENT

Acquisition – Storage – **Access** – Processing – Publication – Archival

Collaboration, sharing, exchange

- Selective access control read/write
- Internal access for team members
  - Includes collaboration on projects
- Sharing with external partners
- Microscope systems
- Shared workstations
- PCs

- Mobile devices
- Services
- External storage media
- Which software, OS, protocols?

- Standard protocols
- Wide range of access pathways
- External sharing via Nextcloud

REQUIREMENTS DATA MANAGEMENT

Shared compute systems

- High performance computers with fast data access
- Computation close to the data
  - Remote desktop, Jupyter, LiberTEM web GUI
- Gateway for transfers

- NFS (Linux), CIFS (Windows, MacOS)
- Fast network (WIP, currently small “island” but extension imminent)

REQUIREMENTS DATA MANAGEMENT

Acquisition – Storage – Access – **Processing** – Publication – Archival

PCs, laptops

- Testing, development, travel, data exchange, contingency
- Helpful if user’s devices allow basic data analysis
- Work pretty well for smaller data

- Nextcloud local sync
- CIFS, SSHFS

REQUIREMENTS DATA MANAGEMENT

Keep data that is not in active use anymore

- Store safely for at least the mandated period of time
- Possibly, perpetual storage
  - National archives and such?
- Keep all relevant meta-information connected
  - ELN entries, notes, software, provenance, specimens, …
  - Information to allow others to understand what is stored there when the original authors are gone
- Make sure there’s a record of what is in the archive

- Currently deferred, keep everything in hot storage for now