Pitschi: Rolling Out an Imaging Big Dataset Management Framework for Advanced Microscopy

Mark Endrei

Nishanthi Dasanayaka

Research Computing Centre
The University of Queensland

CMM@UQ

Centre for Microscopy and Microanalysis

Microscopy and microanalysis related services, research and development

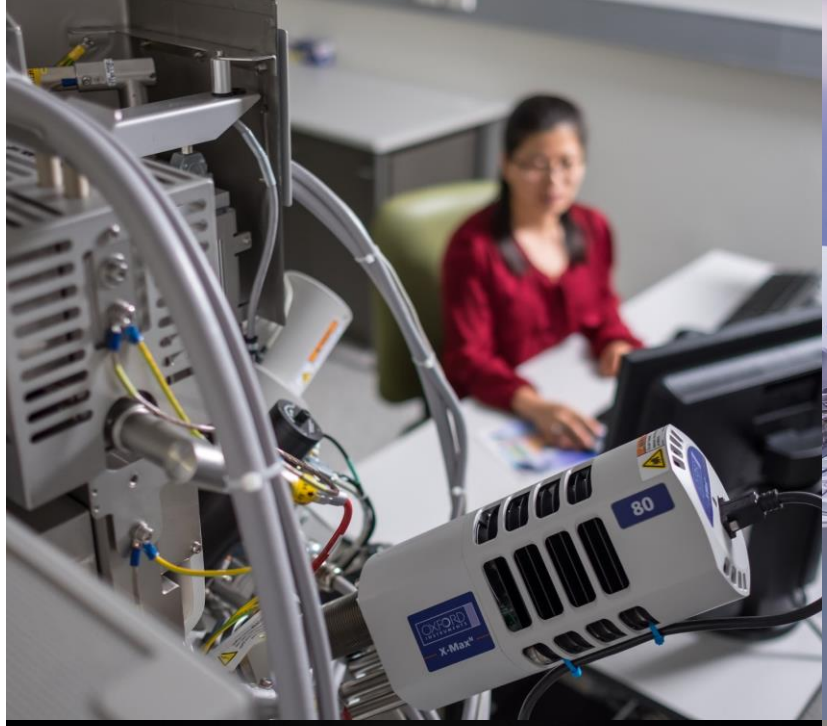
42 specialist staff

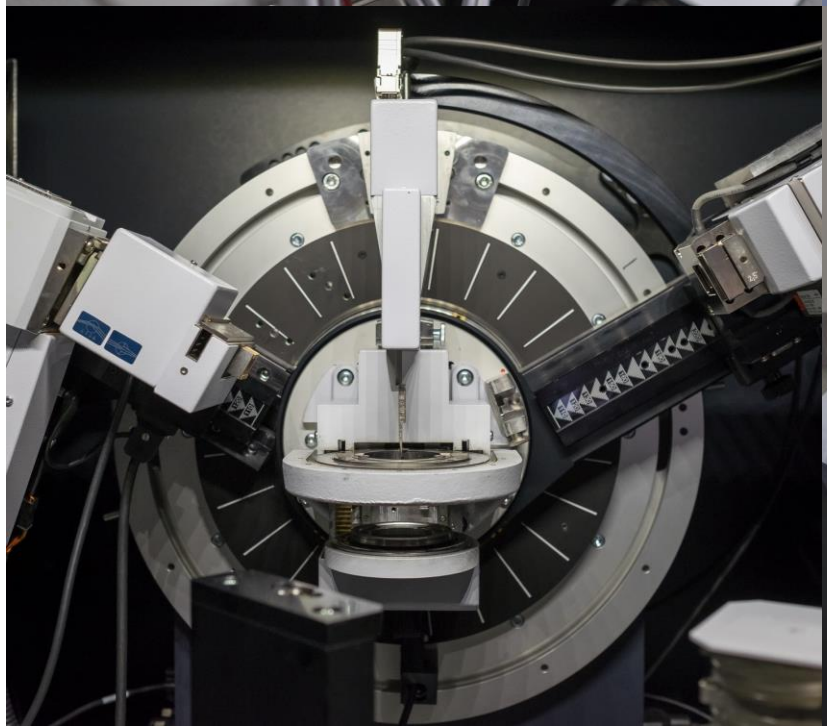
Users

- **Students, teaching and research staff**
- **Industry users and clients**

Facilities

- **Life Science and Soft Matter Facility**
  AIBN lab
- **Material Science Microscopy and Novel Imaging Technologies**
  Hawken lab
- **X-ray Material Science and Spectrometry**
  Chemistry lab
- **Structural Biology and Protein Crystallography**
  QBP lab





Pitschi objectives

Make it as easy and simple as possible for users to deposit their data and metadata

End-to-end process for

- **Capture, transfer raw data to storage collections, index at image repository**
- **UQ RIMS instrument booking system of record**

One storage collection per project

- **Project owner controls access to data**
- **Move away from one collection per instrument model**

Harvest as much metadata as possible – automatically

- **Allow users to view/search their data and process it on UQ HPC cluster**

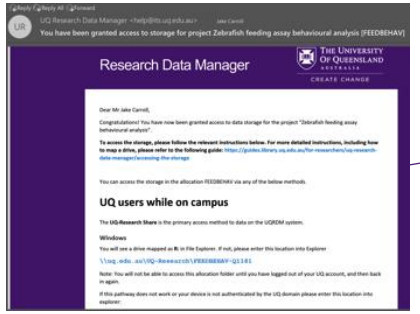
A central repository that adheres FAIR data principles

Typical workflow at UQ

Researcher obtains a collection.

Research Data Manager

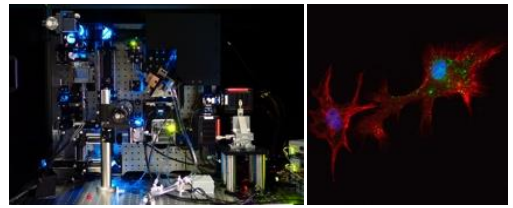
Designed at The University of Queensland



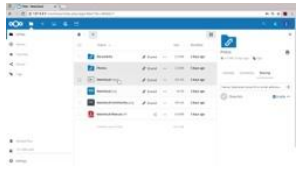
Researcher is sent an email explaining access instructions.

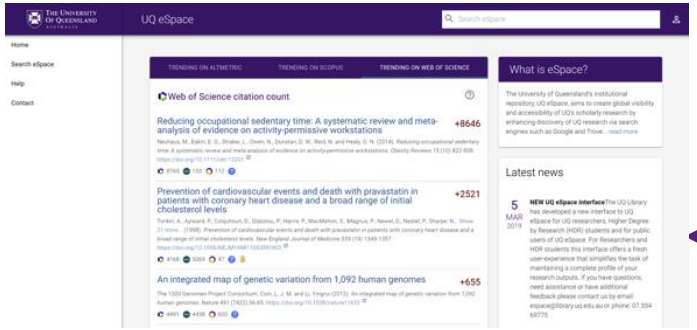
Collection then mounted on the instrument.

Researcher then acquires data and stores



UQ Data fabrics (MeDICI)

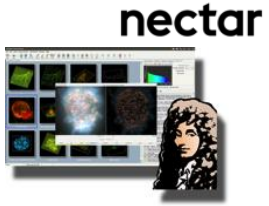




DOIs can be minted, published to UQ eSpace.
RDM can facilitate data linking back for durable URL.

PYTORCH

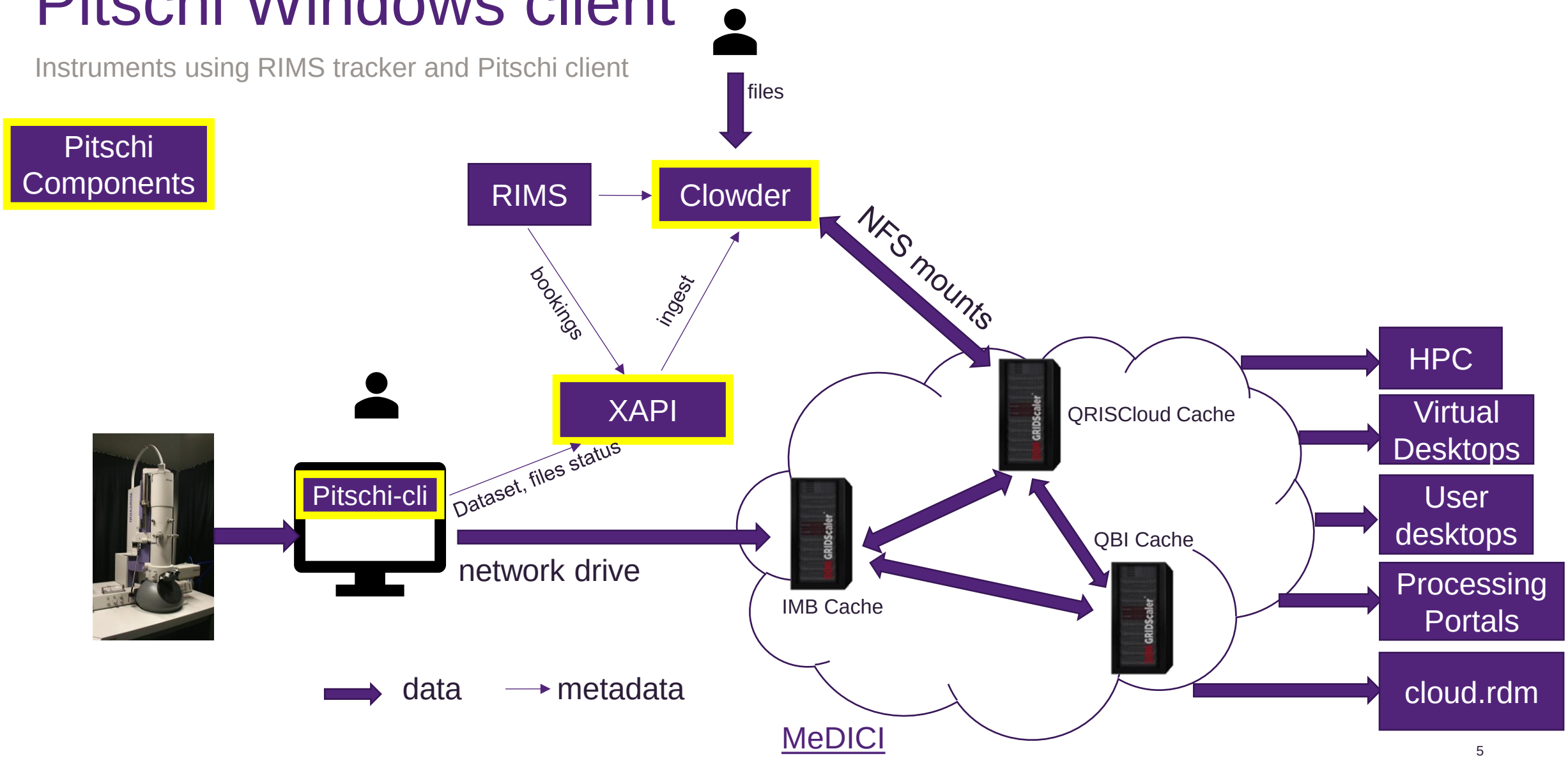
nectar



It is already on the fabric. Can be accessed via CVL, supercomputers, other.

Pitschi Windows client

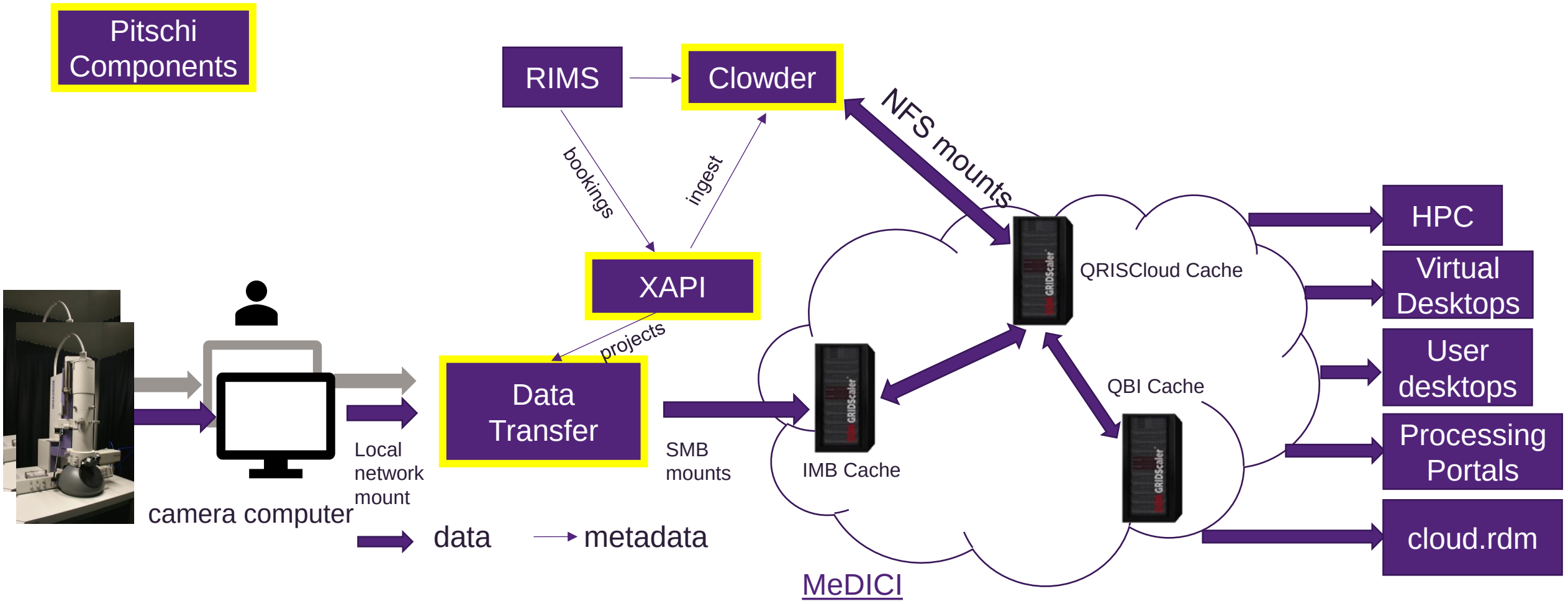
Instruments using RIMS tracker and Pitschi client



Pitschi datamover

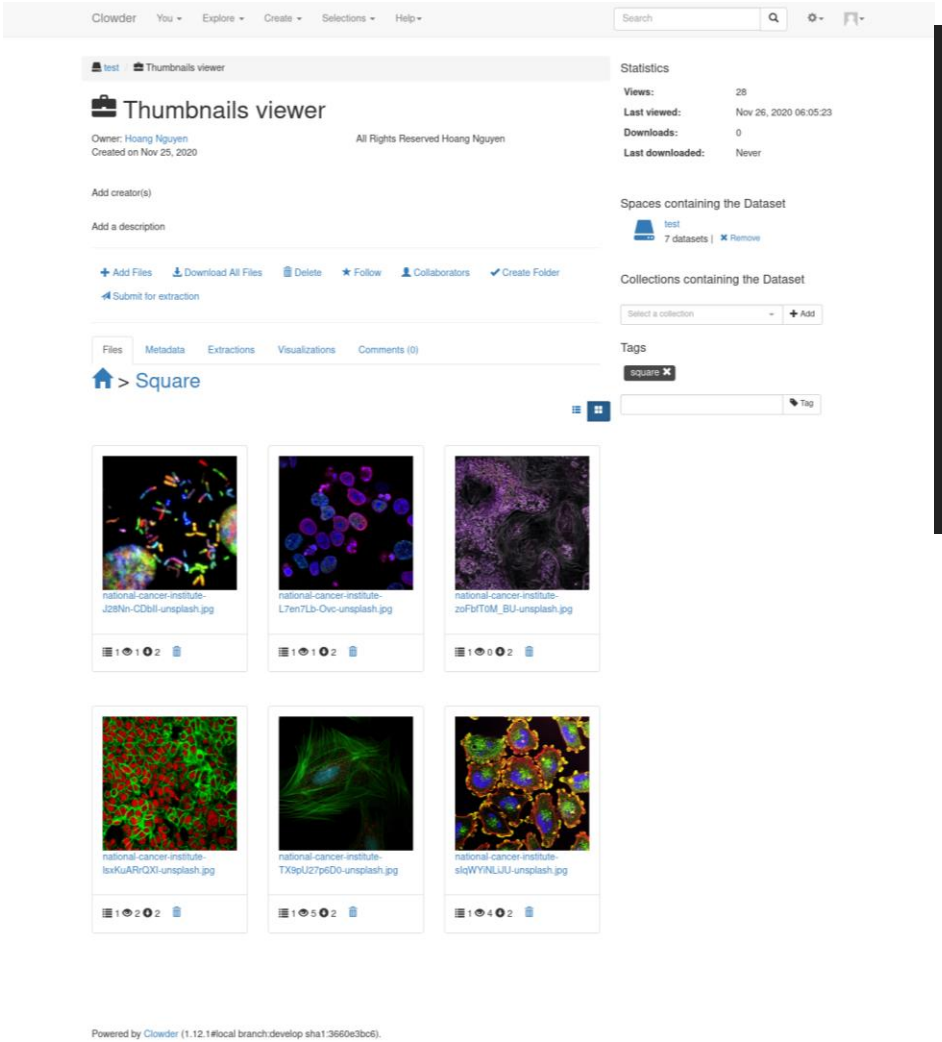
Instruments not using RIMS tracker

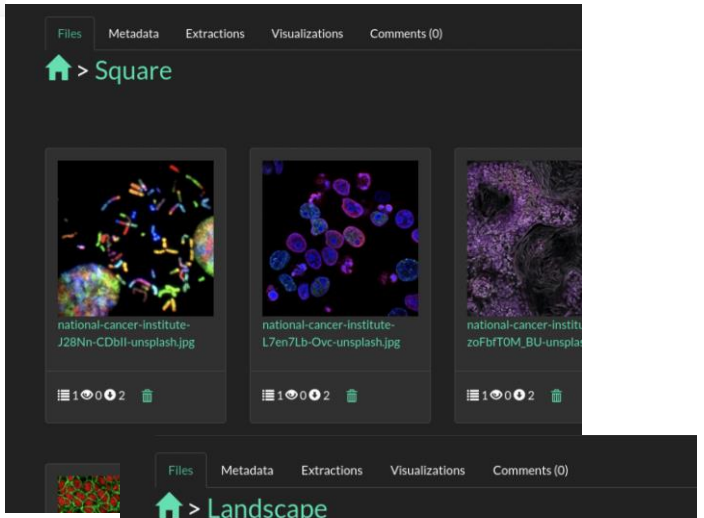
DEMO https://youtu.be/Xya0XzPfDnA

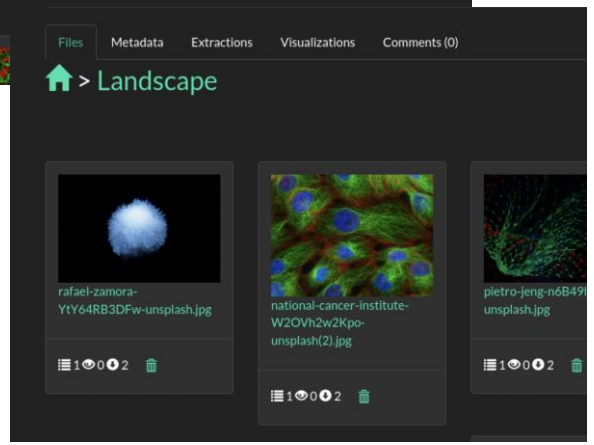


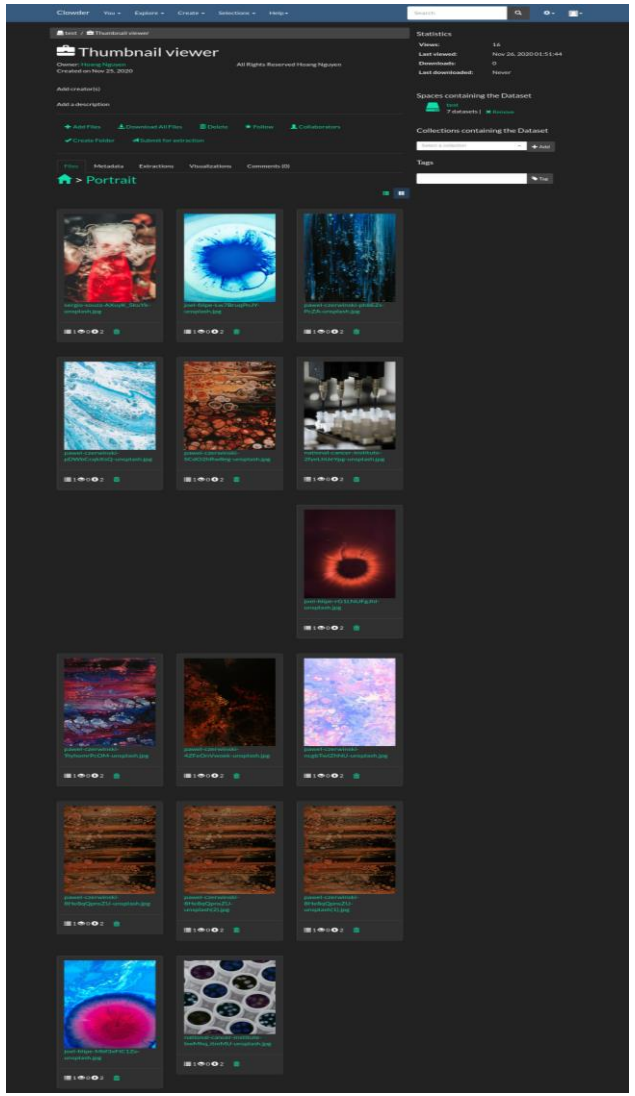
Clowder image browser

https://pitschi.rcc.uq.edu.au/









Rollout progress

Metrics and highlights

	2022	2024
Labs	4	5
Instruments	22	50
Projects	100	400
Users	160	280
Datasets	270	5,000
Files	650,000	4,000,000
Data (TB)	80	360

Cryogenic electron microscopes are the biggest data generators at CMM

Weekly training is part of the on-boarding process for new users

Key enhancements

Since 2022

Enhancement	Win client	Datamover	XAPI
Training sessions	✓		✓
Supervised / Fee-for-service sessions		✓	
Multiple cameras support		✓	
User feedback option		✓	
RIMS data sync improvements		✓	✓
Session recovery improvements	✓	✓	✓
Remote logging	✓	✓	
Auto-update	✓		
Container deployment		✓	

Clowder 1

Emerging challenges

Aging software stack

- **Out-of-support libraries**
- **DevOps skills availability**

Kubernetes compatibility

- **External storage integration**

Clowder 2 features

User highlights

New web interface

Shareable links

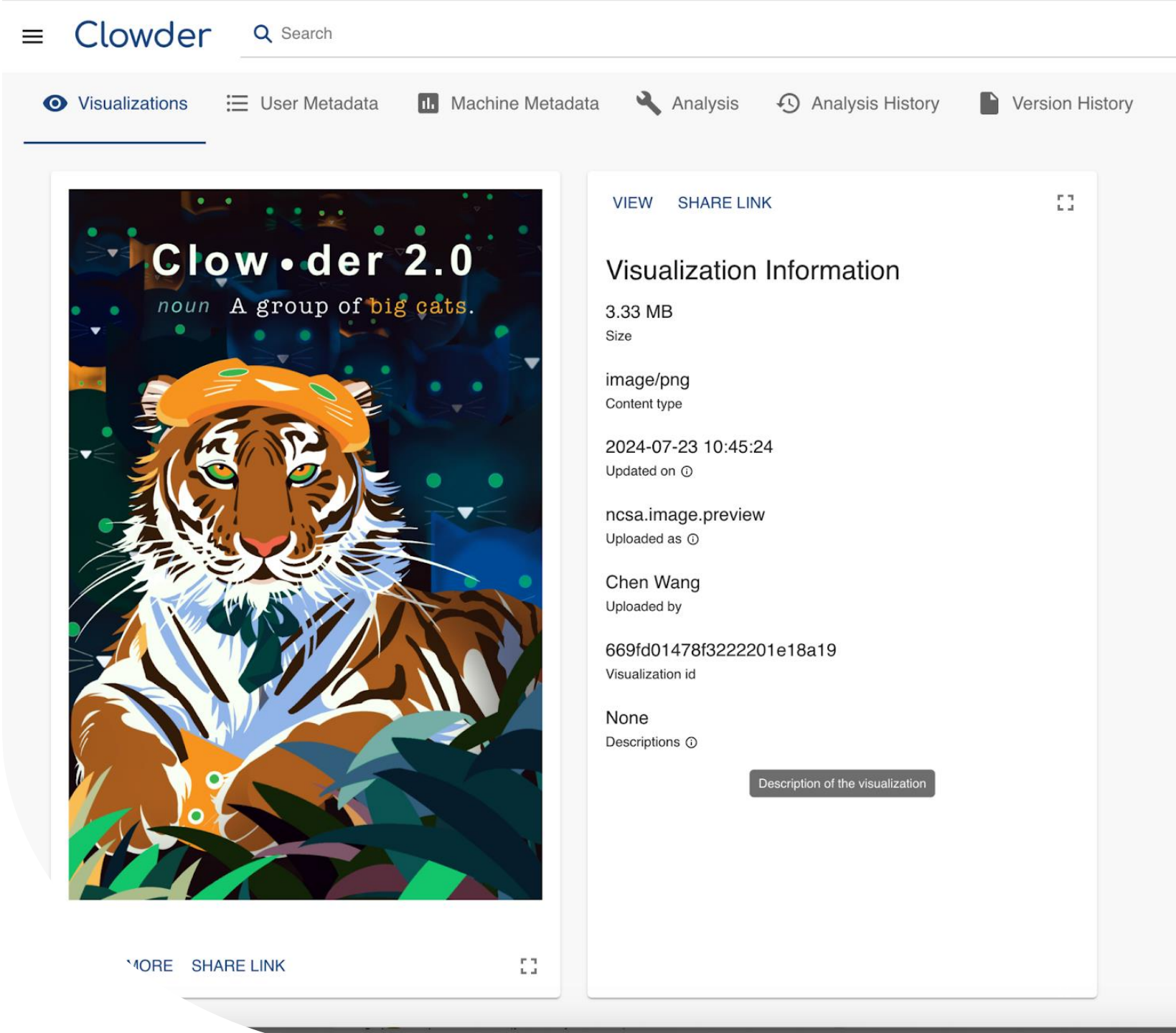
Dataset versioning and history

Custom metadata

Mandatory metadata

Collaborator groups

Jupyter integration



Clowder 2

Expected benefits

New software stack

- **React frontend**
- **Python FastAPI backend**
- **S3 compatible object storage**
- **Keycloak OIDC id management**
- **Kubernetes/Helm deployment**

Research tools integration

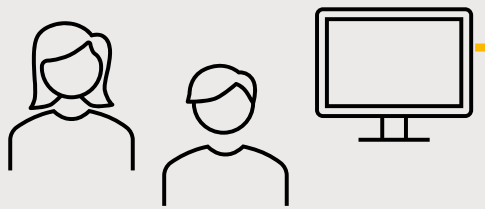
Open source

Community

Clowder 2 migration plan

Target data access and workload distribution approach

Researchers/Collaborators and Lab Instruments



Kubernetes Cluster

Interactive

Clowder 2

API

XAPI

Batch

RIMS Data Sync

GPU User Jobs

Jupyterhub

Notebooks

HPC Cluster

slurm workload manager

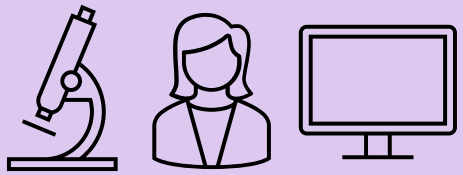
User Jobs

Research Data Manager

Researcher Data Collections via Protocol Gateway

Rollout successes

What went well



Researchers

Data management – instrument to collection sync

- **Drop and forget**
- **Minimise risk of data loss**

External collaboration

- **AAF logins**

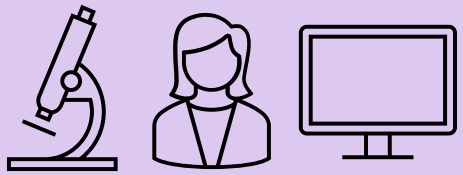
DevOps

CI/CD lifecyle

- **GitHub, DockerHub, Ansible, Docker Swarm**

Rollout challenges

What needed further work



Researchers

Data management use

- **need leadership champions**
- **limit legacy workflow access**

Portal use

- **need convincing case on benefits**

DevOps

Network/infrastructure resilience

- **need monitoring/alerting, auto-retry/recovery**

Multi-site support

- **remote access, centralised logging, auto-update**

Self-help

- **need FAQs, on-site users with elevated access**

Dev and test environments

Next steps

For 2024 and 1H2025

Phase 4 and 5 instruments

- Locked down instrument PCs
- Unsupported operating systems
- Datamover running on SFF Linux box

Clowder 2

- Release due Q4 2024
- Migration to Object Storage

Kubernetes on Nectar Magnum

- In line with other UQ-RCC research portals

Atypical instruments

Clowder 2

kubernetes

Thank you

UQ RCC

Mark Endrei
Nishanthi Dasanayaka
Jake Carroll

UQ CMM

Rubbiya Ali
Tom Mason
Prof Roger Wepf