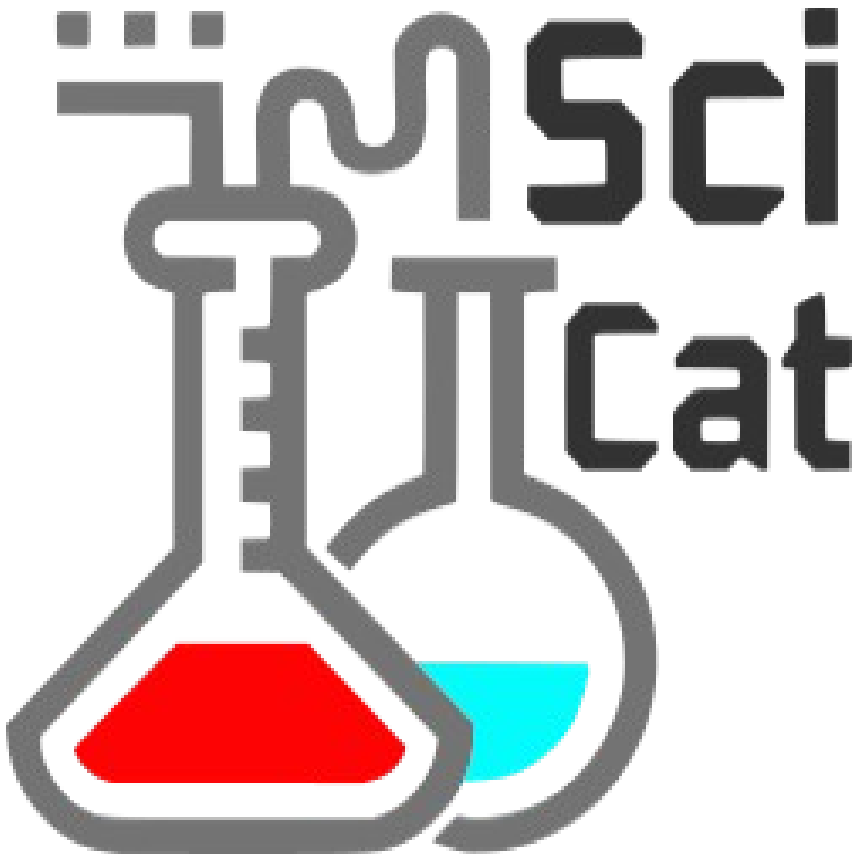
PAUL SCHERRER INSTITUT

PSI



PAUL SCHERRER INSTITUT

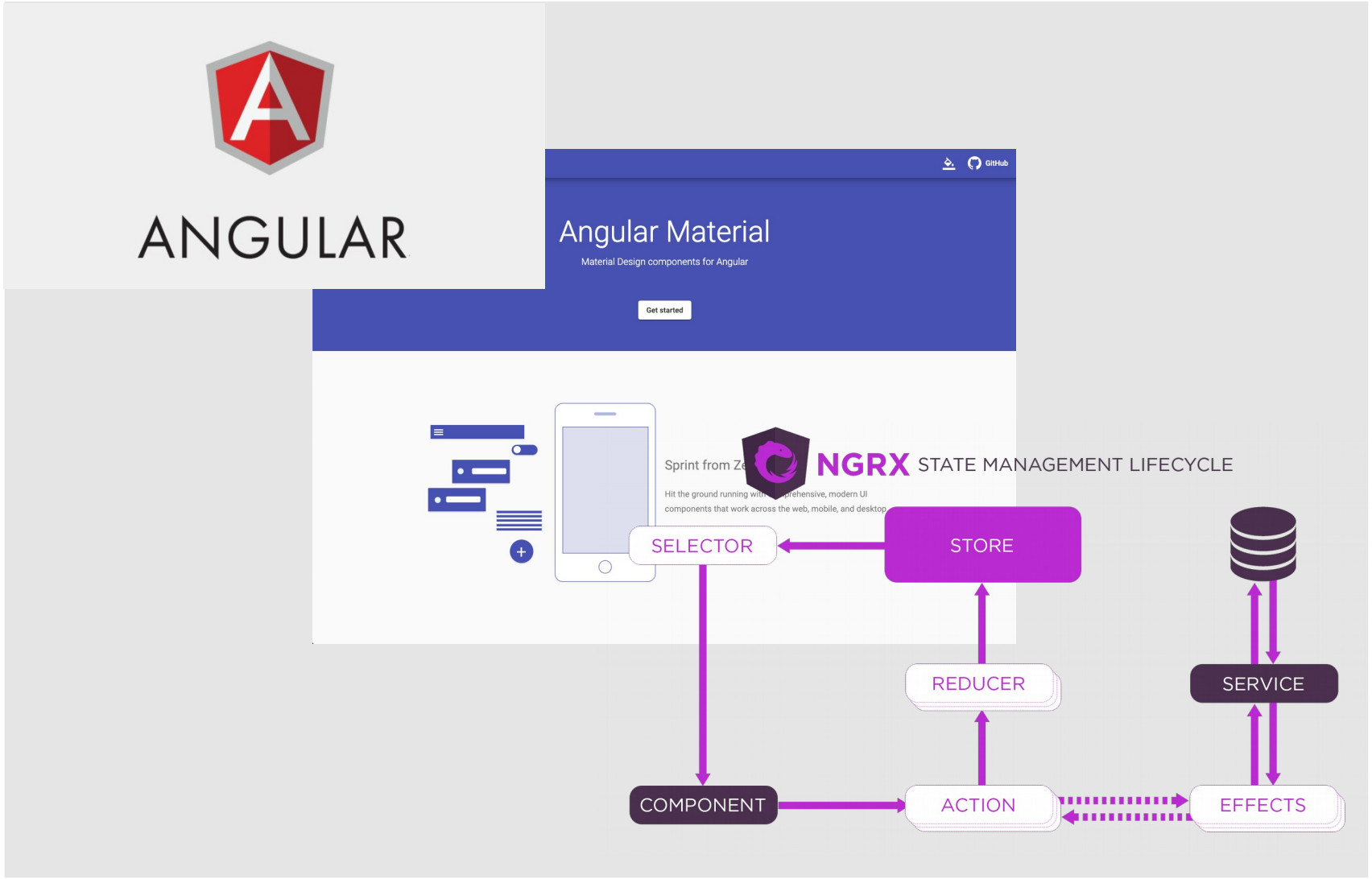
PSI

Luke Gorman – Software Engineer (Paul Scherrer Institut)

What is SciCat

- Data catalog management system for scientific data
- Aims to support the management of the whole data lifecycle
- Open source (https://github.com/scicatproject)
- Modular design to cope with evolving requirements from scientists
- Technical architecture based on Microservices with the latest web technologies

The Stack - Client



Frontend

- Built using Angular and Ngrx
- Created with Typescript
- Component based architecture to reuse throughout the application
- Based on Angular Material Widget Set
- Responsive, Standards compliant

The Stack – Back End

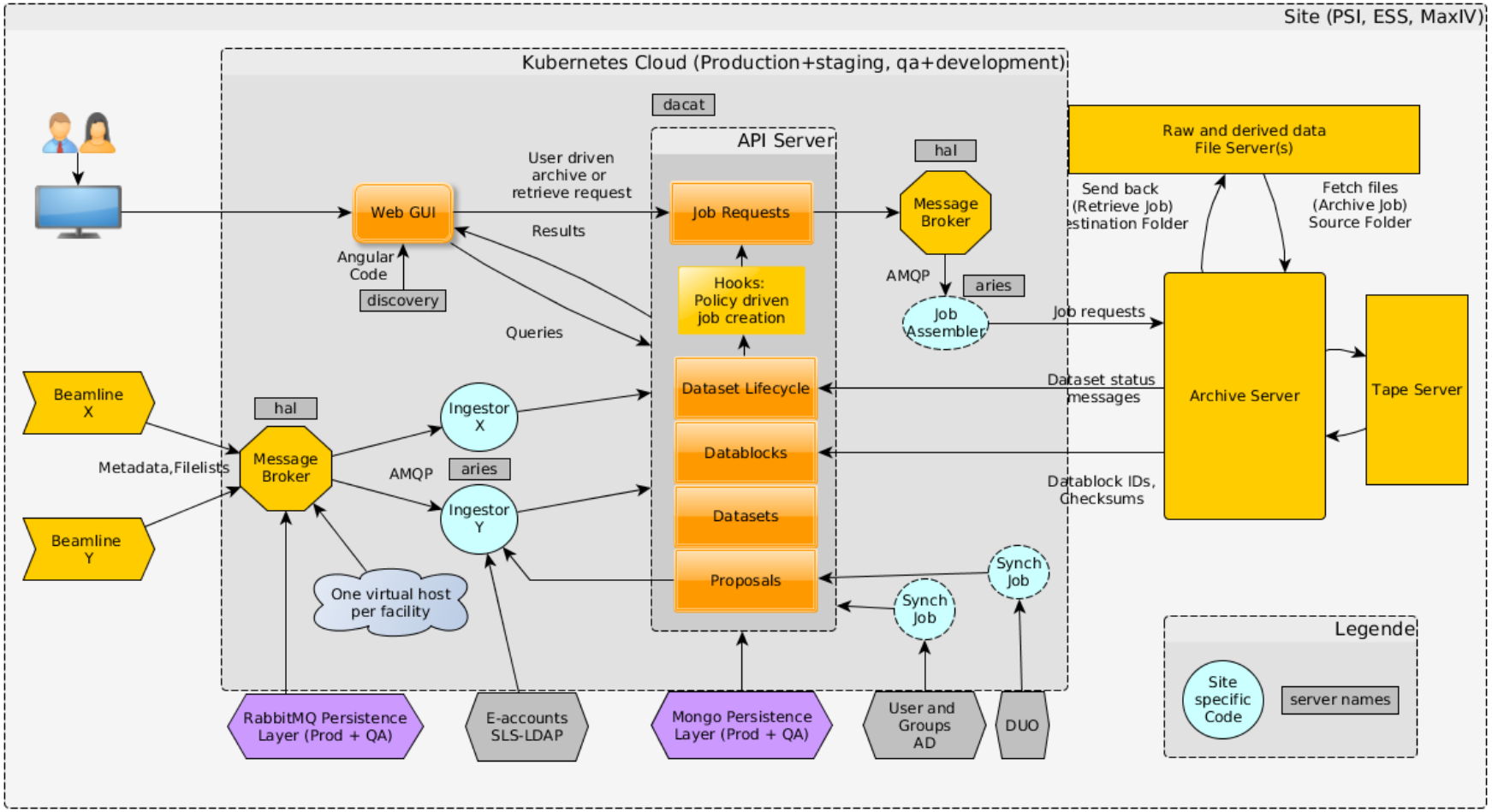
mongoDB

LoopBack

Express

Fast, unopinionated, minimalist
web framework for Node.js

Overall Data Catalog Architecture



Why use it?

- Allows to add metadata to your files to give them more meaning
- Allows you to organize data and make the data findable
- Allows you to publish your data and make it citable via a digital object identifier (DOI)
- Can be used both for raw (experiment) data and derived (analysis results) data

1. Define Datasets

- Simply list all the file names that logically belong to a measurement (e.g. a data taking run). In the simplest case this is just the name of a folder: all files within are taken as the dataset files
- Define data **ownership** by assigning a p-group via Digital user Office DUO. Only members of the pgroup get access to the data. Membership defined by beamline manager BM or principal investigator PI
- Define minimal administrative metadata (more is possible)

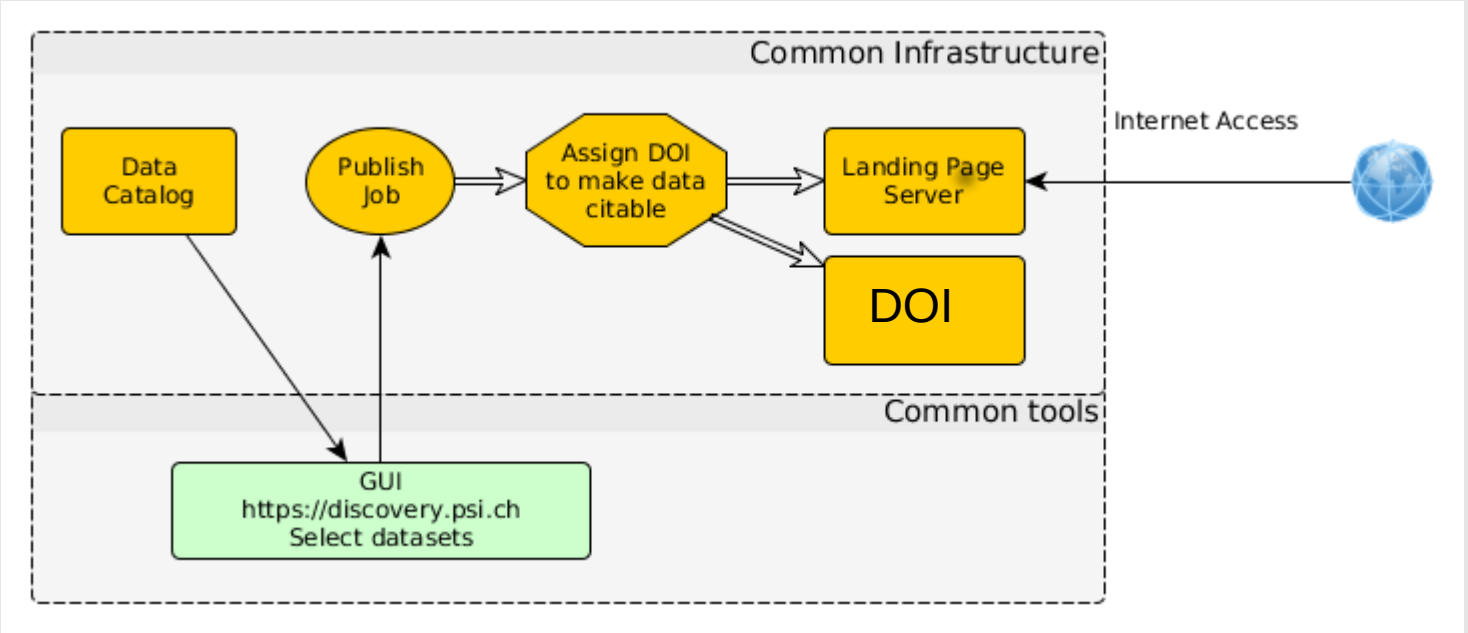
```
{
    "principalInvestigator": "federica.marone@psi.ch",
    "creationLocation": "/PSI/SLS/TOMCAT",
    "sourceFolder": "/data/experiment/run/17",
    "type": "raw",
    "ownerGroup":"p16623"
}
```

2. Define Scientific Metadata

- The definition of scientific meta data is completely up to the scientific discipline.
- Ideally follow a standard if it exists, e.g. Nexus based HDF5 files
- Scientific metadata can also be added later
- Just an example:

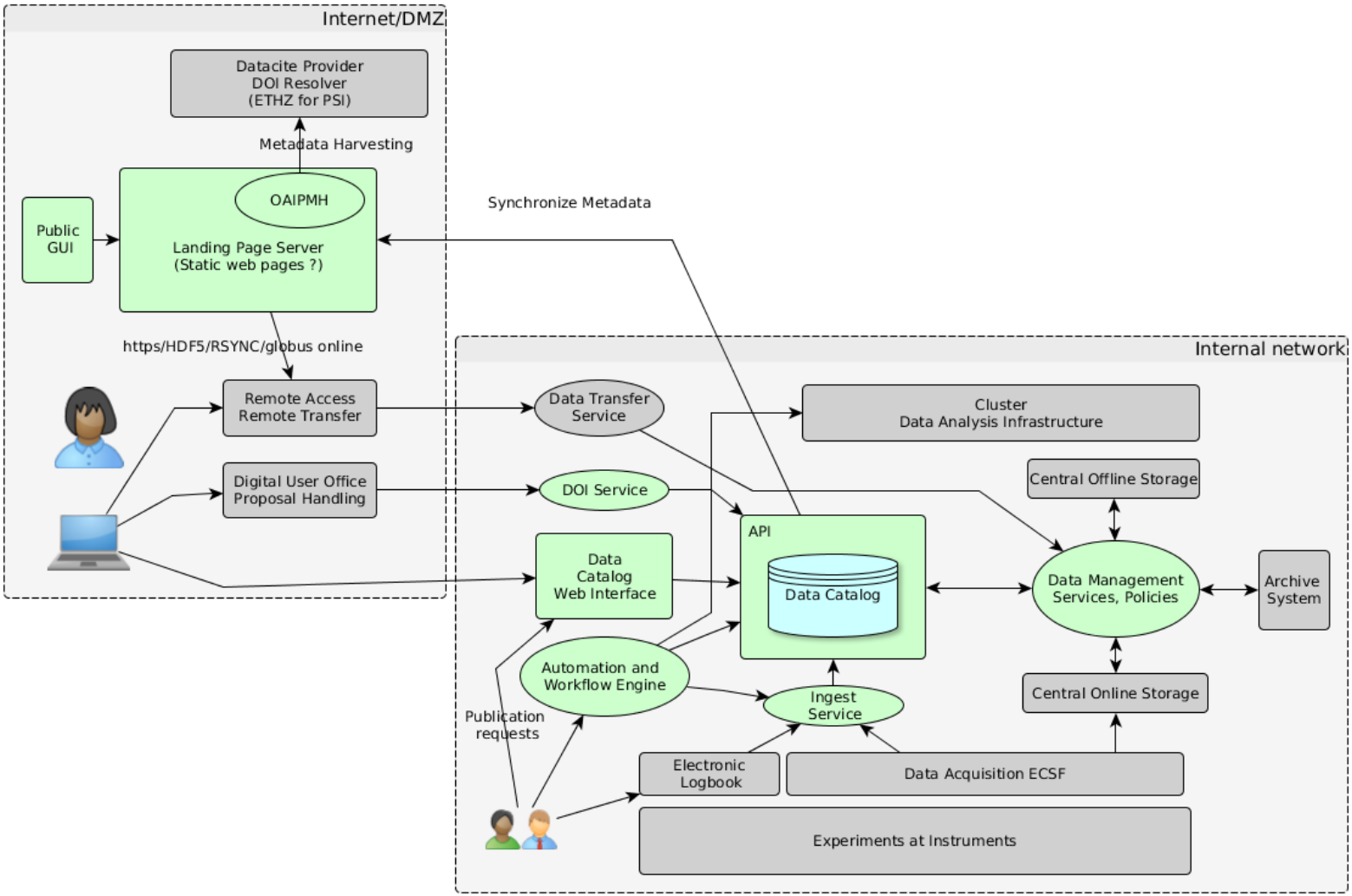
```
"beamlineParameters": {
    "Monostripe": "Ru/C",
    "Ring current": {
        "v": 0.402246,
        "u": "A"
    },
    "Beam energy": {
        "v": 22595,
        "u": "eV"
    }
},
"detectorParameters": {
    "Objective": 20,
    "Scintillator": "LAG 20um",
    "Exposure time": {
        "v": 0.4,
        "u": "s"
    }
}…
```

How it works: publishing data

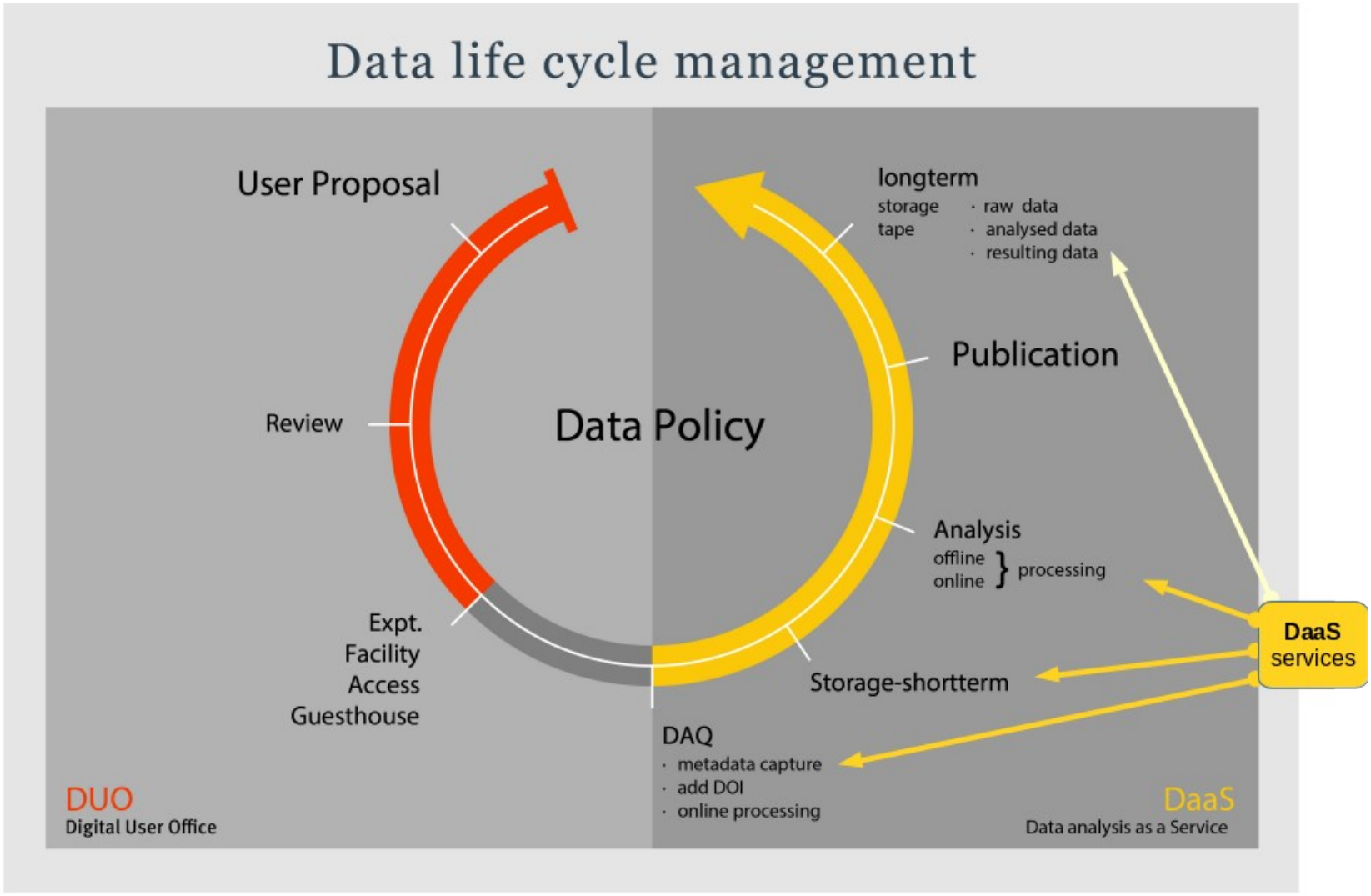


Driven by researchers who want to publish in a journal

Architecture Update for Publication Workflow



Bigger Picture:Data Lifecycle

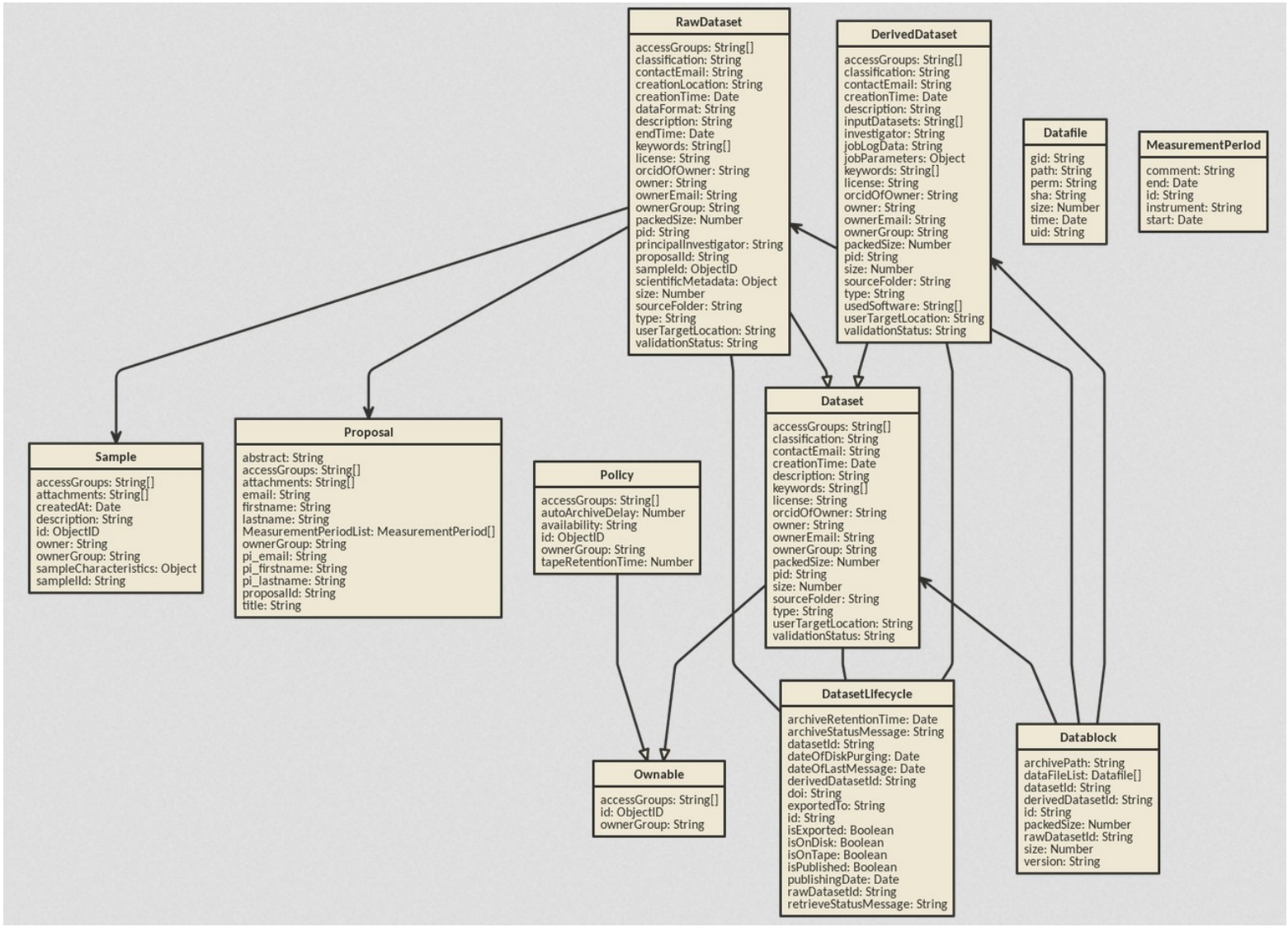


PAUL SCHERRER INSTITUT

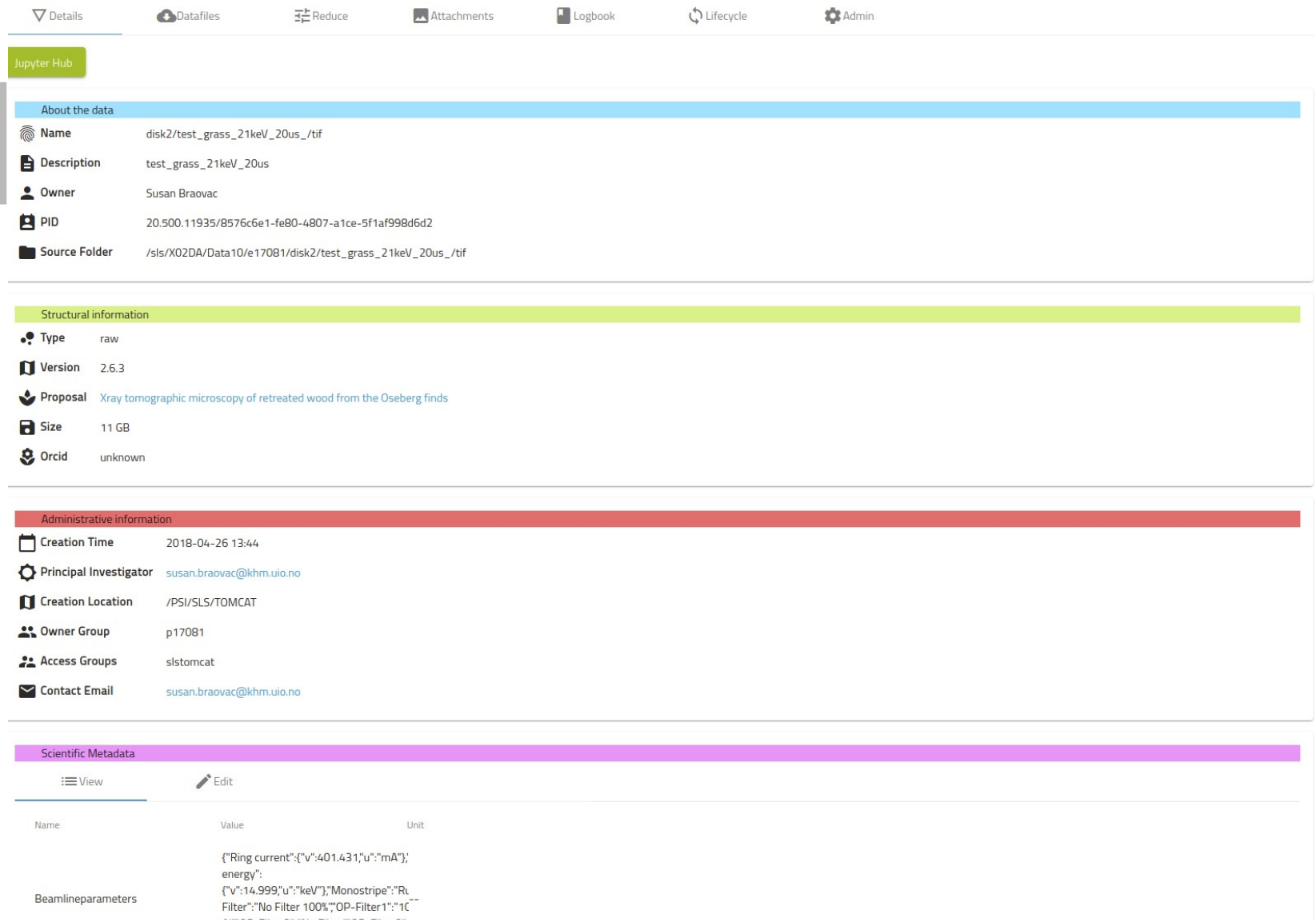
Datastore

- MongoDB backend
- NoSQL - Document based storage, collections,"schemaless"
- Map/Reduce/Facet Aggregation queries
- Powerful indexing and support for file storage
- Fault tolerant and drivers for most languages

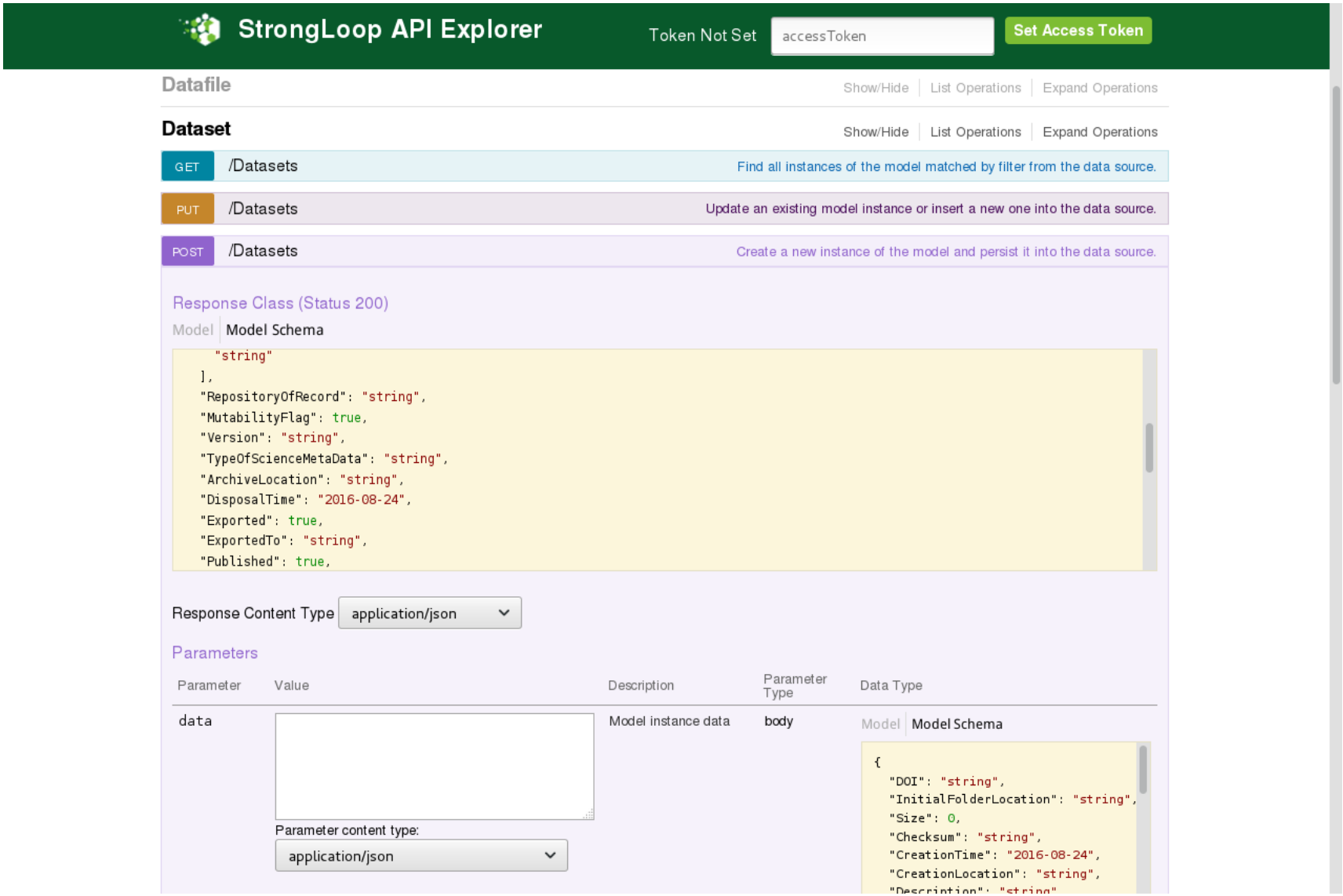
Data Model: Raw/Derived Datasets, Proposals…



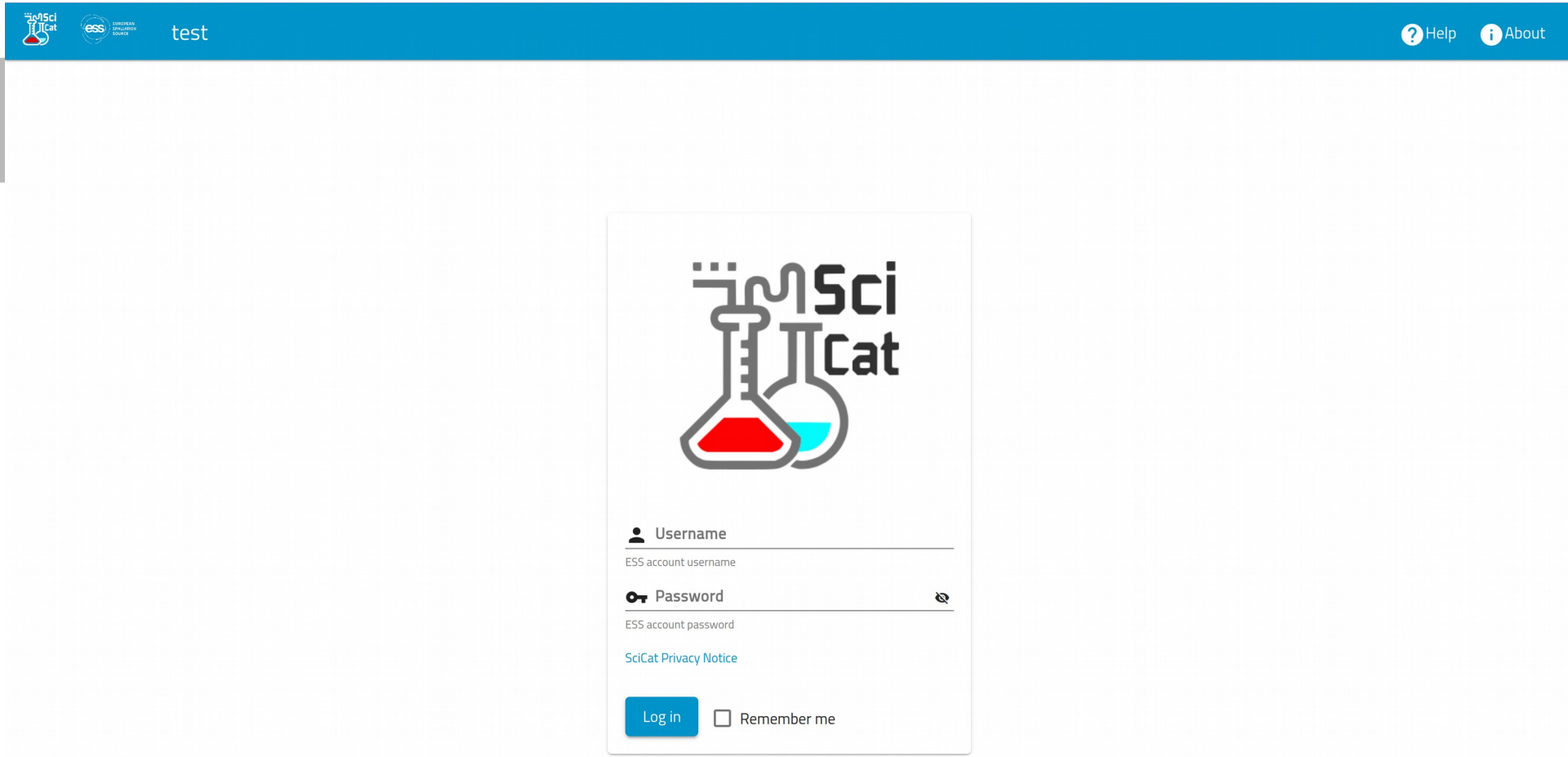
Metadata


Details
Datafiles
Reduce
Attachments
Logbook
Lifecycle
Admin
Jupyter Hub
About the data
Name
disk2/test_grass_21keV_20us_/tif
Description
test_grass_21keV_20us
Owner
Susan Braovac
PID
20.500.11935/8576c6e1-fe80-4807-a1ce-5f1af998d6d2
Source Folder
/sls/X02DA/Data10/e17081/disk2/test_grass_21keV_20us_/tif
Structural information
Type
raw
Version
2.6.3
Proposal
Xray tomographic microscopy of retreated wood from the Oseberg finds
Size
11 GB
Orcid
unknown
Administrative information
Creation Time
2018-04-26 13:44
Principal Investigator
susan.braovac@khm.uio.no
Creation Location
/PSI/SLS/TOMCAT
Owner Group
p17081
Access Groups
slstomcat
Contact Email
susan.braovac@khm.uio.no
Scientific Metadata
View
Edit
Name
Value
Unit
Beamlineparameters
{"Ring current":{"v":401.431,"u":"mA"},'
energy":
{"v":14.999,"u":"keV"},"Monostripe":"R
Filter":"No Filter 100%","OP-Filter1":"1

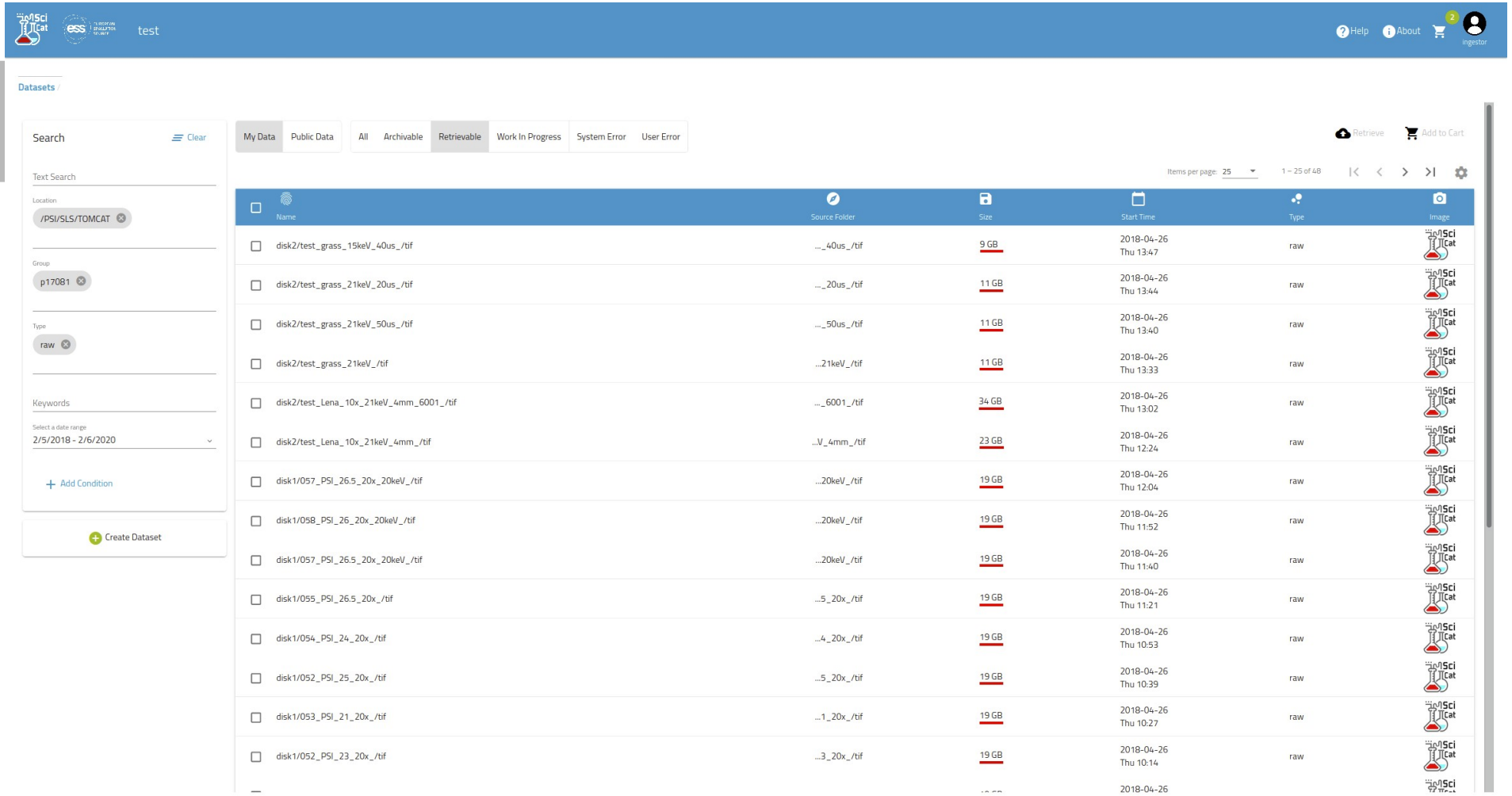
Screenshot of API Explorer



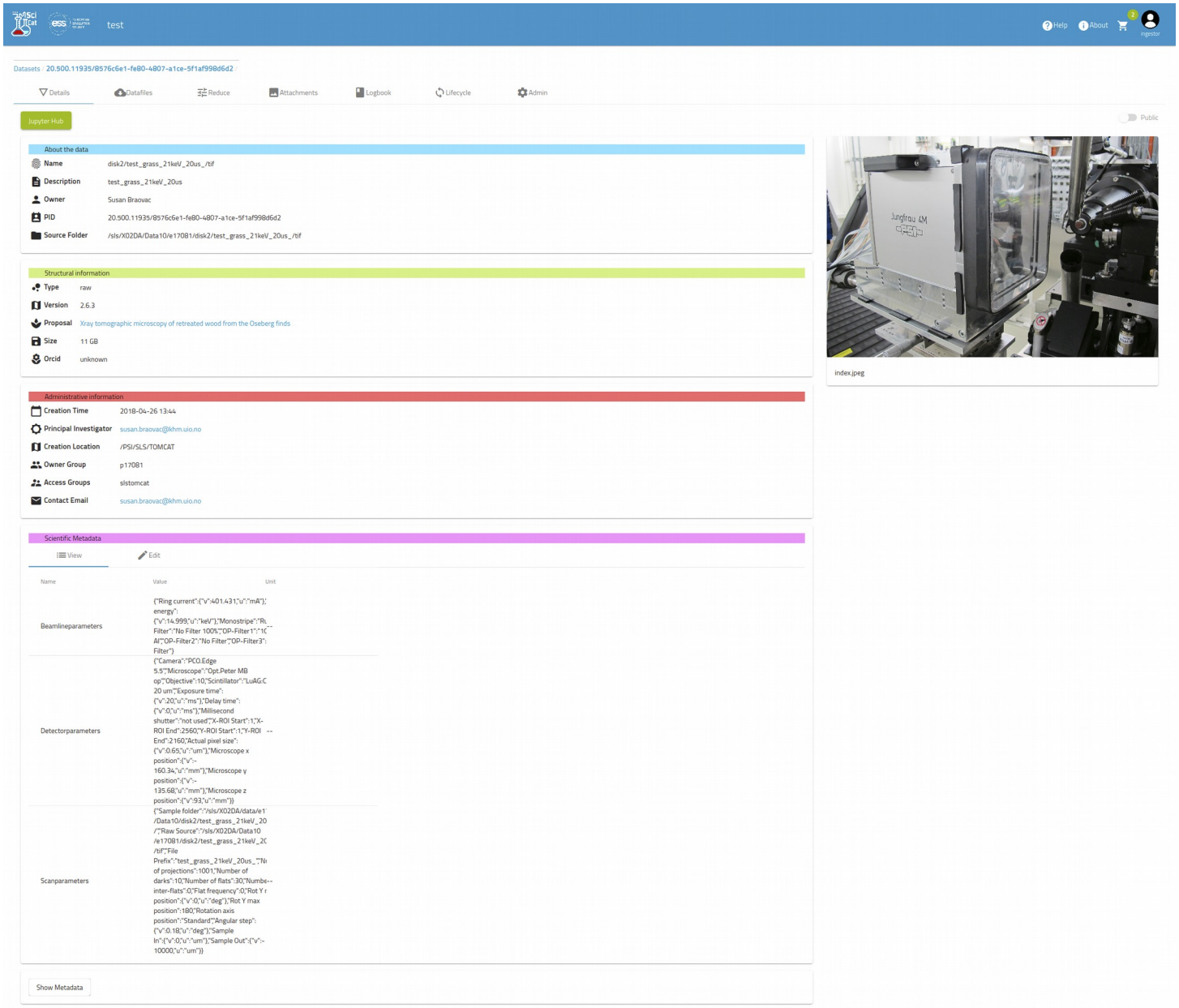
Frontend Login



Dataset Listing and Navigation

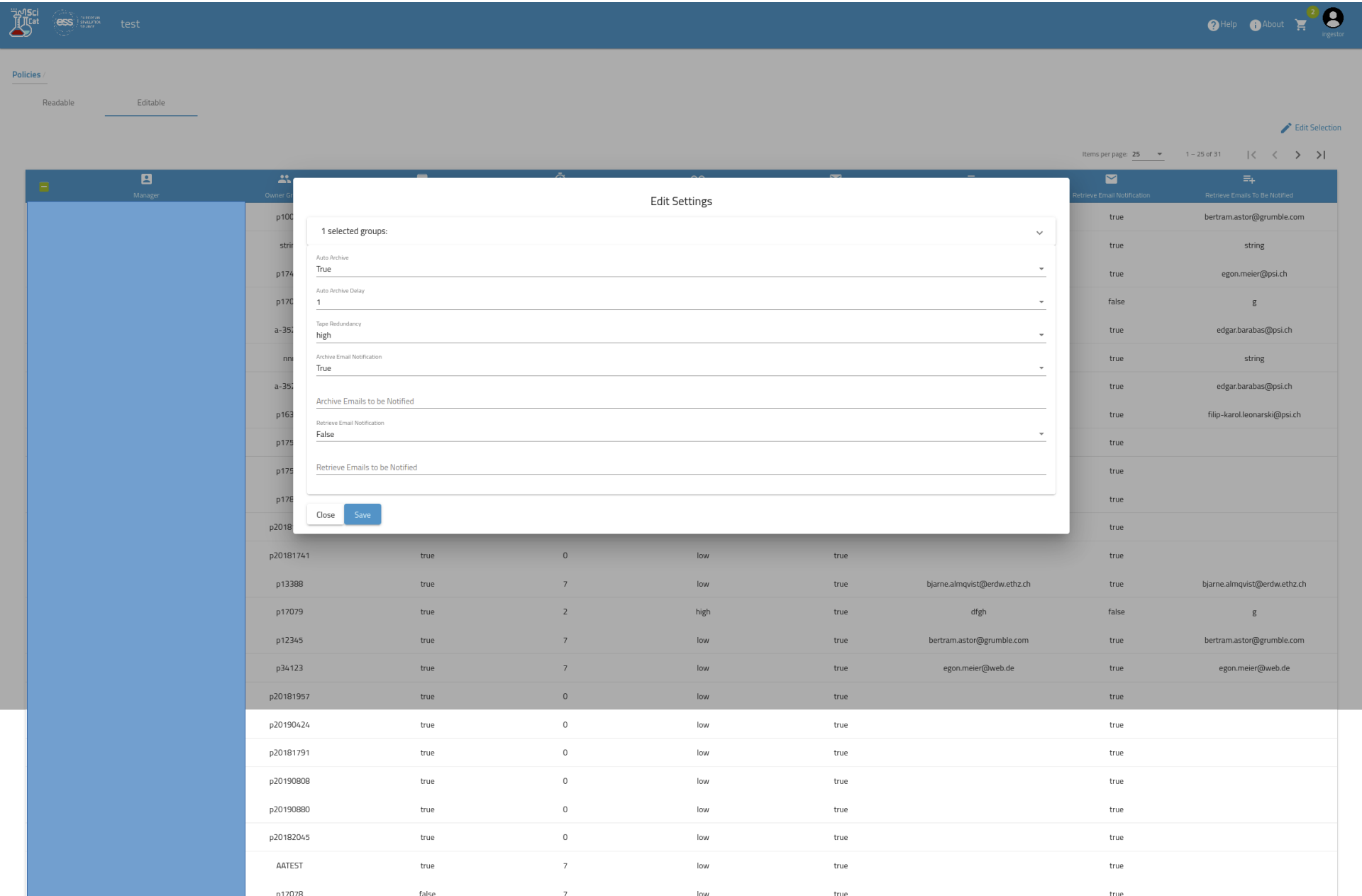


Dataset and Attachments



PAUL SCHERRER INSTITUT PSI

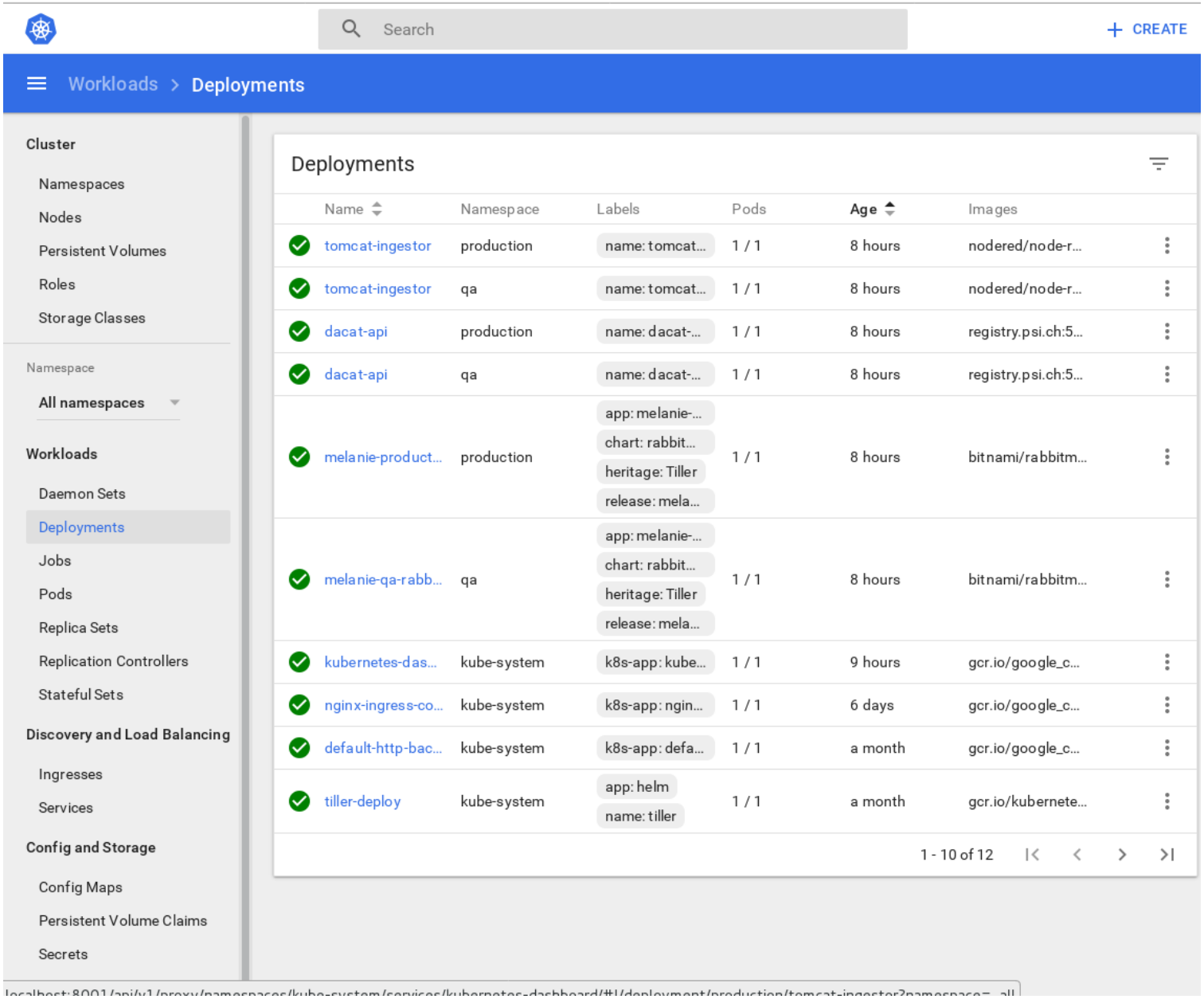
Policy



Management and deployment

- Different deployment options, with and without container technology are available
- Preferred method is via a Kubernetes cluster
- Each microservice is built into a docker file and saved in a registry
- Kubernetes and Helm package manager deploys the containers and handles:
  - Routing
  - Scaling
  - Server failures
  - Updates

Kubernetes Management Dashboard



SciChat – Log Book

lukegorman

Explore Filter

PEOPLE
- henrikjohansson
- stephanegli
- Gareth and henrikjohansson
- Empty room

ROOMS
- scicat
- Gareth

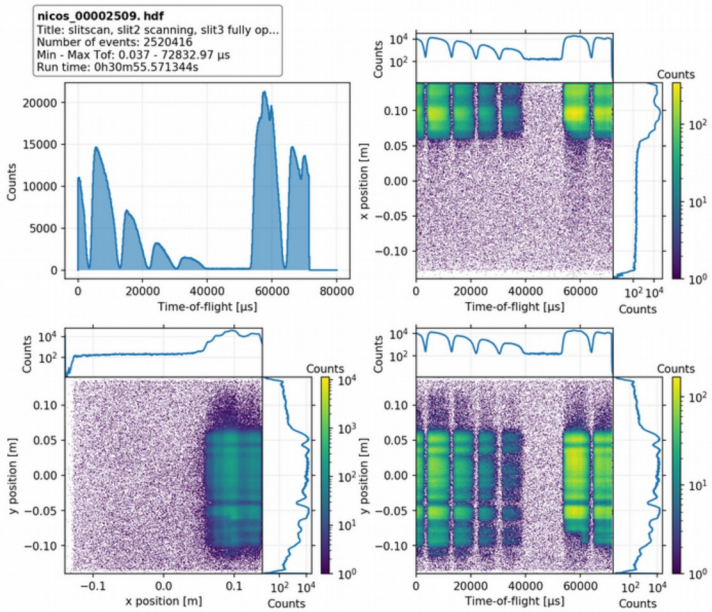
LOW PRIORITY
- YC7SZ5

YC7SZ5 V20 <2019-11-27>

scicatbot

The file /data/kafka-to-nexus/nicos_00002509.hdf was created. See https://scicat.esss.se/datasets/20.500.12269%2F2509nicos_00002509.hdf for details

The file /data/kafka-to-nexus/nicos_00002509.hdf was created. See https://scicat.esss.se/datasets/20.500.12269%2F2509nicos_00002509.hdf for details



Download plot of the data

The file /data/kafka-to-nexus/nicos_00002510.hdf was created. See https://scicat.esss.se/datasets/20.500.12269%2F2510nicos_00002510.hdf for details

Send a message (unencrypted)...

Invite to this room

- tobiasrichter
- henrikjohansson
- fredrikbolmsten
- stephanegli
- michelebrambilla
- Afonso
- Matthew Jones
- @Jonas
- matt.clarke
- lukegorman
- Gareth
- anderspettersson
- scicatbot
- Oliver
- judithhouston
- andrewjackson
- danielnixon
- neilvaytet

Filter room members