Facility Update DESY

DESY specifics in metadata management

Regina Hinzmann, IT-InFa, for DESY.
Input from IT-RIC, IT-InFa, Library, FS-EC, FS-SC
2025-06-24

SciCatCON 2025
DTU Denmark

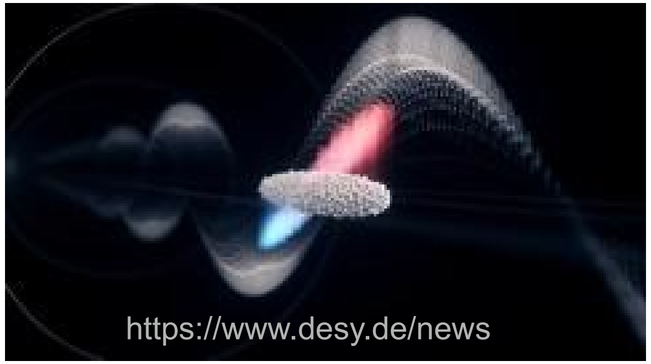
Metadata management at DESY FS

General

DESY serves with its two light sources PETRA III and FLASH a broad scientific community collaborating in many national and international projects and and industrial cooperations.

- More than 3000 researchers visit DESY yearly.
- It operates 23+2 PETRA* III beamlines and 4 for FLASH**.
  - FLASH has a DAQ system to be able to record each high frequency photon pulse while at PETRA beamlines no direct DAQ is needed ⇒ different metadata capturing. Vital for FLASH, as they cannot use their data w/ metadata.
  - PETRA III data : on tape (dCache): a bit more than 27 PB (incl. 1 copy); on disk: 13 PB
  - FLASH data on dCache about 0.4 PB.
- Accelerator research and PETRA IV project are in front of DESY's door.

    (100x higher brilliance)



Laser-plasma accelerated electron bunch show record beam quality.

*Positron–Electron Tandem Ring Accelerator   **Free Electron LASer in Hamburg

Metadata management at DESY FS

Brief history and status

When I joined DESY-IT in August 2023, SciCat was already 2 years around (migration to new backend).

- Main setups at dedicated **demonstrator beamlines**: one at a PETRA and one at a FLASH beamline.
  - At FLASH they used an inhouse metadata system which ceased when the person left, they integrated SciCat as core element - which was not sustainable.
  - At P08 they dedicated beamline staff to write an ingestor and actively persue until today metadata ingestion into SciCat.

Today, 2 years later

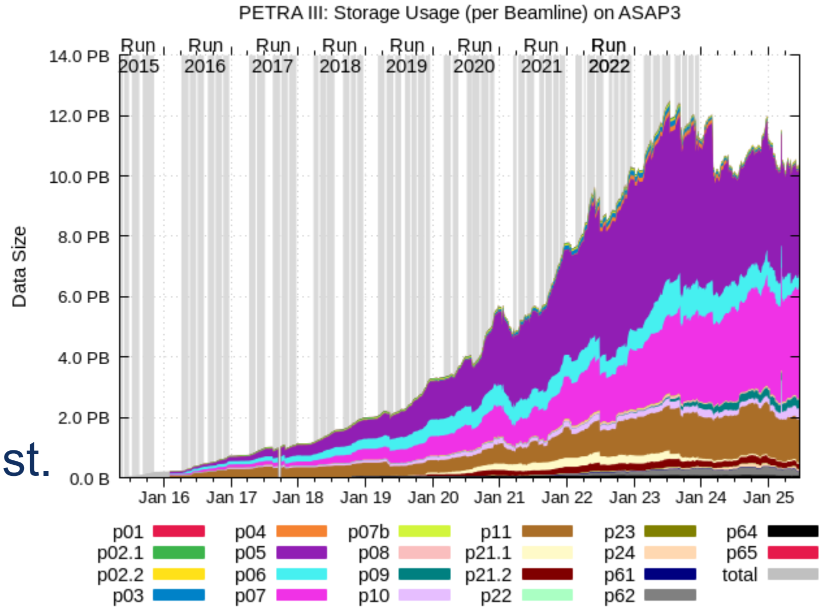
- FLASH had contributed to improvments of SciCat but decided to wait for any metadata catalogue will be provided by DESY. Until then, they are happy to “survive beamtimes”.
- More PETRA beamlines (P08, P05 and P10) work now with SciCat.

(Immediate) usefulness of SciCat@DESY

- Primary: Catalogue functionality (search and find, filter, select), also very important: issue DOIs.

Usefulness of SciCat @DESY

What are our questions to answer?



Increased usefulness of SciCat @ DESY

➢ Service for systematic FS data access.

Which data should be accessible?

- Operation of PETRA has increasingly produced more data over the past.

Who should access the data?

- Depends on ***when?*** Data under embargo? Hard to say w/o data policy. Clear: it's ***not immediate***, but what would be the workflow once the embargo period ends?

- Journals require already ***now*** a DOI; ***reviewer*** may need to access it, ***other scientists*** as well when paper is published they'd like to access the data of that specific DOI.

SciCat: The most promissing tool for DESY?

One goal, two SciCat instances.

Address with two setups of SciCat that run at different timescales for public data.

- IT-RIC deals with ***open data***:

  public-data.desy.de

  provides access to actual data in HiFis dCache

- IT-InFa deals with ***metadata under embargo***:

  scicat.desy.de

  provide catalogue functionality, but also pathway to data publication, first manually, but eventually automatically?

Workflow from embargo to publicly accessible data – independent of past, current and future data taking?

public-data.desy.de
no login required

public-doi.desy.de
detailed landing page with link to DataCite

scicat.desy.de
can only see either published or one own's private dataset records
currently one DB for all beamlines ... (P08, P05, FLASH have 114,000)
doi.desy.de
should only see registered records (with DOI).

SciCat for DESY’s public-data portal

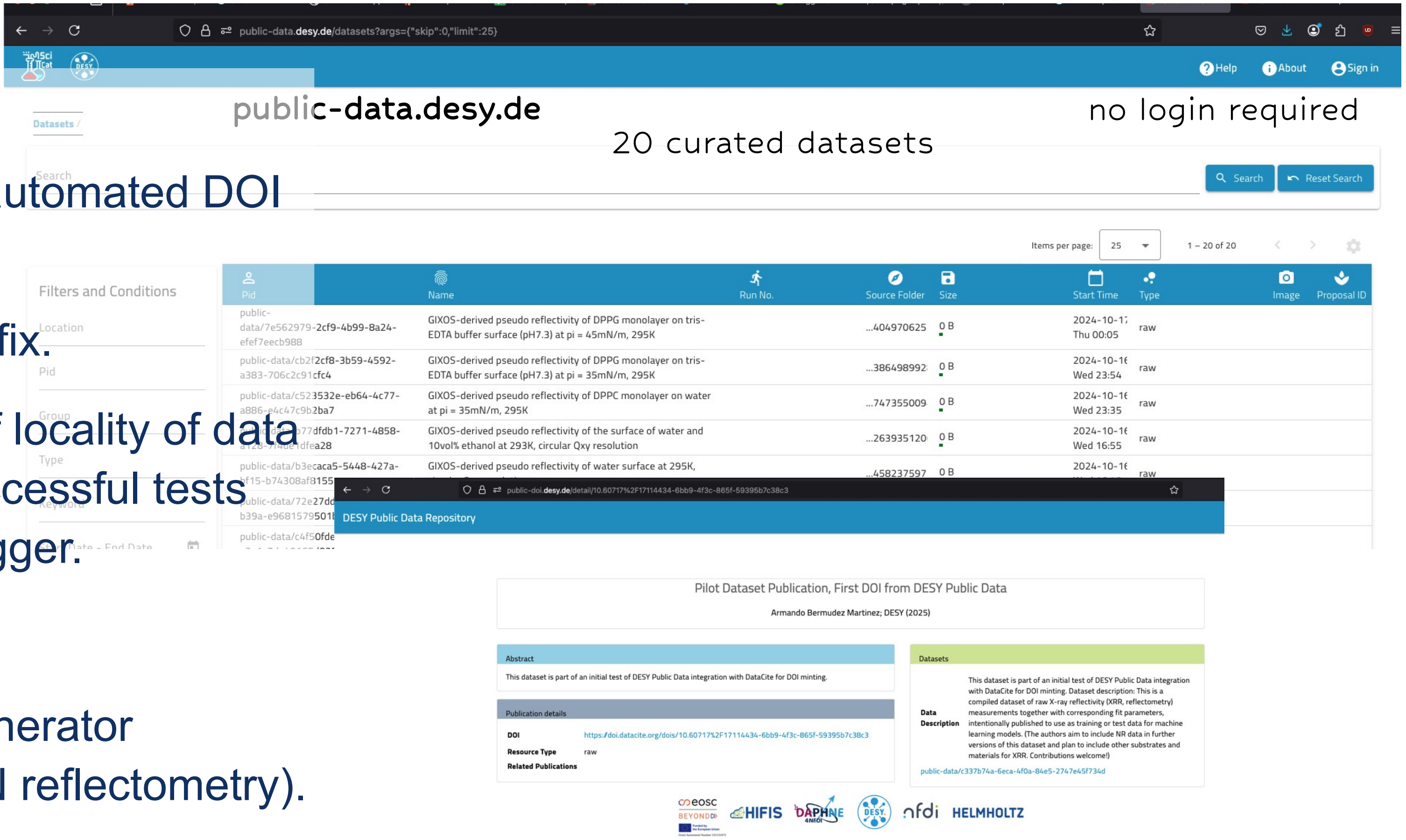
Beyond DESY scope.

Advanced pilot features

- **Manual DOI minting** is set up, as final automated DOI minting as agreed with L is still missing.
  - Uses already productive DataCite prefix.
- Extended landing page by information of locality of data by which enables download by DOI. Successful tests with **download tools** aria2 and DataHugger.

Emphasis on data curation

- Advancement planned of ingest form generator sisyphos (community driven GUI for PaN reflectometry).
- Plan to improve data curation and more user friendly functionalities.



public-data.desy.de/datasets?args={"skip":0,"limit":25}

Items per page: 25 1 – 20 of 20

Pid	Name	Run No.	Source Folder	Size	Start Time	Type	Image	Proposal ID
public-data/7e562979-2cf9-4b99-8a24-efef7eecb988	GIXOS-derived pseudo reflectivity of DPPG monolayer on tris-EDTA buffer surface (pH7.3) at pi = 45mN/m, 295K		...404970625	0 B	2024-10-17 Thu 00:05	raw		
public-data/cb2f2cf8-3b59-4592-a383-706c2c91cfc4	GIXOS-derived pseudo reflectivity of DPPG monolayer on tris-EDTA buffer surface (pH7.3) at pi = 35mN/m, 295K		...386498992	0 B	2024-10-16 Wed 23:54	raw		
public-data/c523532e-eb64-4c77-a886-e4c47c9b2ba7	GIXOS-derived pseudo reflectivity of DPPC monolayer on water at pi = 35mN/m, 295K		...747355009	0 B	2024-10-16 Wed 23:35	raw		
public-data/b77dfdb1-7271-4858-a128-7f4de1dfea28	GIXOS-derived pseudo reflectivity of the surface of water and 10vol% ethanol at 293K, circular Qxy resolution		...263935120	0 B	2024-10-16 Wed 16:55	raw		
public-data/b3ecaca5-5448-427a-bf15-b74308af8155	GIXOS-derived pseudo reflectivity of water surface at 295K,		...458237597	0 B	2024-10-16	raw		
public-data/72e27dd b39a-e9681579501								
public-data/c4f50fde								

public-data.desy.de

no login required

20 curated datasets

public-doi.desy.de/detail/10.60717%2F17114434-6bb9-4f3c-865f-59395b7c38c3

DESY Public Data Repository

Pilot Dataset Publication, First DOI from DESY Public Data

Armando Bermudez Martinez; DESY (2025)

Abstract

This dataset is part of an initial test of DESY Public Data integration with DataCite for DOI minting.

Publication details

DOI	https://doi.datacite.org/dois/10.60717%2F17114434-6bb9-4f3c-865f-59395b7c38c3
Resource Type	raw
Related Publications	

Datasets

Data Description	This dataset is part of an initial test of DESY Public Data integration with DataCite for DOI minting. Dataset description: This is a compiled dataset of raw X-ray reflectivity (XRR, reflectometry) measurements together with corresponding fit parameters, intentionally published to use as training or test data for machine learning models. (The authors aim to include NR data in further versions of this dataset and plan to include other substrates and materials for XRR. Contributions welcome!)

public-data/c337b74a-6eca-4f0a-84e5-2747e45f734d

public-doi.desy.de

detailed landing page with DataCite entry

Usefulness of SciCat @ DESY

DESY FS scope: What are our questions to answer?

Data taking time axis and correlation with storage: systems GPFS (fast access, limited disk) and tape.

t_0: Moment beamtime is started, fix ressources on GPFS: proposal metadata, ownership, access rights, etc.

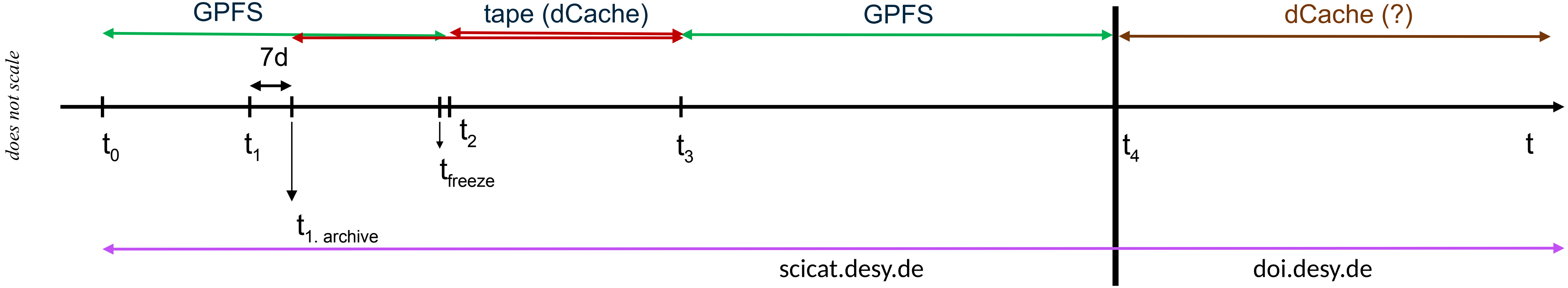
t_1: Moment beamtime is stopped. Data is on GPFS until *dwell timer* expires and not activly accessed (t_2).

$t_{1.\ archive}$: First copy to tape is triggered and data is removed from GPFS.

t_{freeze}: Only a filelist of all data in that beamtime is left in directory on GPFS.

t_3: User requests re-stage from tape to GPFS.

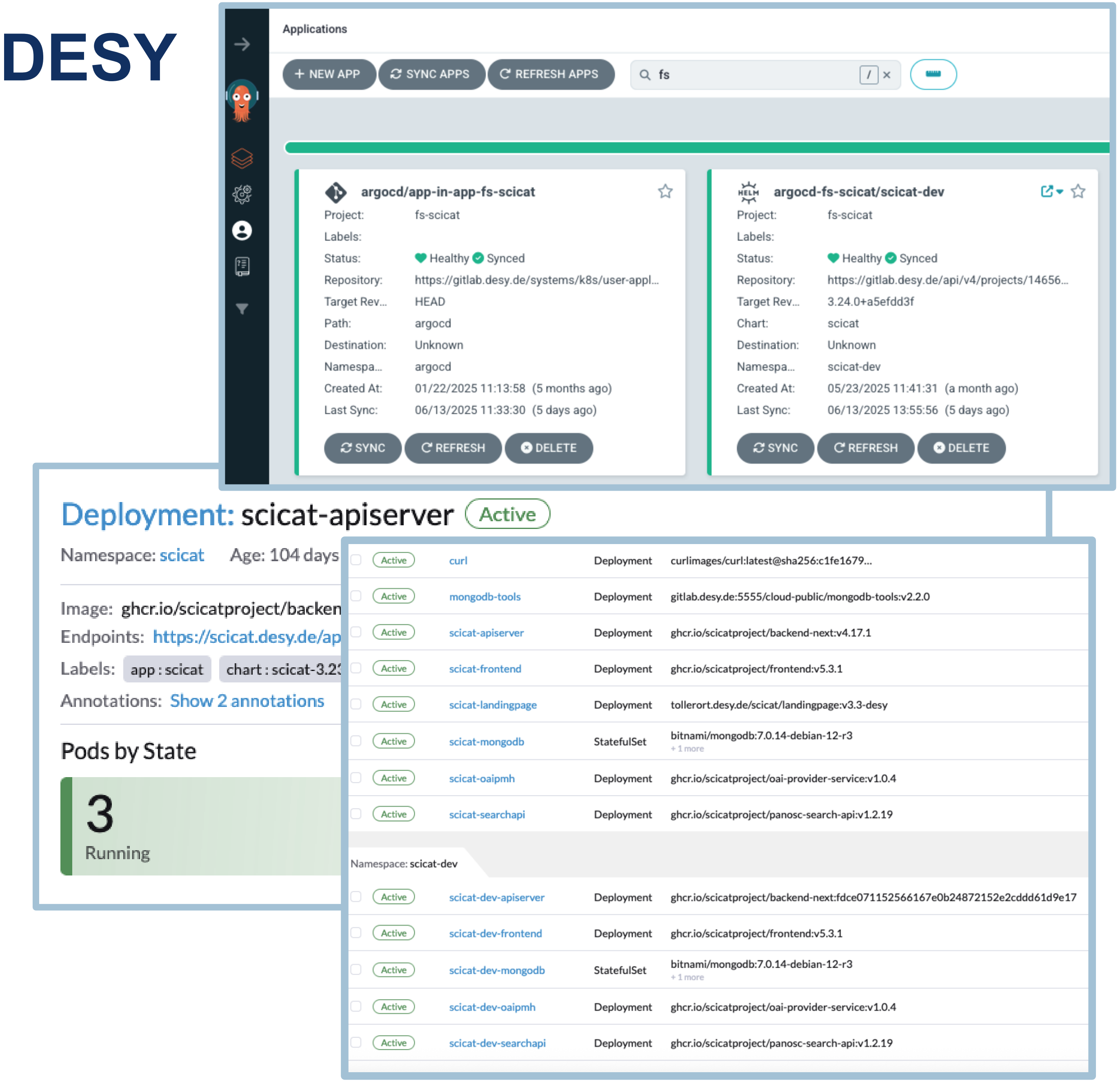
t_4: User publishes and a DOI is minted for the data. Should data be accessible? To whom, for how long?



Systems setup of SciCat @ DESY

scicat.desy.de

- Run (both) SciCat in Kubernetes cluster in using argoCD.
- Own Helm Chart, published to OCI registry oci://tollerort.desy.de/scicat/scicat
- Use official images for prod. Key facts for prod: 3 APIs, 3 MongoDB, 3 traefik ingress controller, 1 frontend.
- Plan more detailed performance studies, see Keynote talk by Igor K.
- Re-design to have separation of embargo-ed and publicly accessible data.



argocd/app-in-app-fs-scicat

Project:	fs-scicat
Labels:	
Status:	Healthy Synced
Repository:	https://gitlab.desy.de/systems/k8s/user-appl...
Target Rev...	HEAD
Path:	argocd
Destination:	Unknown
Namespa...	argocd
Created At:	01/22/2025 11:13:58 (5 months ago)
Last Sync:	06/13/2025 11:33:30 (5 days ago)

argocd-fs-scicat/scicat-dev

Project:	fs-scicat
Labels:	
Status:	Healthy Synced
Repository:	https://gitlab.desy.de/api/v4/projects/14656...
Target Rev...	3.24.0+a5efdd3f
Chart:	scicat
Destination:	Unknown
Namespa...	scicat-dev
Created At:	05/23/2025 11:41:31 (a month ago)
Last Sync:	06/13/2025 13:55:56 (5 days ago)

Deployment: scicat-apiserver Active

Namespace: scicat Age: 104 days

Image: ghcr.io/scicatproject/backen

Endpoints: https://scicat.desy.de/ap

Labels: app : scicat chart : scicat-3.2

Annotations: Show 2 annotations

Pods by State

3 Running

Active	curl	Deployment	curlimages/curl:latest@sha256:c1fe1679...
Active	mongodb-tools	Deployment	gitlab.desy.de:5555/cloud-public/mongodb-tools:v2.2.0
Active	scicat-apiserver	Deployment	ghcr.io/scicatproject/backend-next:v4.17.1
Active	scicat-frontend	Deployment	ghcr.io/scicatproject/frontend:v5.3.1
Active	scicat-landingpage	Deployment	tollerort.desy.de/scicat/landingpage:v3.3-desy
Active	scicat-mongodb	StatefulSet	bitnami/mongodb:7.0.14-debian-12-r3 + 1 more
Active	scicat-oaipmh	Deployment	ghcr.io/scicatproject/oai-provider-service:v1.0.4
Active	scicat-searchapi	Deployment	ghcr.io/scicatproject/panosc-search-api:v1.2.19

Namespace: scicat-dev

Active	scicat-dev-apiserver	Deployment	ghcr.io/scicatproject/backend-next:fdce071152566167e0b24872152e2cddd61d9e17
Active	scicat-dev-frontend	Deployment	ghcr.io/scicatproject/frontend:v5.3.1
Active	scicat-dev-mongodb	StatefulSet	bitnami/mongodb:7.0.14-debian-12-r3 + 1 more
Active	scicat-dev-oaipmh	Deployment	ghcr.io/scicatproject/oai-provider-service:v1.0.4
Active	scicat-dev-searchapi	Deployment	ghcr.io/scicatproject/panosc-search-api:v1.2.19

Outlook

Where will we be in 2 or 5 years?

- Continue with the strategy demonstrator beamlines: gain happy users by providing nice search functionality.
- Work out how to handle past FS data. Investigate “1 BT = 1DS = 1 DOI”?
- Work out how to handle current FS data.
- By then we can have sufficient experience to handle future FS data.
- Thanks to DAPHNE project many important issues could be and are being addressed (proposal hierarchy).
- Dream or FS-IT goal: Provide functional, reliable and useful catalogue and support for DESY FS users beyond project funding.

? How do other labs handle data publication and data access?

Acknowledgments*: **DESY IT** Peter van der Reest, Johannes Reppin, Stefan Dietrich, Dirk Jahnke-Zumbusch, Martin Gasthuber, Philipp Neumann, Jürgen Hannapel, Frank Schlünzen, Tim Wetzel, Armando B. Martinez, Tobias Klann, Yves Kemp. **FS:** Jan Kotanski, Igor Khokhriakov, Linus Pithan, Florian Betram, Bridget Murphy, Anton Barthy, Edgar Weckert. **L**: Martin Köhler, Kirsten Sachs, Robert Thiele. **DAPHNE4NFDI** Nicolas Hayen (CAU).*

Thank you

for your attention ❤️

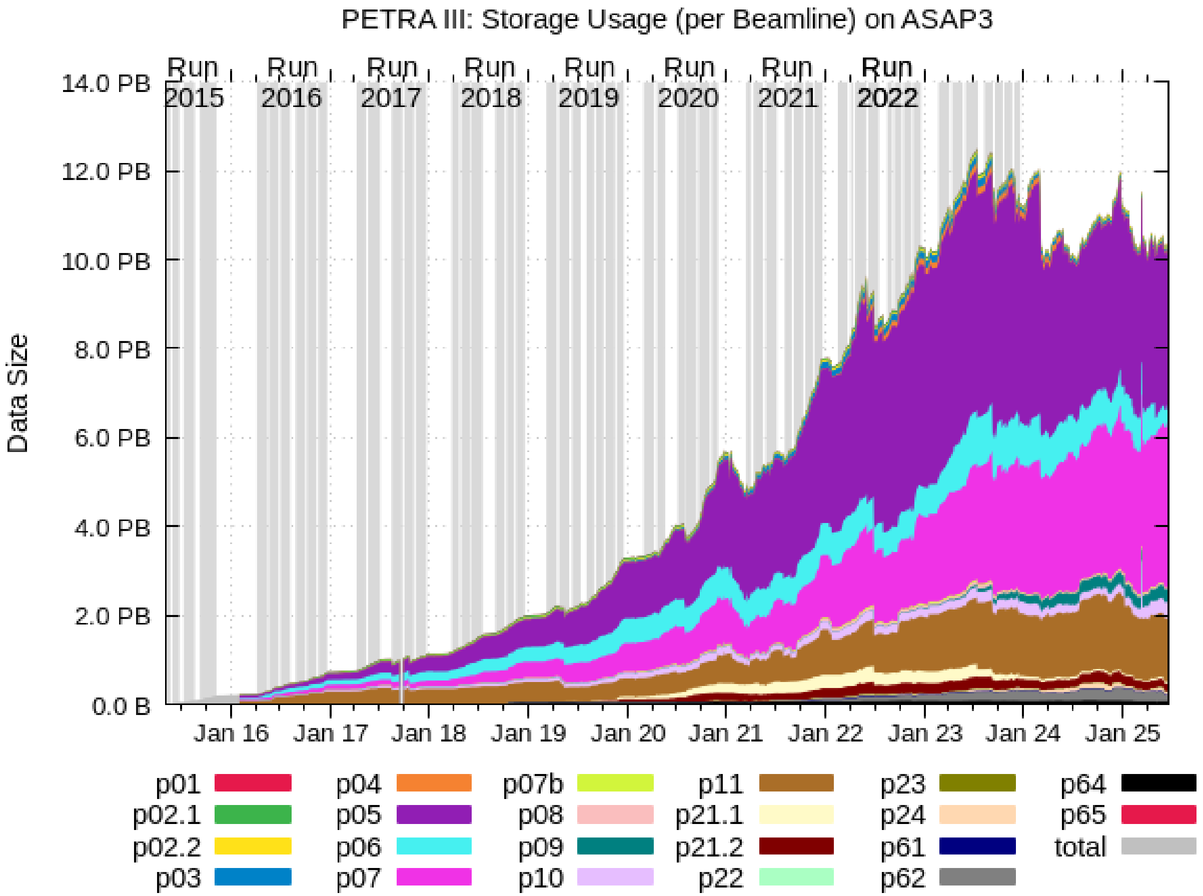
List of issues

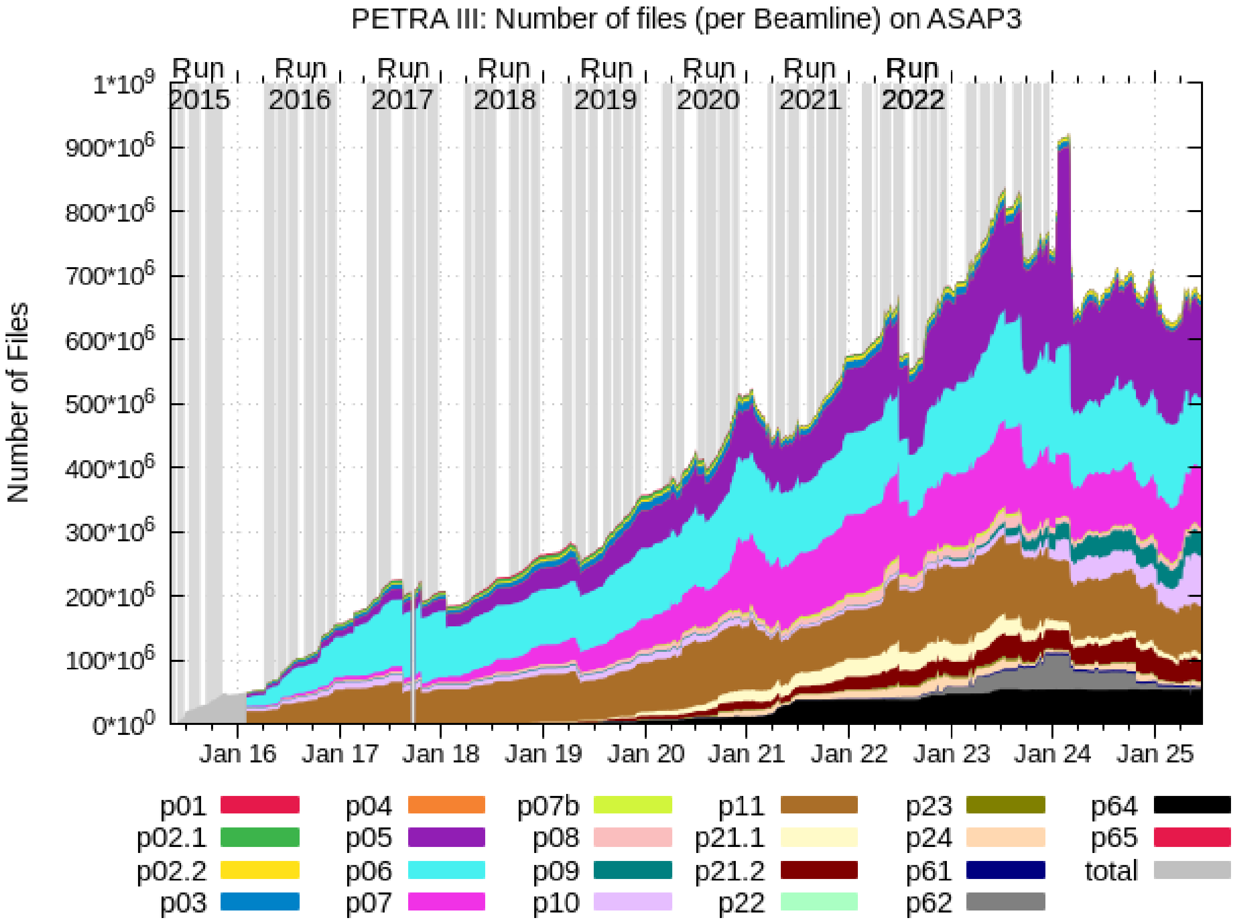
Catalogue functionality:

- Elastic Search
- Scientific metadata search

Data Taking Rate

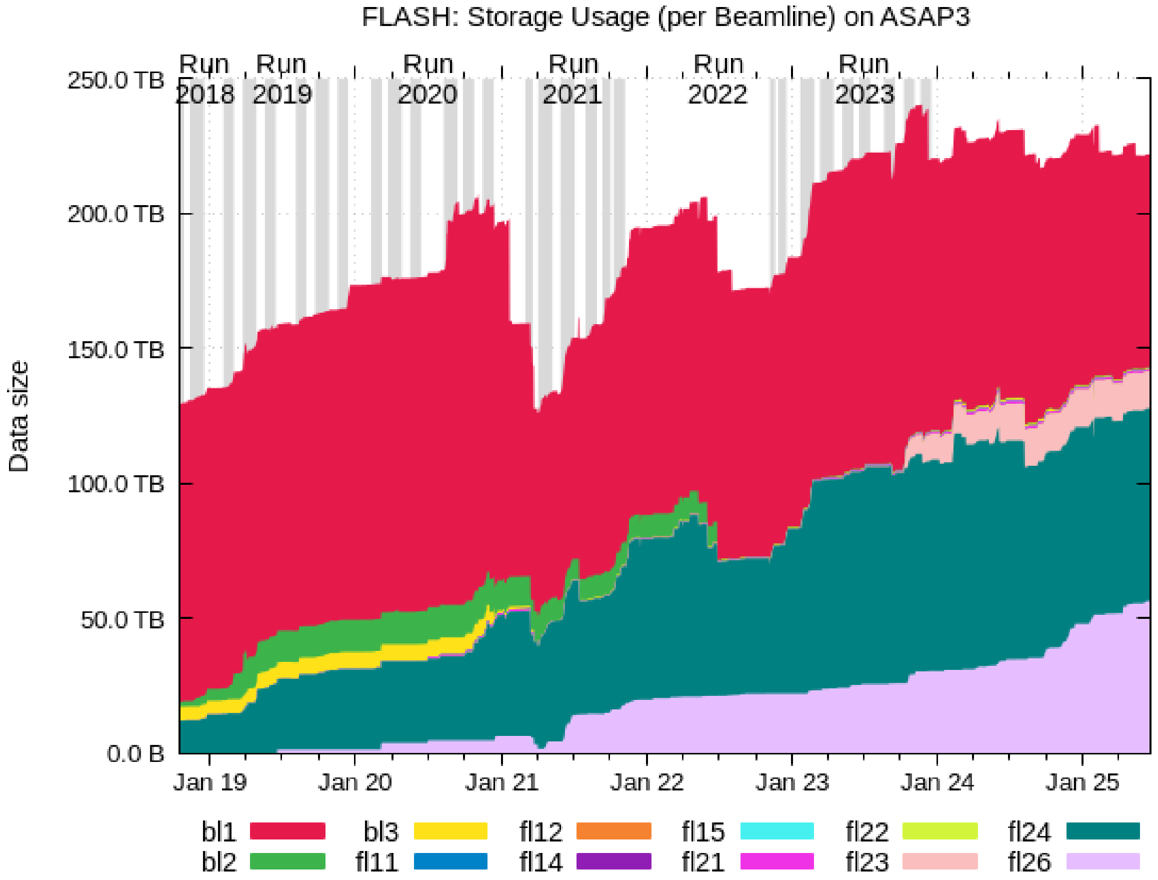
PETRA data on GPFS

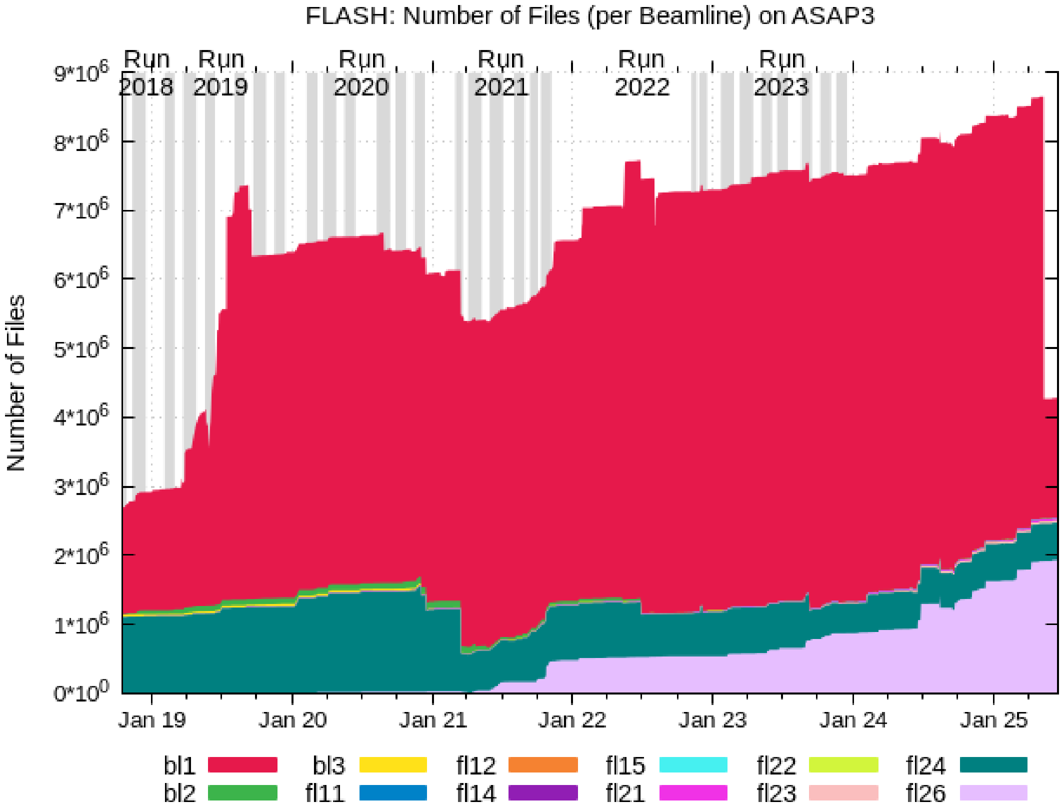




Data Taking Rate

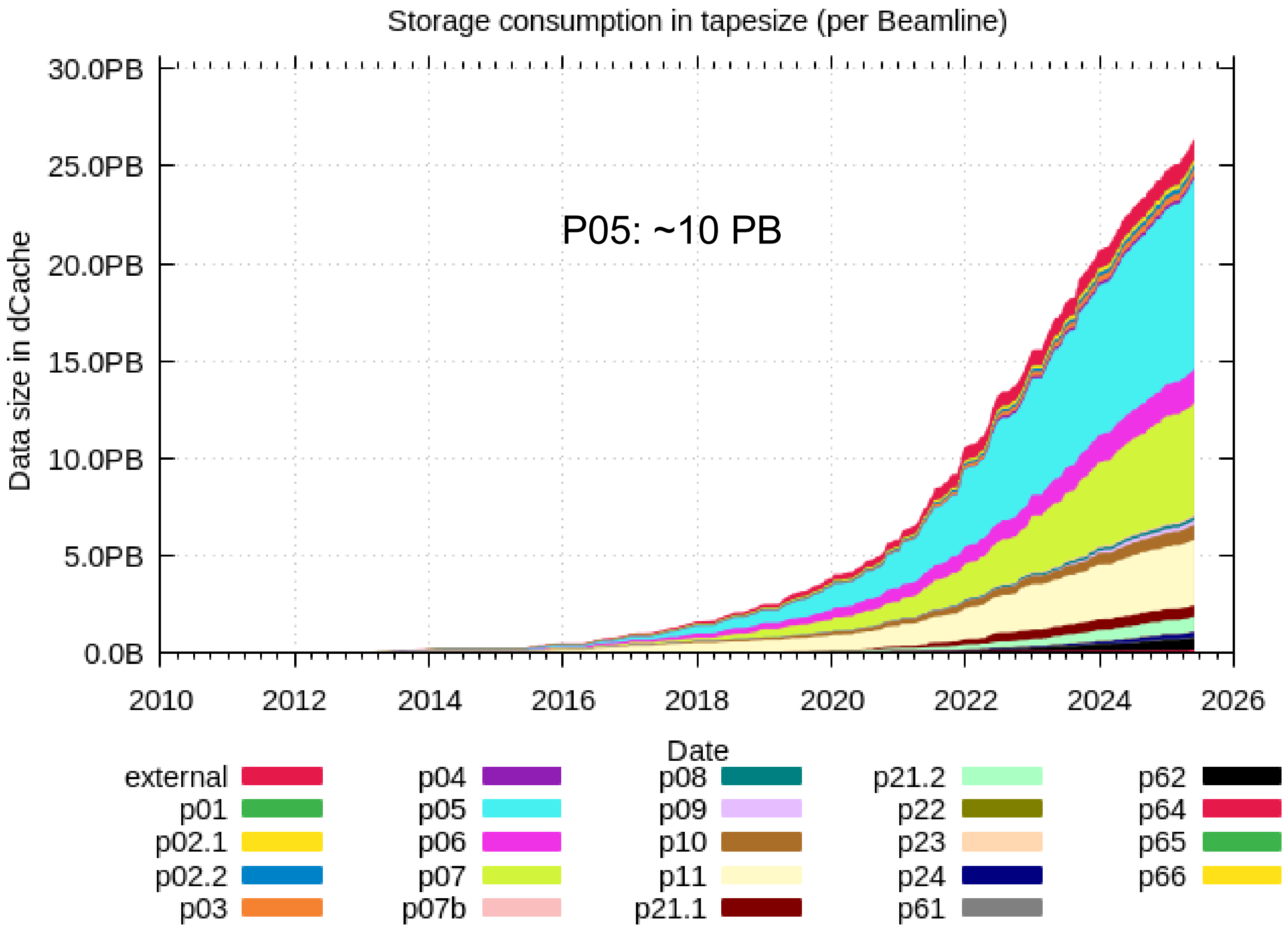
FLASH data on GPFS



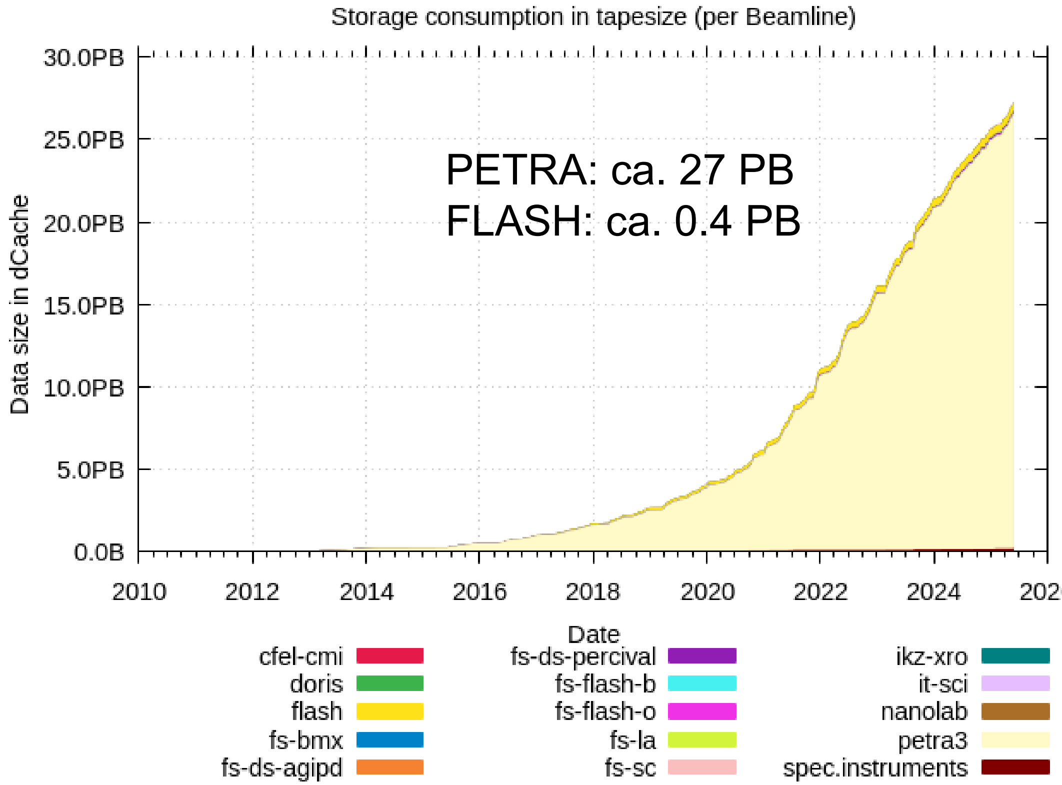


Data on Tape (dCache) incl. 1 copy

Data on tape per beamline



P05: ~10 PB



PETRA: ca. 27 PB
FLASH: ca. 0.4 PB