Where are we with SciCat @DESY?

And where do we want to go in 2024? A summary and plan.

Regina Hinzmann

Hamburg, 27.11.2023

SciCat @ DESY

Why?

Driving forces: DESY needs a catalogue

to make DESY data ***FAIR***. This covers the entire data life cycle, from the moment of data taking to the end of its life cycle.

At DESY, it was decided to go with SciCat now two years ago as other labs within Photon Science already use it and since there is a broad community behind it.

SciCat @ DESY

Touched a number of topics within the past 3 months

Identify the missing pieces and single steps for SciCat to be useful to the users.

- Investigations of
  - Who is the user?
  - What is the status now?
  - What would be really cool to have?
- Started hands-on
  - Reproduced together errors.
  - Brought forward particular issues, which have been worked on remote.
  - Getting more familiar with the scicat code.

Things begin getting clearer!

What did we achieve sofar …

A recapitulation

What did we do over the past 3 to 4 months?

Kick-off SciCat meeting on 7th September 2023

- Created a frame for all scicat acitivities: SCG (3) + SCT (6)
- Strategy to **start with beamlines at DESY** is well underway: setup 4 new test instances, 5th is to come (on top of 4 actively used instances) of 13.
  - P08 started ingesting in the new instance
  - FLASH is in the middle of transition from old to new backend
  - Addressed general issues (e.g. measurementPeriodList, authentication issues)
  - DAPHNE participation in general SciCat developments
  - Started to addess DOI provision (FS and DESY library agreed. Next step: implementation of the service that talks to SciCat and DataCite, execise workflow)

… and where are we heading to?

outlook for 2024

For the first good half of 2024:

- Make SciCat a stable service even for single beamlines
  - Follow up on experiences and practices from other labs (as presented on 20.11.2023 in SCG, needs agreement, work in progress)
  - Work on frontend/search of SciCat (committment by Igor K)
  - Federated login: user management, integration of federated user accounts within DOOR, the DESY internal proposal management system (DOOR developers agreed to work on that)
  - Provision of DOIs (FS, Linus P, managed to get an agreement with L, need to prioritise resources for implementation)
  - Provide documentation (ideally by all the questions users have, please feedback to me!)
- Make it the access point to browse/download/access data (currently done by the Gamma-Portal, will cease)
- Leave space for emerging other problems not yet thought through but need to be addressed too.

By the end of 2024 have at least sketched a

- Set-up and run a benchmark performance test (functionality, stability, reliability, scalability) in Continuous Integration and Continuous Deployment
- A roadmap from > 10 TEST to regular PRODUCTION instance(s) (embargoed and open data)

Technical details Momentaufnahme

An overview of activities

Helpful problems

01 P08

–**MeasurementPeriodList**: get empty field when ingesting complete proposal, *upstream a fix (fpotier) - to be tested*

– **ProposalIngestor**: ingestor cannot read proposal ingested by proposalIngestor, *being worked on upstream*

– **Too many data**, specific to P08: many scans are performed within a short time (minutes), not scalable, result is not searchable, scicat is not used. *Alternative data ingestion is being tested.*

–**migration**

02 FLASH

–Login as user **cannot access the data**

–**migration**

03 DOIs

– needed by FS, **FS joined forces with DESY library and IT**. Pathway is being worked out.

04 Jobs

–Development work has a use case for DESY

–Started with URLs, needs extensions to Kafka

05 Tests

–Documentation of how to run a test of SciCat code

06 Documentation

Demonstrator beamline P08

Helpful problems

MeasurementPeriodList:

Empty list when ingesting a proposal data

With Stefan Dietrich I spent time to investigate this problem. We could trace back the reason. Found an upstream fix - to be tested.

How to reduce the vast amount of data ?

We needs to reduce the data stored in SciCat, also for scalability reasons. Ansatz solution

- A parallel test instance is being setup
  - goal: test if one can ingest data in an alternative way by reducing metadata, requires input from the party that does the data analysis, followed up within FS (Linus P)

Frontend search issues will be addressed in the next release.

Also from external talks (SCG 20.11.) they presented a **user customized interface**, to be seen if useful for DESY, too.