Where are we with SciCat @DESY?

Updates, status reports and involvements: PP and DDD.

Regina Hinzmann

Hamburg, 28.10.2024

P10 Follow Up

since Moras presentation last week

Major points to get scicat instance up at P10 :

1. Ingestion part by Jan K

    1. started scicat ingestion at P10, now P10 scientists need to define what should go in.

    2. found an error due to renaming of the instance which has been fixed (ingest also prosposalIDs)

    2. IT part by Johannes R: solved error at login and keykloak forwarding

3. Mora and Fabian W work to

    1. get their desired meta data in (key parameters are transmission and exposure times), iterating with the users to fix a first draft of meta data.

    2. produce a schema in LinkML in order to fit into automatic data validation (in contact with resp. P08 person, Philipp Jordt)

P08 visit outcome

2024-10-16 during CAU (Kiel) beamtime

P08 SciCat misses some features to be really useful to them so they actually use it.

- They want to identify certain datasets/scans that form e.g. a "reflectivity" or a "time scan", consisting of SciCat datasets, ie several nexus or fio files (several points in a measurement plot). A button to group these files and a possibility to tag them would be great.

- Then, for publication, several such reflectivity pre-selected datasets if identifiable, they'd like be able associate them to the paper. Example: In a publication, they have 10 reflectivities and 3 time scans.

- Functionality is desired to group already grouped datasets and tag those too, so they can find all associated datasets by seeing 10 + 3 entries and more details are shown when they select one of them.

- Improved search functionalities:

  - An AND combined query (==applying several filters) is missing in the scientific metadata. Example ScanCommand: can one enable queries by just selecting pattern of the fields in that ScanCommand? This has the potential to help significantly!

- Ideally, one would be able to issue a DOI for these grouped datasets and download them.

big thank you to CAU team, in particular Svenja H and Bridget M

User documentation

Migration of official SciCat documentation for Users

Official User documentation: links are broken, page navigation broken, links to videos are broken, etc.
Discussion started almost a year ago. Carlo did a great job to update documentation of scicatlive. Conclusion sofar: Have 3 categories of documentation: 1 User 2 Site-Admin/Operator 3 Developer.

repository in github

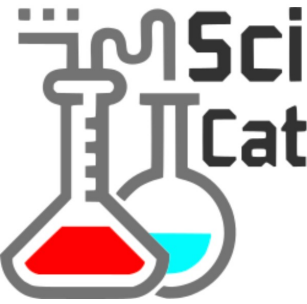scicatlive (Public)
Simple getting started

Jupyter Notebook ☆ 9 11 10 ⊙ 16 11 1 Updated 4 days ago

user-documentation (Public)
For all the user facing documentation. Developer documentation will sit alongside the code

☆ 0 ⚖ Apache-2.0 11 0 ⊙ 1 11 0 Updated last week

current view

SciCat Metadata Catalogue

Build Status

Why SciCat ?

See the SciCat Home Webpage for an overview of how SciCat can help to manage scientific data.

SciCat User Manual

Welcome to SciCat Users Manual

This is a short guide to most common questions users of SciCat can face. First of all: What is SciCat? And why should I use it? SciCat, a *science catalogue*, should serve you to find back your data, a bookkeeping tool accompanying some critical steps during the entire data life cycle which are: getting an overview of datasets, data analysis, re-analysis, publishing datasets, publication. In SciCat, only the meta data is stored, descriptions to identify a certain measurement. We highlight most prominent features that SciCat offers, but note that SciCat is a general software layer which when integrated with site-specific application develops its full potential.

Features: quick links to How-To's

Your data may be of type raw or derived, you may want to login or just browse what's in the catalogue. Here is how to find more on how to proceed:

- Login
- Search and find your data
- How to change some fields after ingestion
- How to view history of changes to a dataset
- How to group and tag datasets
- How to group and tag grouped datasets

Structure of SciCat

These main classes determine the functionality of SciCat:

1. Datasets: This class is the most elaborated one.
2. Proposals: are used to link datasets to the proposal under which beamtime was granted.

local prelimary preview

DAPHNE contribution to SciCat Development

first meeting with Max Friday

Plan was to start task 1 (*18d*),

see <u>doc</u> or <u>agenda</u>

Refactor frontend (FE) to use new autogenerated typescript SDK.
This will align FE to the current backend (BE) API code leveraging BE features and increasing maintainability.
Produce stable release.

5 subtasks.

conclusion first week

- ahead of time wrt refactoring code

- tests to be defined (95% use cases, edge cases)

Subtasks:

1. explore changes needed.
   Estimation: 2d - ***Completed: 1d***
2. review and validation of findings
   Estimation: 1d - ***Completed: 0.5d***
3. create roadmap
   Estimation: 2d - ***Completed: 0.5d***
   See in appendix 1 for plan's details
4. frontend code refactoring, CI tests passing
   Estimation: 8d – ***In Progress***
   1. Updated Angular to the latest version and also all the packages
   2. First working draft with new SDK
   3. Starting with second pass: fixing details and data typing
   4. SDK version strategy:
      1. stable release will pull the correct SDK version from NPM repository
      2. for development purposes, user can decide which version they would like to use. Documentation on options and how to implement them will be provided.
         Options:
         1. build SDK locally from local copy of BE
         2. select specific version from NPM
5. review current frontend tests, listing missing tests and add relevant ones
   Estimation: 5d

DOI status and plans

Since last report in pre-summer 2024

Several fronts:

1. Too big dataset for 1 DOI: with Johannes R we could reproduce the error, tried changing some configuration of the cluster, started trying ingesting to new scicat cluster where we can exclude disturbances from other processes. Need to do that more systematically.

2. Will meet Tim who started workflow of issuing DOIs with DataCite for his public datasets. Plan is to discuss the general workflow and setup how to design such a service using e.g. scicatlive. Evaluate other possibilities. Prototyping means for me

   1. We issue DOIs without direct data access, but providing info that data is available on request and details of contact

   2. We publish meta data of datasets of an entire beamtime first, without indicating which datasets are less useful while waiting for implementation of (-- as outcome of FS and L agreement)

      1. functionality to select and cluster a subset of datasets of a beamtime (user awareness condition of L)

      2. mapping more SciCat meta data fields according to DataCite (BE #1192)

Summary

Clear goal remains: DESY wants to offer a useful SciCat service to the user

- Many fronts in working progress towards our goal

  - probably there is much more behind the scenes I'm not aware of, esp. by FS/Linus P: need of regular exchanges.

  - I experience good IT support (InFa, Systems, RIC).

- DESY's involvement in SciCat developments is increasing

- also pushed by DAPHNE involvement

  - on-site travel forseen in cw 48 or 49 (between last and first week of this and next month resp.)

- First ideas for 2025 features:

  - Data access

  - security of SciCat