Updates from DESY

News, status and achievements, reoccurring problems and perspectives.

Regina Hinzmann for SciCat team at DESY
2024-07-02

HELMHOLTZ

DAPHNE 4NFDI

DESY.

Overview

Reports.

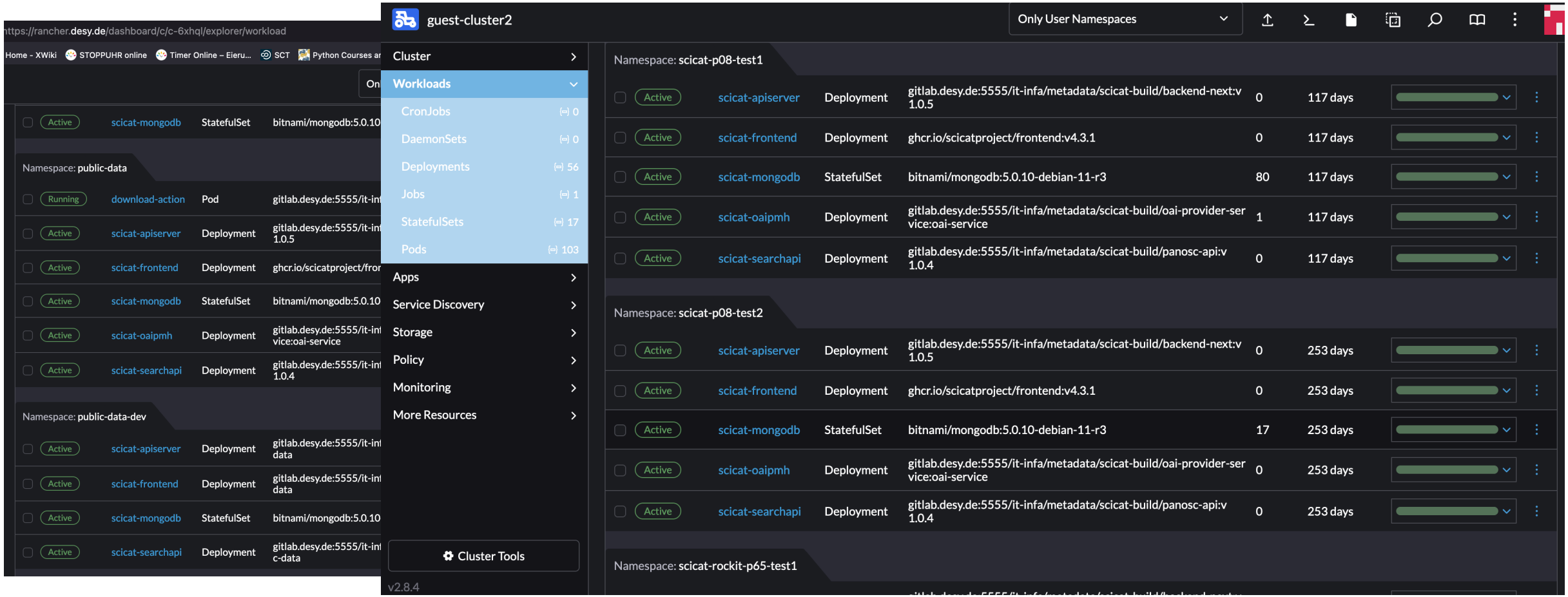
- Our achievements wrt to last year.
- Our plans with SciCat at DESY.
- Our challenges at DESY.
- Some points to discuss.

Highlights of last year’s activities

Summary of past year's activities

Achievements – what do we have now ?

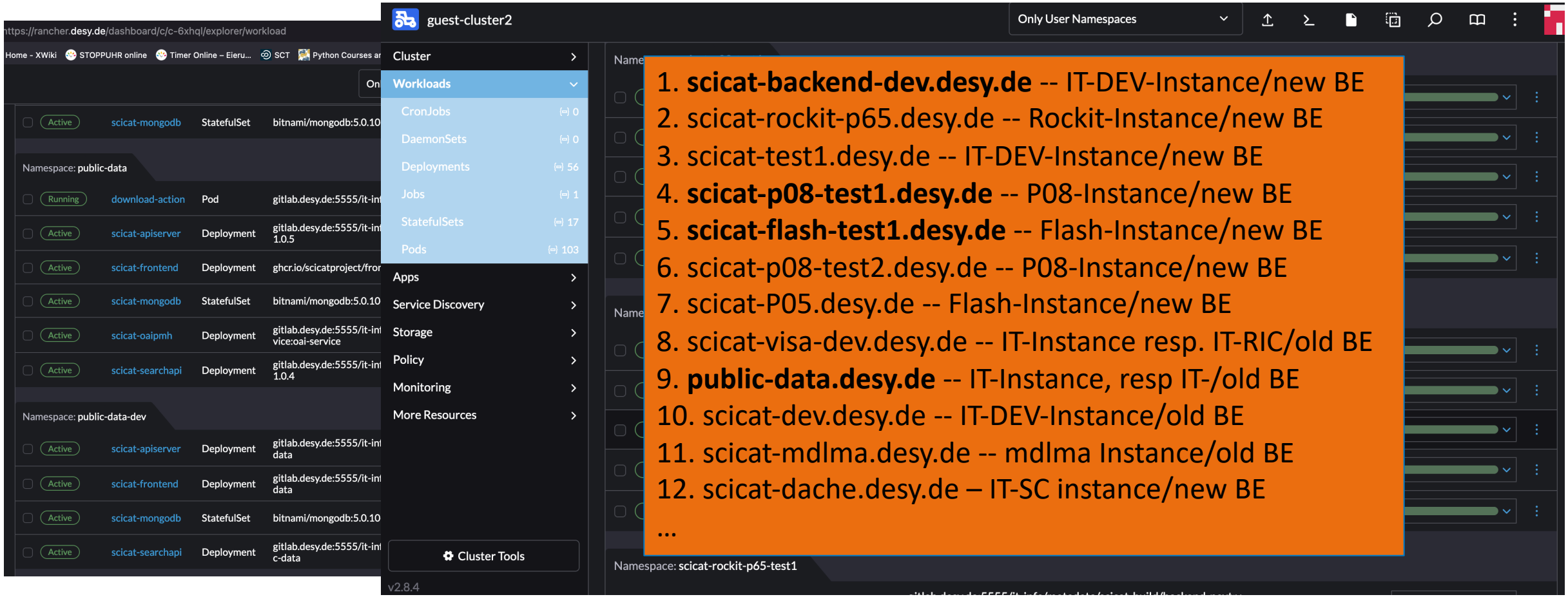
A CI/CD as been setup to **mirror official stable releases of SciCat repos**, several SciCat instances for beamlines are **managed in a Kubernetes OpenStack** environment. They run smoothly with other services: **authentication via keycloak**, federated login (Helmholtz AAI), distributed storage using CEPH, etc.



Summary of past year's activities

Achievements – what do we have now ?

A CI/CD as been setup to **mirror official stable releases of SciCat repos**, several SciCat instances for beamlines are **managed in a Kubernetes OpenStack** environment. They run smoothly with other services: **authentication via keycloak**, federated login (Helmholtz AAI), distributed storage using CEPH, etc.



Achievements

Current status

Our goal is: *make SciCat useful to the user*.
Tackle problems individually, gain experience, build up expertise.

1. internal purposes **2. external world**

1. Test instances at the beamline: **demonstrator beamlines**
- **@Petra III and @FLASH**
- Work in progress for **a setup common for all beamlines.**

2. FS test instances for more **general tasks**
- **DOI minting**: Different DESY groups (IT, FS and L- library) **collaborated to agreed** on a service to gain experience in the workflow
- **Public data** at DESY (IT-RIC): a **pilot service** is being set up that combines DESY and Helmholtz products (dCache HIFIS storage) with SciCat.

1. **scicat-backend-dev.desy.de** -- IT-DEV-Instance/new BE
2. scicat-rockit-p65.desy.de -- Rockit-Instance/new BE
3. scicat-test1.desy.de -- IT-DEV-Instance/new BE
4. **scicat-p08-test1.desy.de** -- P08-Instance/new BE
5. **scicat-flash-test1.desy.de** -- Flash-Instance/new BE
6. scicat-p08-test2.desy.de -- P08-Instance/new BE
7. scicat-P05.desy.de -- Flash-Instance/new BE
8. scicat-visa-dev.desy.de -- IT-Instance resp. IT-RIC/old BE
9. **public-data.desy.de** -- IT-Instance, resp IT-/old BE
10. scicat-dev.desy.de -- IT-DEV-Instance/old BE
11. scicat-mdlma.desy.de -- mdlma Instance/old BE
12. scicat-dache.desy.de – IT-SC instance/new BE

…

What's new since last year

We have regular meetings!

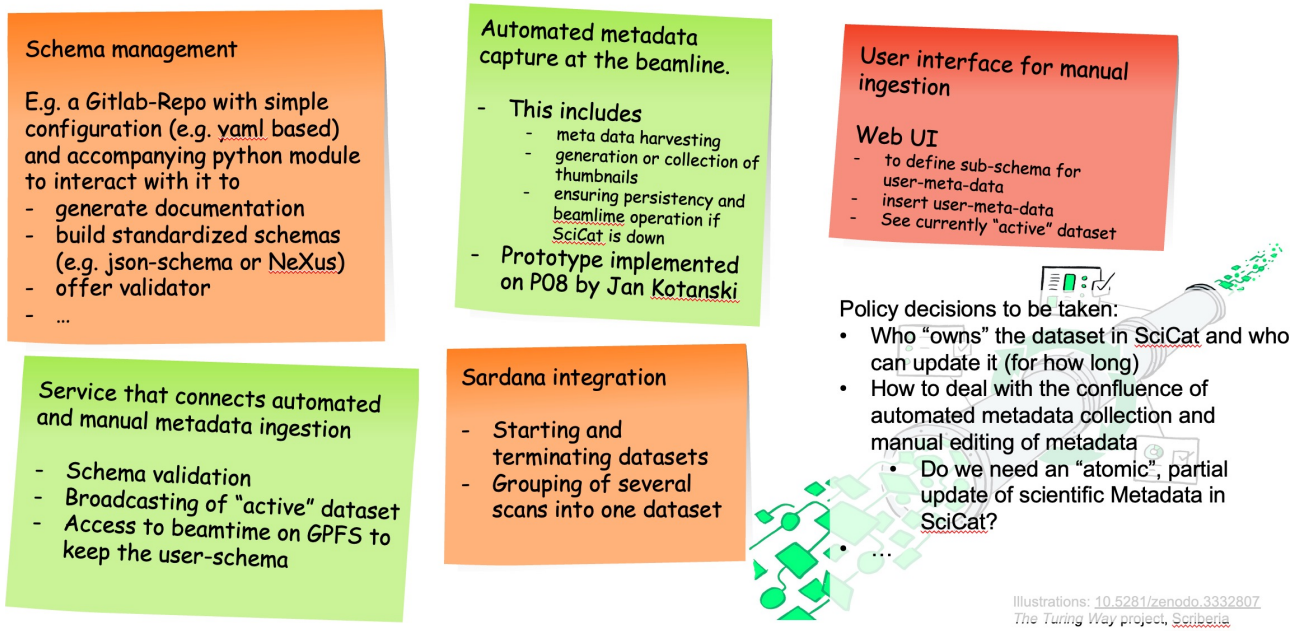
We run **regular** a meetings on metadata topics, specifically for SciCat

- Weekly: SciCat Technicals, SCTs, run by IT, complemented by
- Monthly: SciCat General meetings led by FS and IT.

Results

- Provide room for discussions and questions regarding SciCat features.
- Have a frame on which decisions are not only taken but also followed and progress is monitored.

Ingredients for the data ingestion pipeline at DESY



Slide by Linus Pithan @SCT on 2023-12-11, see indico.desy.de/category/1045

Where does DESY want to go in the next year? Our plans.

DESY FS and IT Plans for SciCat in upcoming year.

make SciCat useful to the user

All work targets concretely this goal: **Have SciCat run and operated in production**.

We'd like to have an idea how to setup **SciCat such that it becomes available at all DESY beamlines**. Readiness for PETRA IV (DESYs next generation synchrotron).

- **This year's main goals**
  - An example: Make it easier to find datasets (currently use `grep` of part of filename strings in `ls` directory). Thanks to elastic search this feature makes our users happy. ✓
  - Setup of a DOI minting service for datasets just like PSI has done.

  Very helpful collaboration with PSI -- **Thank you, Carlo**! ✗

- **Next years goal**:
  - Now (2024) they are not yet connected amongst each other.
  - Set up of a performant, reliable system requires installation of monitoring tools *NOW*. Users do complain about *"slow down due to the currently massif metadatasets".* Need of quantified figures. Install monitoring tools.

  Infrastructure of performant and reliable systems (helm, Kubernetes, OpenStack + monitoring tools).

Some more of our challenges.

Is this DESY specific?

Topics which the same user community is facing?

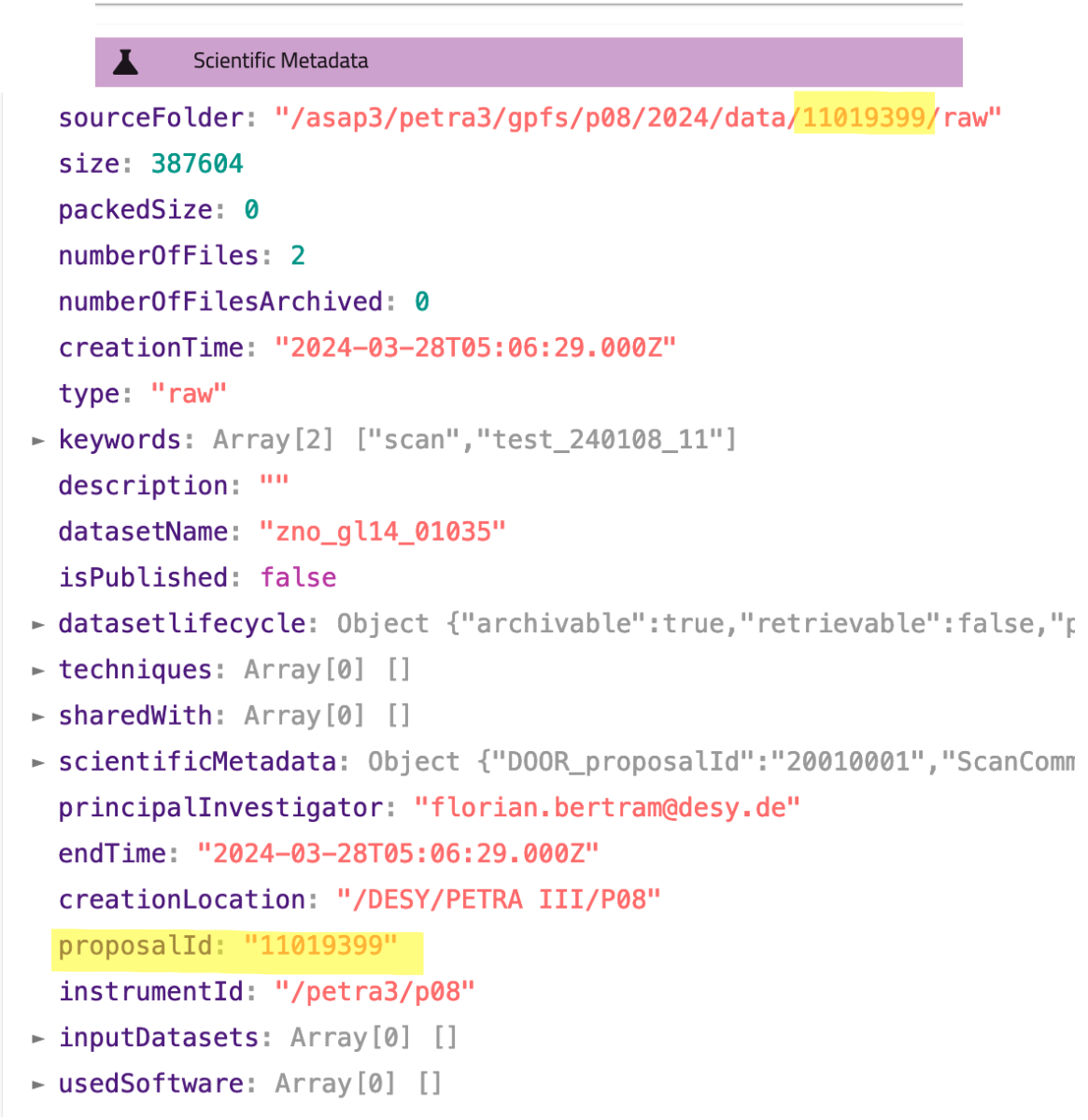
Data structure at DESY is **1:N**

Usually several beamtimes per accepted proposal are allocated. The typical workflow to take data at PETRA is

- User register in our proposal system, DOOR, and
  1. Choses between six types of proposal if not more and submits it.
  2. If accepted, **receive 1 proposal ID**.
     1. **N** (N >1) beamtimes are allocated
     2. Provide **info**: PI, who else, what sample (sample declarations) etc **per beamtime ID**

- Data taken is sorted according to beamtime ID, proposal ID is not useful for finding back data.

- In SciCat has proposal ID is more prominent. Sofar implemented work-around: a DESY mapping to SciCat names, looks like



Several proposed solutions.

Topics which the same user community is facing?

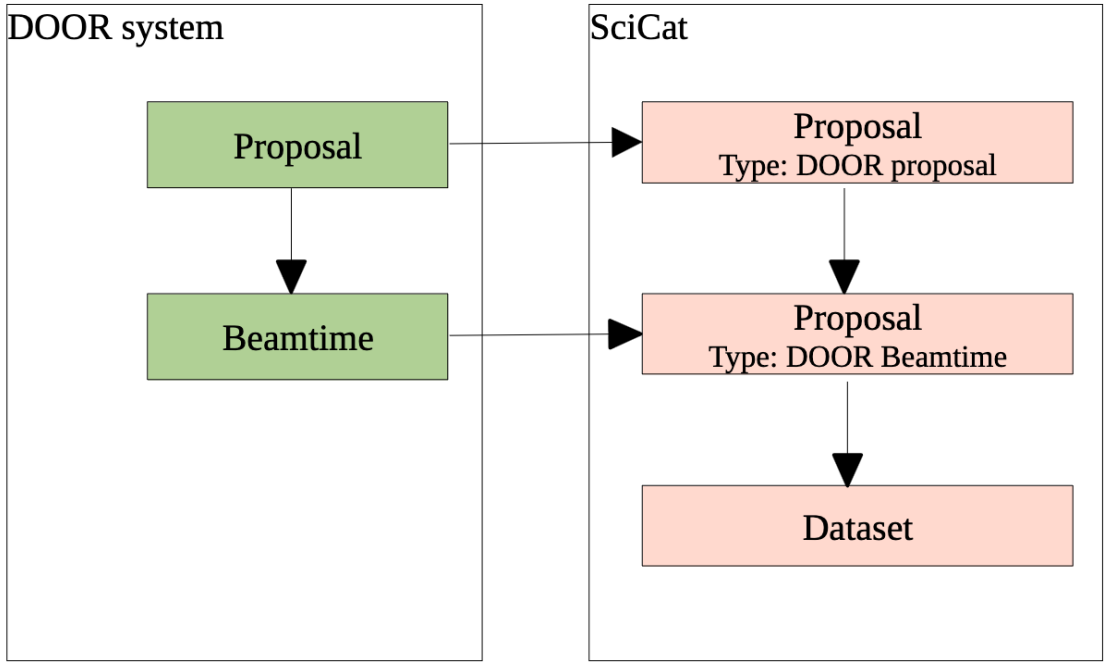
But this confuses users… -**Thank You, Max!**

- Update **`proposalClass by`** adding 2-3 new fields

  1. (Type: can go to properties)
  2. Relationships: relationships with other proposals.
  3. Properties
     unconstrained list of user defined properties. Similar to the datasets relationships.

Was presented yesterday to SciCat audience at DESY.
Apparently it was already addressed in April –
THANKS Frederic Potier!

Plan is to make use of it and implement two types as properties (3.).



General very long term proposal

Topics which the same user community is facing?

Configurable frontend labels – VERY LONG TERM solution
Allows localisation, different languages.

DESY needs to

- Create github issue
- Wait for current release is completed.
- Seek for resources
- Plan, design, implement

Summary of discussion points

How do other labs…

- Priorties for this and next year

1. Landing page prototyping
2. Github issues (most relevant maybe atomic patches)
3. Performance monitoring large scale setup.

Thank you

- …view some of our issues mentioned, how useful would they be for you? Eg.
  - update of proposal class,
  - introduction of DatasetCollection,
  - extend DOI fields according to DataCite
- Do we want a **PID/DOI also for proposals**?
  - Would it have an impact on dataset DOIs?
  - What is scientifically significant for having PIDs for proposals?
- Where do other labs see the role of a ***data curator***?
- We have a lot in the pipeline with very little expertise (and man power) yet.

Thank you

Extra slides

(extraction of github issues)

01 contains_string, 2024-05-28, Linus P

#138, NEW BE #119,OLD BE #680, FE #974

- Intial work dates back to Jul/Aug 2022 and fell right into the change to the new backend:

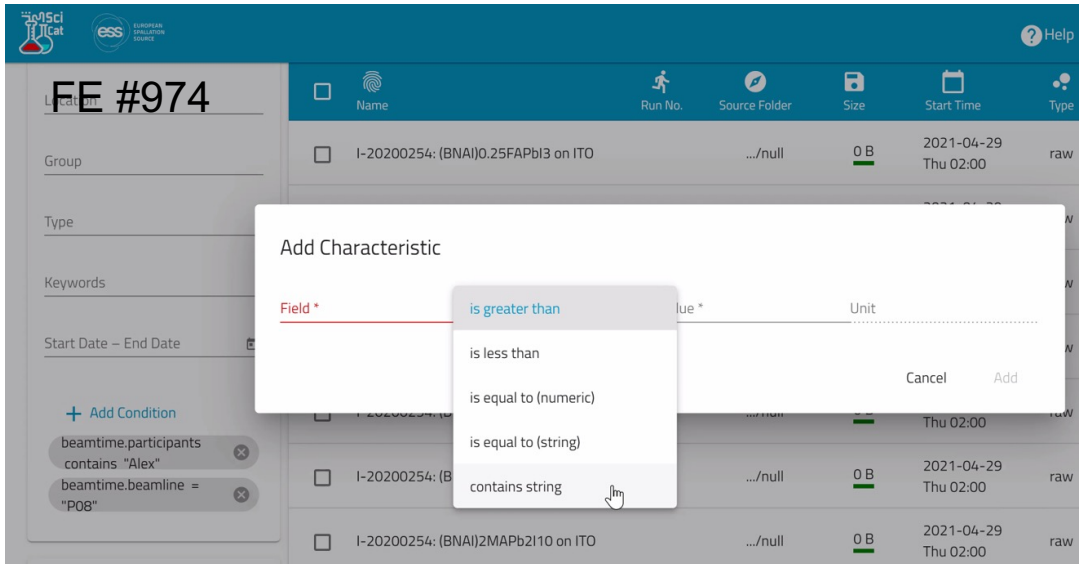
  #119: Partial matching of strings in generateScientificExpression (#680 in old BE).

  #138: discussion of implementation of a general helper function able to deal with different options of values, values and unit, values and optional unit. Problem is that this field may appear at several places and a superior function would be beneficial.

  Person needed to code.

  # 974 related frontend changes

  Person needed to code (Igor can do that after his search UI release is out, end of Nov 24).

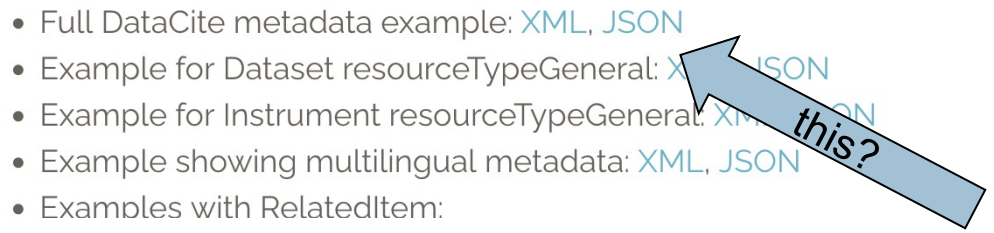


03 publish data in standard schema, 2024-04-25, Paul M

#1192 in BE

Summary

- SciCat has the concept of published data; that is, a set of one or more datasets that, collectively, are described by certain metadata fields. This metadata description is stored as a MongoDB document with the class PublishedData.
- The backend has the ability to map this information to DataCite's XML schema, but only does this when making DataCite API requests for DOI activity. This ability to map PublishedData to a corresponding DataCite XML description isn't exposed by a SciCat API.
- Perhaps because of this lack of exposing the DataCite description, the oai-provider-service reimplements the same mapping functionality (albeit not completely consistently). […]



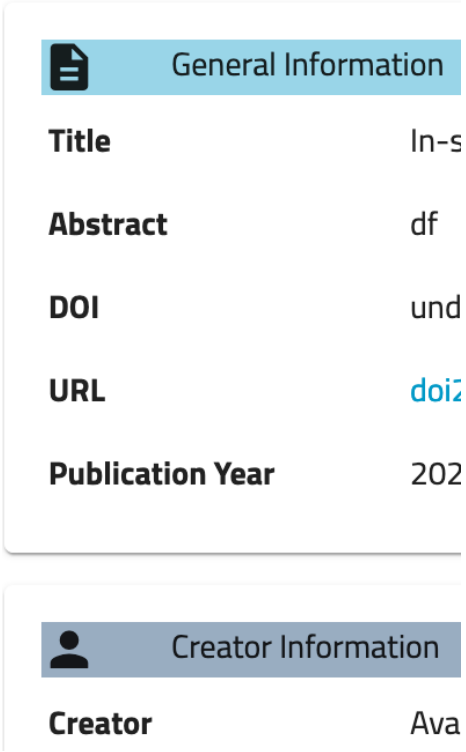
03 publish data in standard schema, 2024-04-25, Paul M

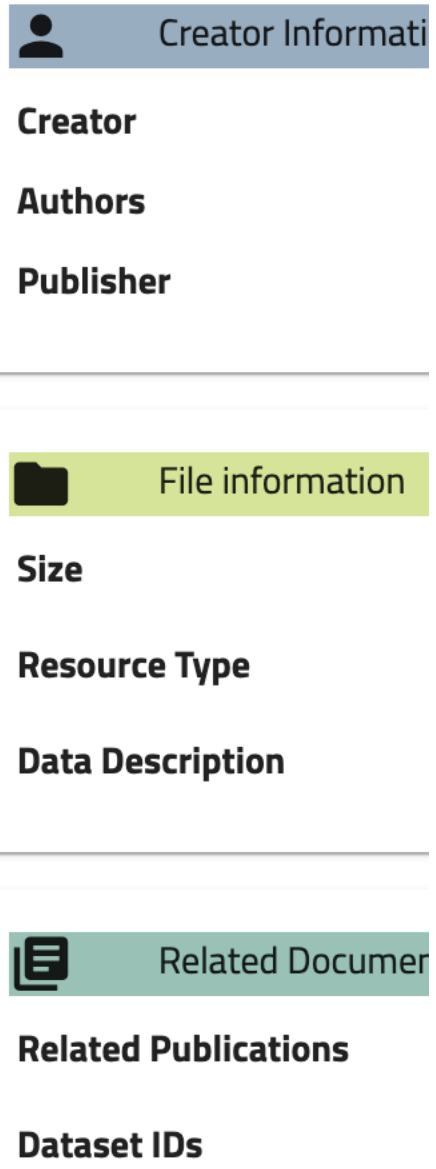
#1192 in BE

- We have these fields available
- There can be more, that is what Paul requested in this issue
- Examples: other PIDs (persistent identifiers) than DOIs are
  - Id for author ORCID
  - Id for research topic
  - Id for sample ISGN
  - Id for proposal
  - …

Nice-to-have but not a showstopper to setup infrastructure *now*.

How do other labs think about it?





04 add schemaless field in proposalclass, 2024-04-17, RH

#1178 in BE and #1455 in FE

- This was raised by Linus: like Dataset class has scientificMetadata, a field of free format, Proposalclass missed this feature. After requesting it, fpotier implemented the BE part. Now the FE is missing.
- Needs person to code. Igor has only from autumn onwards capacities.