Current work areas with and for SciCat

Status on DOOR IDs and DOIs related work for SciCat

Regina Hinzmann
2024-04-08

HELMHOLTZ

DESY.

Announcements

General updates

SciCat developer meetings

- Igor, Noel and me join regulary these biweekly meetings
  - Gitub issues are discussed
  - Design decisions are taken
  - Next releases: 1. Release jobs, tests (by Spencer and me), 2. Search UI (Igor)
  - Other work areas: SciCatLive (Carlo M), Documentation (Laura S)
- Created indico page with meeting notes by Max Novelli: https://indico.desy.de/category/1079/

SciCat Conference 2024, July 2nd to 4th, at PSI

https://indico.psi.ch/event/15861/

- Institute use cases (DAY 2), I will give an update for DESY
- My plan: collect input from topics relevant for DESY
  - Flexibility in naming fields
  - Partial patching of scientific metadata fields (#954)

PLEASE, IF YOU HAVE A CASE, COME and SEE ME! .. so I understand the priority within all the tasks and fronts we currently work on!

List of TO DOs

Tasks and fronts

We want SciCat to be useful to the user. This means:

- Deploy on all DESY beamlines
- Ensure stability and performance
- Have a standardised way of selecting meta data
- Maintain and provide support for current instances, ie document present setup.
- Mint DOIs
- Port download functionality over to SciCat
- Contribute to SciCat development

- …

DOOR business

Identifiers of proposals and beamtime in SciCat

Different vocabular entagled

Problem

Naming of some fields within SciCat are **unfavourable** (potential source of confusion):

- There are many **types of proposals** in DOOR. One can have - more often than in the past - per proposal ID multiple beamtimes, thus multiple beamtime IDs: 1:N, proposalID:beamtimeID.
- To find the data, the proposal ID from DOOR is not really relevant outside of DOOR. Thus, it was decided in 2022 by FS-SC to just **use the beamtime ID in the SciCat field proposalID and have DOOR_proposalID separtely**.

Scientific Metadata


```
sourceFolder: "/asap3/petra3/gpfs/p08/2024/data/11019399/raw"
size: 387604
packedSize: 0
numberOfFiles: 2
numberOfFilesArchived: 0
creationTime: "2024-03-28T05:06:29.000Z"
type: "raw"
► keywords: Array[2] ["scan","test_240108_11"]
description: ""
datasetName: "zno_gl14_01035"
isPublished: false
► datasetlifecycle: Object {"archivable":true,"retrievable":false,"p
► techniques: Array[0] []
► sharedWith: Array[0] []
► scientificMetadata: Object {"DOOR_proposalId":"20010001","ScanComm
principalInvestigator: "florian.bertram@desy.de"
endTime: "2024-03-28T05:06:29.000Z"
creationLocation: "/DESY/PETRA III/P08"
proposalId: "11019399"
instrumentId: "/petra3/p08"
► inputDatasets: Array[0] []
► usedSoftware: Array[0] []
```

Identifiers of proposals and beamtime in SciCat

Possible solutions

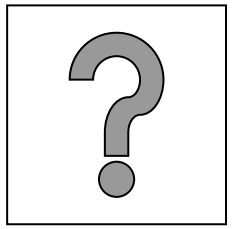
1. Modify fields in **`ProposalClass`** and **`MeasurementPeriodListClass`**

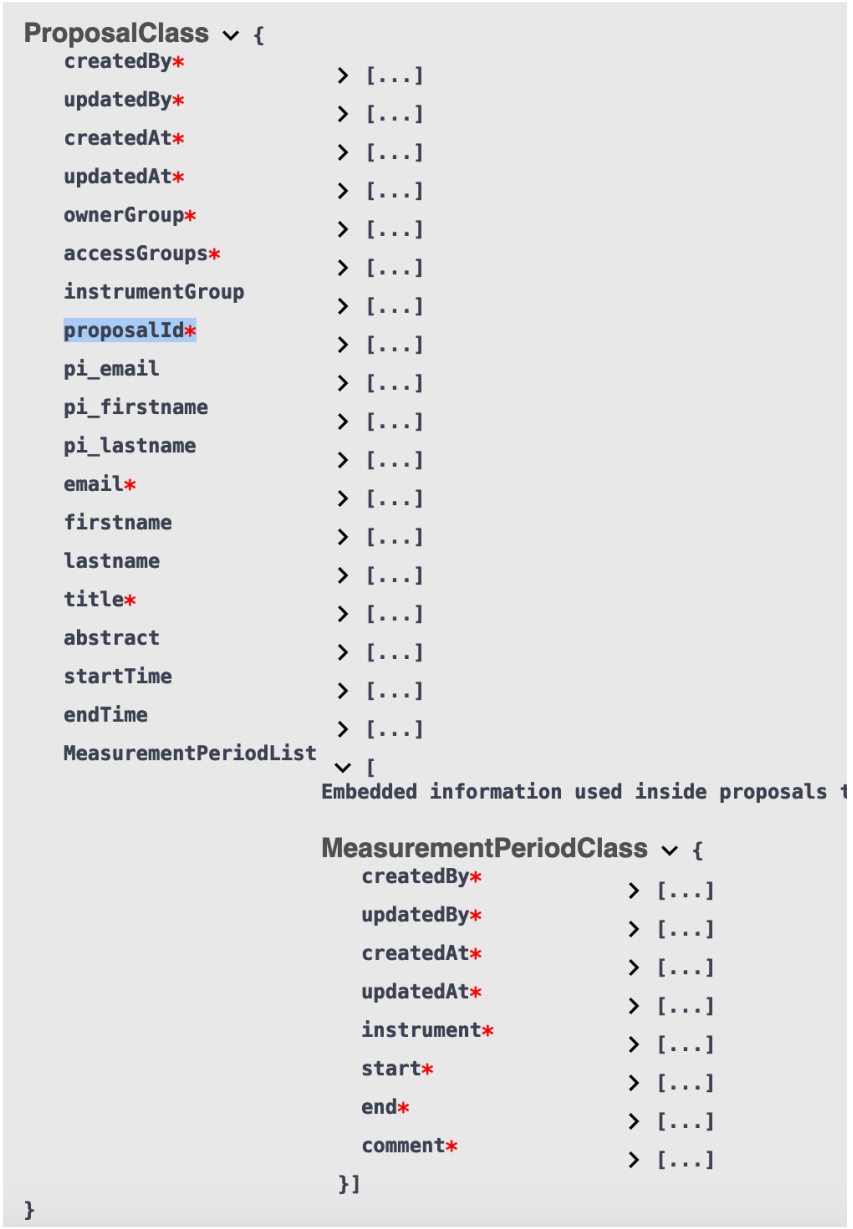
   - In ProposalClass have the real proposalID (from DOOR) and use MeasurementPeriodList for beamtime metadata:
     - One would have to add the beamtime ID in there
     - Probably change data access rights
     - Case of changing PIs not even covered
     - …

   Major change within SciCat, backwards compatibility not given, even if this would be implemented for DESY, then one would branch from main SciCat development, one could use any other catalogue ⇒ not practical.

2. Try to make names configurable in our frontend, adoptable per laboratory, for DESY, eg:

   - proposalID ⇒ beamtimeID
   - Instrument ⇒ beamline





DOI business

DOIs in Photon Science

Why and how?

DOIs

- They should not have any semantics, but be opaque, ie one should not be able to read anything inside of them to avoid confusion (from M Koehler). In L, they nevertheless use YYYY_number in some DOIs.

don't use sub-DOIs.

Workflow

1. User selects datasets in SciCat
2. Launches request of DOI creation:
   - Do we want to have them publicly available? If so, one would have to transfer to the public server.

From:
https://scicatproject.github.io/documentation/Users/Publishing.html

Data that is known to the data catalog can be published. The publication workflow does the following:

1. It defines a **set** of datasets to be published
2. It assigns metadata relevant to the publication of the datasets, such as author, abstract etc
3. It assigns a **digital object identifier** DOI to the published data, which can e.g. be used to link from a journal article to the data
4. It makes the data publicly available by providing a **landing page** that describes the data.
5. It publishes the DOI to the worldwide DOI system , e.g. from Datacite

DOIs with SciCat

Where are we? Status and plans.

What FS and L agreed on to meet user needs:

- 1 beamtime = 1 dataset

This approach has technical challenges which were discussed in FS-IT meetings. The problem file list of 1 beamtime with metadata is O(TB), ie too big.

~44000 OrigDatablocks with 1000 files per OrigDatablock (“worst case beamtime”)

```
sourceFolder: "/asap3/petra3/gpfs/p08/2024/data/11019399/raw"
```

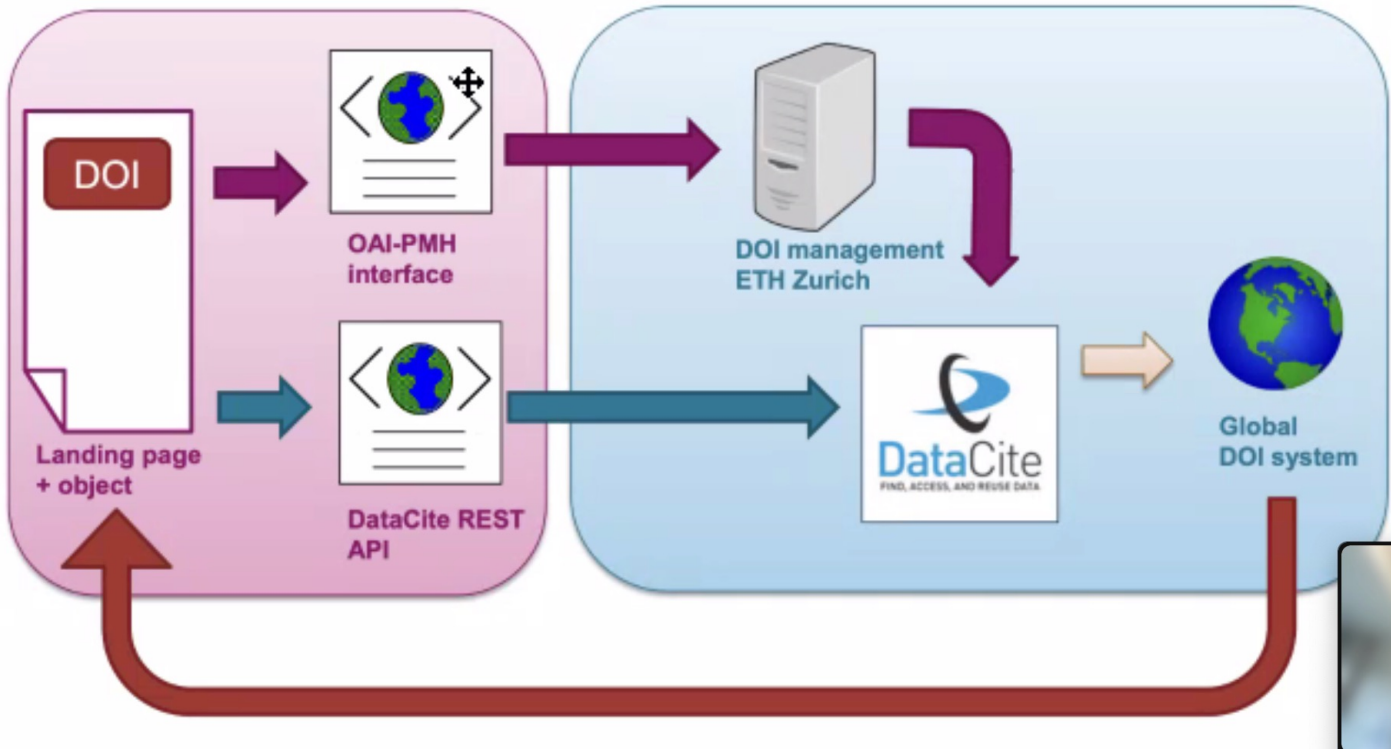
SD will check if only part of data path in GPFS can be stored.

Once this issue is solved one can chose either to use directly DataCite or SciCats OAI-PMH interface and DataCite to initiate DOI minting.

I’m investingating how DOIs are issued using SciCat at other labs, e.g. PSI, ESS, MaxIV…

DOIs at PSI

Ways to mint a DOI



- Top workflow (PSI):

1. Within SciCat chose datasets for which you want a DOI (== publish)
2. Follow OAIPMH standards
3. Tell ETH server, it looks for any publications
4. Request a DOI from DataCite
5. Once issued, data is part of global DOI system
6. Landing page with info about published dataset is created.

Conclusion and outlook

Tasks and fronts

Many things to follow up, many good people with lot of useful knowledge.

Thanks for everyone sharing their knowledge with me, for past weeks in particular DOOR people; Ulrike L, Jan-Peter K; Frank S; Linus P, Anjali A; Martin K. and also to Noel B.

For this month, my plan is to focus on work (1) on tests for jobs with Despina, (2) document our high-end Kubernetes infrastructure with Noel and Johannes and follow up on the (3) DOI workflow.

Please come and talk to me with any issues, your ideas, wishes, questions about our meta data catalogue!

Next weeks SCT : tbc

Next SCG: 2024-04-26