SciCat @ DESY

How do we want to structure the data ingestion process

Linus Pithan, FS-EC

HELMHOLTZ

DESY.

Where are we?

in 2023:

- We go a deployable SciCat version based on the new backend with HelmholtzID (AAI) integration … thanks to Noel!
- We “rebooted” the discussion process around SciCat at DESY
- We saw the onboarding of new colleagues that helped to bring new energy in long-lasting discussions … first and foremost Regina
- We multiplied the number of test instances
- Migration of the ingestion process at P08 to the new Backend … thanks to Jan Kotanski
- We started new threads that are to be followed in 2024 (e.g. minting of DOIs, SciCat instance with one dataset per proposal) … thanks to Stefan Dietrich for first steps

Illustrations: 10.5281/zenodo.3332807 *The Turing Way* project, Scriberia

Some challenges and issues for 2024:

(this is a non-exhaustive list)

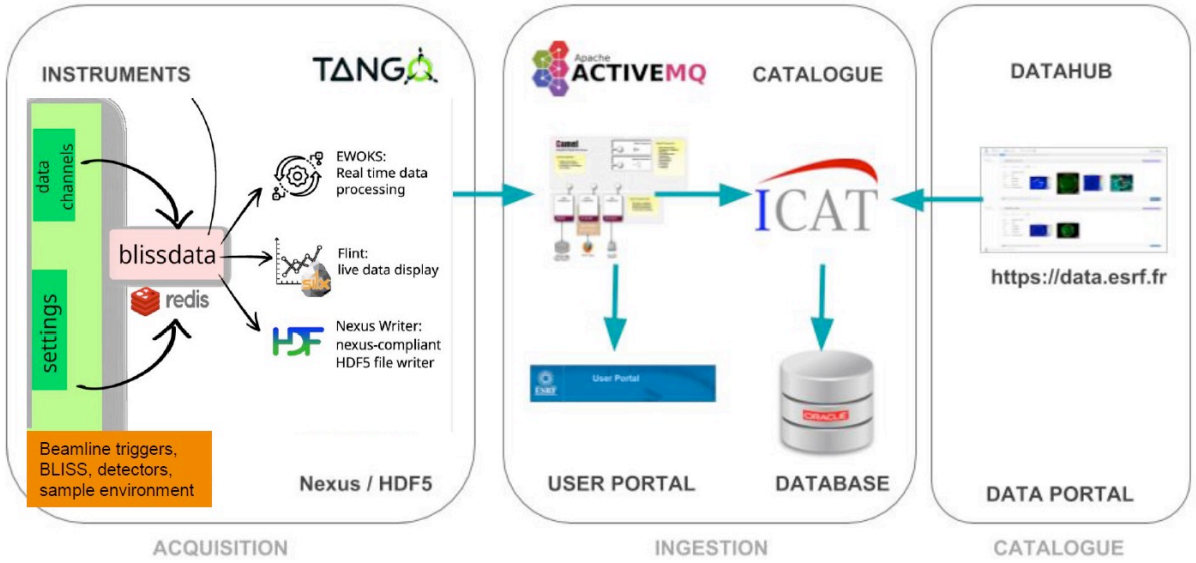
- Agree on a schema for the ‘scientific metadata’
- Define how we plan to structure the ingestion process
- See if there are local adaptation e.g. in the access and permission module (casl) are necessary
- …

Illustrations: 10.5281/zenodo.3332807 *The Turing Way* project, Scriberia

What are the features we want in the (dataset) ingestion process at DESY?

We saw the complexity of the ingestion process at other facilities …

@Maxlab, thanks to Daphne van Dijken



@ESRF, thanks to Andy Goetz

… conclusions to be drawn on Friday…

We saw that we have to acknowledge that

- there are different stake holders in the context of SciCat and different needs to cater for (visiting users, beamline scientists, DAPHNE4NFDI, DESY as facility)

- different levels of standardization in different sections of the meta data stored in SciCat are needed (controlled vocabulary where feasible but not everywhere)

- to cover automated and manual dataset ingestion

… lets keep the schema discussion for another day…

We saw that we need to talk about a policy for data catalogue in order to build the right tools

Exemplary questions in this context are:

- Who can ingest something into the catalogue (the beamline / the user)?
- Do we allow post-experiment editing of metadata?

… now this is the SciCat **technical** meeting, therefore focus on **what** and **how** we want to implement to make sure we find the suitable framework that suits most use cases

Ingredients for the data ingestion pipeline at DESY

Schema management

E.g. a Gitlab-Repo with simple configuration (e.g. yaml based) and accompanying python module to interact with it to

- generate documentation
- build standardized schemas (e.g. json-schema or NeXus)
- offer validator
- …

Automated metadata capture at the beamline.

- This includes
  - meta data harvesting
  - generation or collection of thumbnails
  - ensuring persistency and beamlime operation if SciCat is down
- Prototype implemented on P08 by Jan Kotanski

User interface for manual ingestion

Web UI

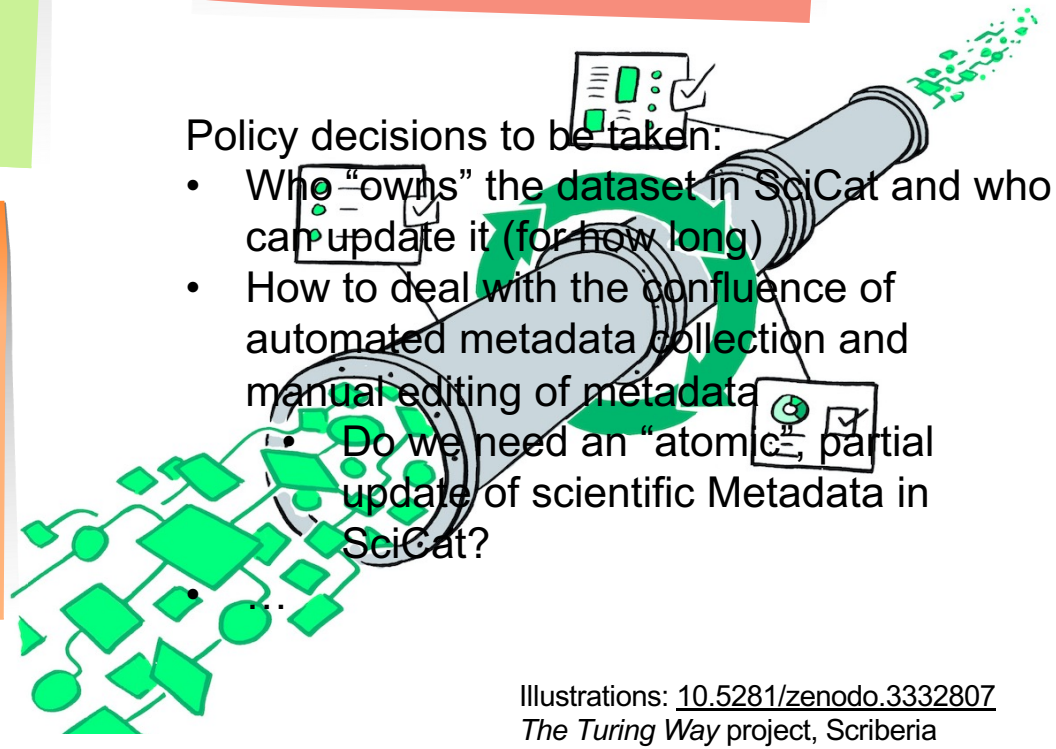
- to define sub-schema for user-meta-data
- insert user-meta-data
- See currently "active" dataset

Service that connects automated and manual metadata ingestion

- Schema validation
- Broadcasting of "active" dataset
- Access to beamtime on GPFS to keep the user-schema

Sardana integration

- Starting and terminating datasets
- Grouping of several scans into one dataset



Policy decisions to be taken:

- Who "owns" the dataset in SciCat and who can update it (for how long)
- How to deal with the confluence of automated metadata collection and manual editing of metadata
  - Do we need an "atomic" partial update of scientific Metadata in SciCat?
- …

Illustrations: 10.5281/zenodo.3332807
The Turing Way project, Scriberia