Blogpost No.1: SampleDB vs. SciCat

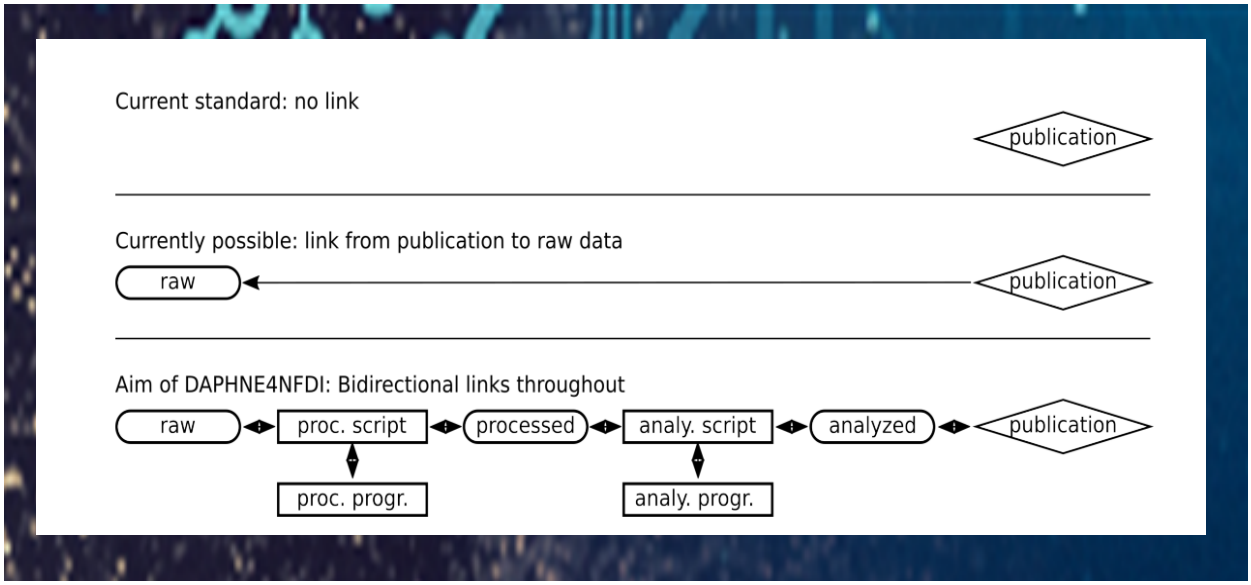
Moritz Hannemann
+4989 158860 780
m.hannemann@fz-juelich.de

Table of Contents

Overview

The user community of DAPHNE4NFDI is highly inhomogeneous. Our users do measurements with neutron- or photon sources, they measure absorption, fluorescence, diffraction, tomography and many other parameters. Furthermore, not only data from the beamlines are connected to a sample, usually also other characterization measurements are being executed.

Consequently, it is challenging to track all relevant data to one experiment. It is therefore necessary to attach metadata relevant for the experiment, so users can later better retrieve data und understand the conditions of the experiment. Metadata are very important to make DAPHNE data adhere to FAIR principles. Adhering to these principles enables everyone around the world to understand under which circumstances the data was collected. This way data will be democratized.



Many cases no connection between published data and acquired data exists. Recently raw data have been connected with published data. Aim of DAPHNE4NFDI is to automatically extract metadata from acquired data and enrich them with more information, so that readers of publications can easily find information related to the publication.

Goal of this study

This work was done in order to fulfill **deliverable 2.1.1** of the DAPHNE proposal:
“Comparison of different repository/catalog systems and recommendations which one to adopt at participating facilities or third-party contributors”

The following work includes an overview over common features of the two presented catalogs here, as well as unique features of them. Additionally, the results of a benchmark study will be shown and explained.

Catalogs

Common Features of SampleDB and SciCat

- Meta data source (Front-end / API)
- Search functionality (Front-end / API)
- Instrument description
- Online visualization
- Online evaluation
- Data life cycle management
- IdM
- ELN functionality
- Associate DOI‘s
- Embargo Time
- Download / Zip-Service

SampleDB

SampleDB is a web-based sample and measurement metadata catalog maintained by the Scientific IT-Systems group at Forschungszentrum Jülich GmbH. The catalog is in use at several institutes at Forschungszentrum Jülich campus, but is also available to others under MIT license. In the early developments, SampleDB was developed with a sample centric-view. This has evolved into an open and flexible structure defined by schemas. The creation of those schemas may need some extra effort in the beginning, but offers the following advantages. Web forms are automatically generated from the schemas and all user input, either by filling those forms or by automatic ingest via the API, is validated, resulting in high quality metadata. Schemas may be stored and shared using setup scripts to facilitate its reuse across different facilities. In addition, SampleDB’s built-in federation support offers new opportunities regarding cross-facility exchange, e.g. when measuring the same sample at different facilities with different techniques, the same sample identified by its PID can be accessed on institutional SampleDB instances using federation. An export from SampleDB to SciCat is currently under development.

Documentation.

Unique Features of SampleDB

- Built-in federation (SampleDB only). More info.
- Extensive search front-end
- Export data to SciCat (under development)

- No initial structure
    - Schema creation / editing
    - Automatic schema validation
    - Automatic web form creation from schema

SciCat

SciCat is a metadata catalog, developed at PSI, ESS and MAXIV, allowing users to access information about experimental results. The catalog will be used at the European Spallation Source (esss.se), but has already been adopted by different facilities. The catalog is available under GPL license. Scientific datasets are linked to proposals and samples. Scientific datasets are linked to publications (DOI, PID). SciCat helps to keep track of data provenance (i.e. the steps leading to the final results). SciCat allows users to find data based on the metadata (both your own data and other peoples' public data). In the long term, SciCat will help to automate scientific analysis workflows.

Documentation.

Unique Features of SciCat

- Initial schema structure
- Event Trigger (Kafka, RabbitMQ)

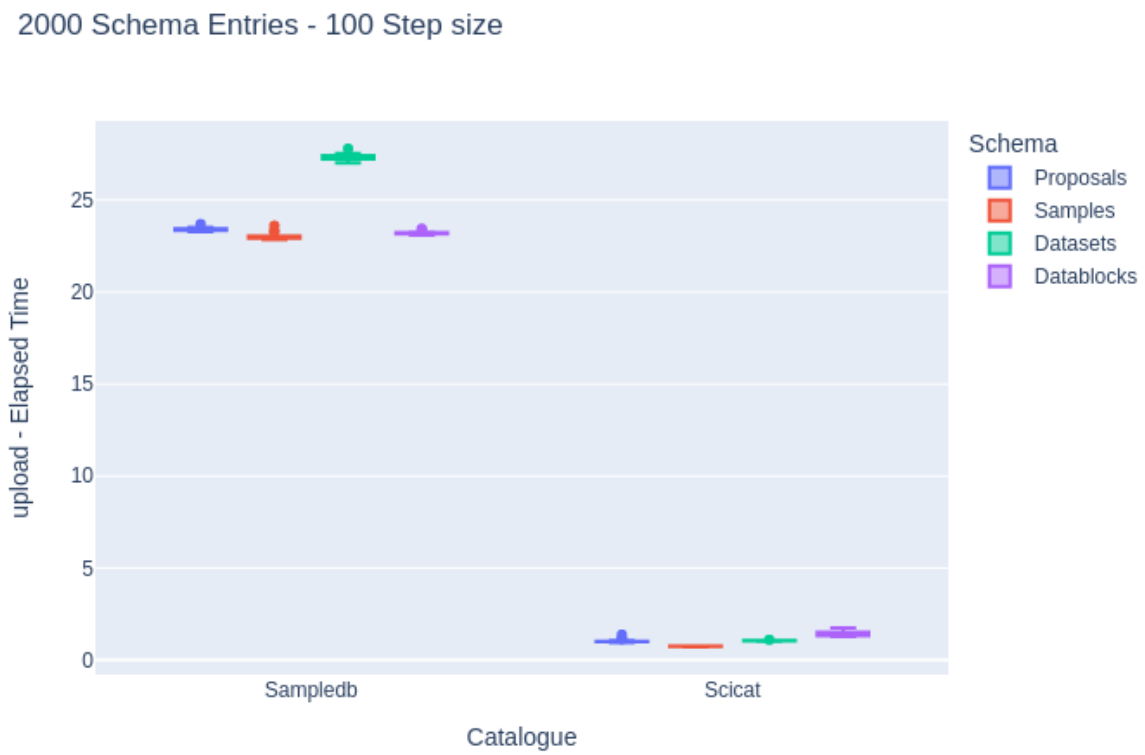
Benchmark Study

For the benchmark study, the structure of SciCat was applied to SampleDB. This way both applications were handling the same data and could be compared quite well. In SampleDB, schemata for proposals, samples, datasets and datablocks were created.

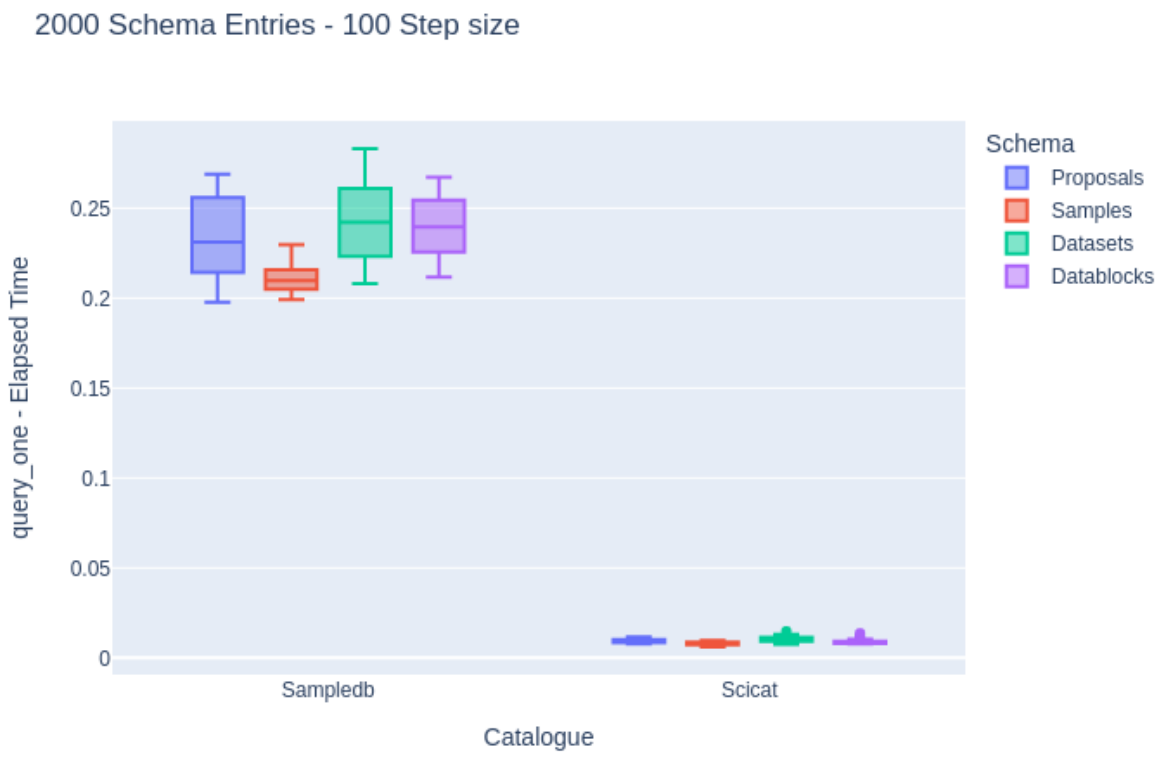
In comparison, SampleDB validates the data on ingest against the given schema. This results in slightly longer upload times, but the biggest time difference is due to the authentication part. In terms of security, SampleDB uses a slow hashing function for long living API tokens, spending upt to 80% of the time in this function for a single API call. This can be improved in the future using a different authentication mechanism for API calls. However, the frontend already uses a different authentication mechanism based on sessions and is not affected by this. The results shown below are limited to the use of API calls.

How the benchmark study was conducted:
In the beginning, both catalogs were empty. Then data was uploaded and after a defined number of single uploads, called steps in this case, queries were run on the catalogs. This enabled us to see if the upload speed dropped over time, when more data accumulated in the catalog. We also observed, how the query speed is affected by the amount of data in the catalog.

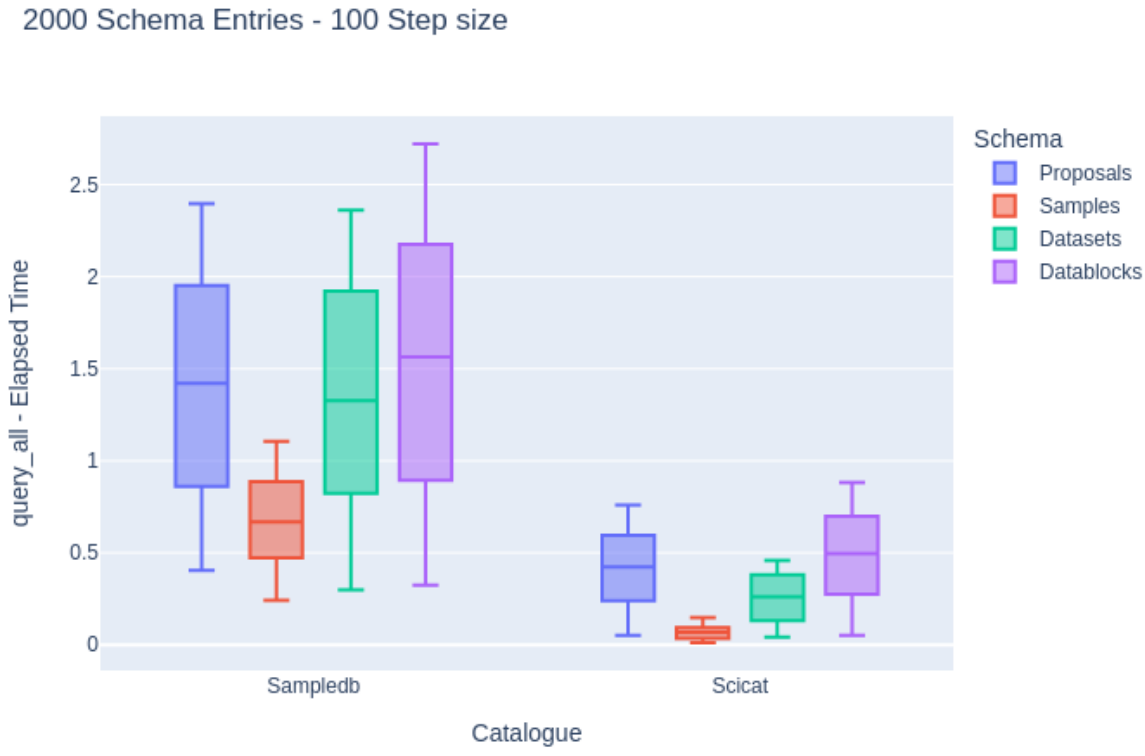
Upload duration for 100 uploads:



Query that returns exactly one dataset:



Query that returns some (~10) Datasets:



Summary

Overall, both catalogs are easy to use. SciCat performance is faster, but does not validate data by itself. SampleDB has a user-friendly user interface to input data by hand while SciCat is more designed for automatic ingestion of data.

One remark about the benchmark. Whereas SciCat performance is definitely faster, there are also options to improve the speed of SampleDB API calls. Please note, that the SampleDB front-end uses a different authentication mechanism not affected by this bottleneck

Both catalogs are actively developed and will evolve over time. New features are added and discussions on how to improve are taking place all the time. SciCat is already well-known in the user community whereas SampleDB is looking forward to engage with the community outside of its own premises at Forschungszentrum Jülich.