Approaching SciCat from a University Group Perspective

Dr. Linus Pithan – Universität Tübingen
linus.pithan@uni-tuebingen.de

04.06.2022 SciCat Meeting @ DESY

- Brief introduction of the DAPHNE4NFDI consortium
- Use cases of SciCat within DAPHNE
- SciCat related developments in DAPHNE
  - Web-Frontend for ingestion
  - Metadata schema management
  - Schema validation

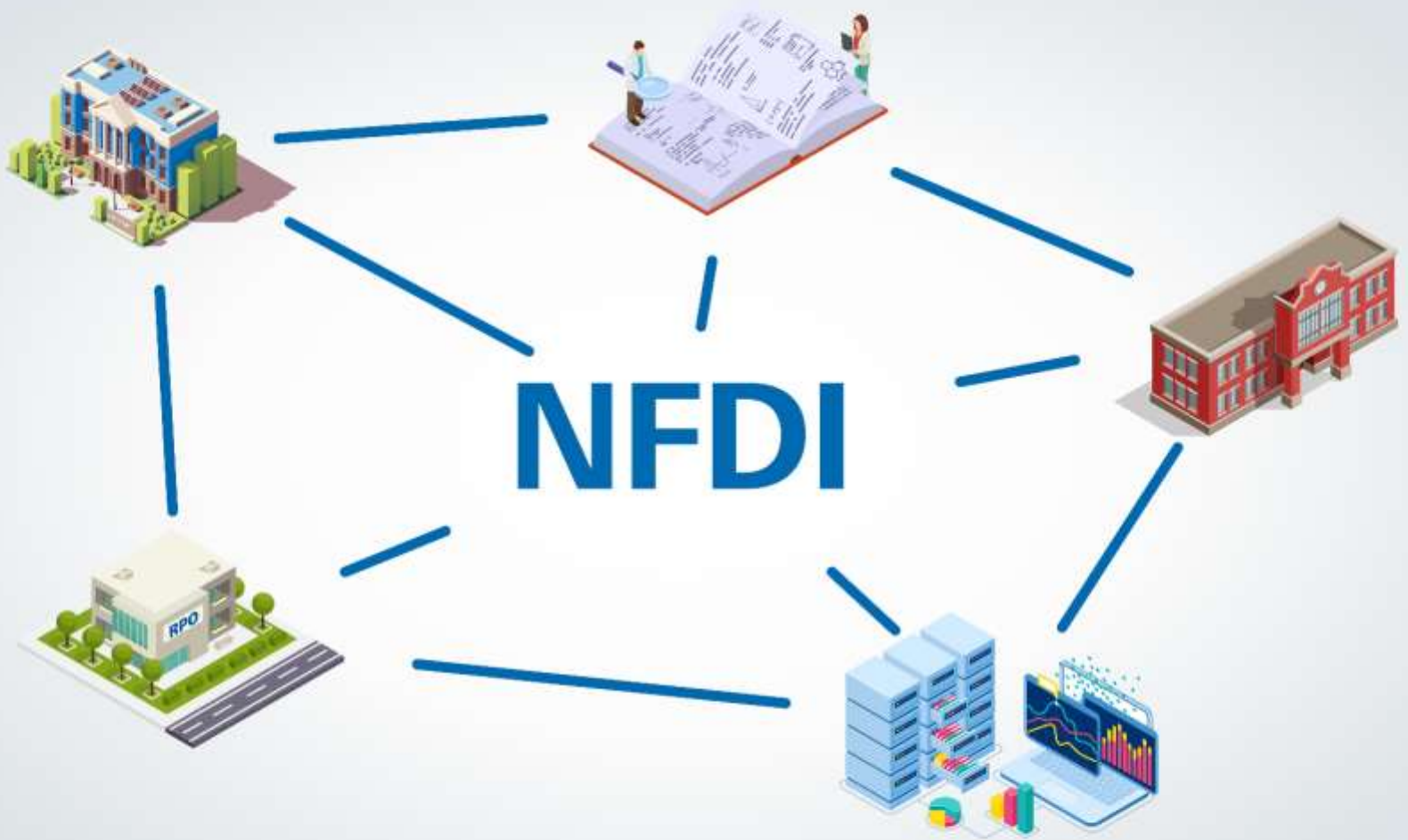
National Research Data Infrastructure

The aim of the NFDI is to systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally.
The NFDI will bring multiple stakeholders together in a coordinated network of consortia tasked with providing science-driven data services to research communities.

DAPHNE is a NFDI consortium that focuses on research with photons and neutrons at large-scale research facilities.

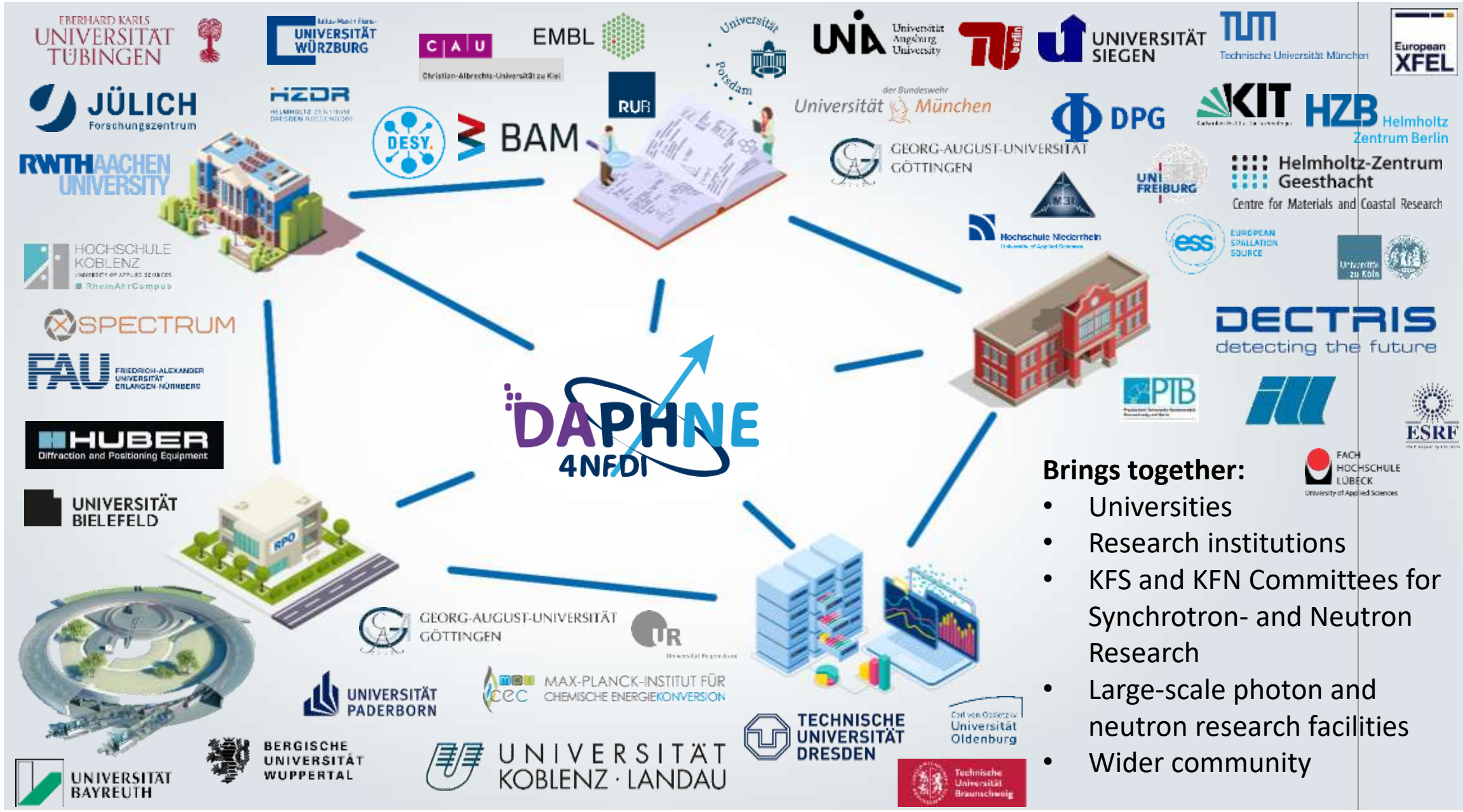
DAta for PHoton and Neutron Experiments

https://www.dfg.de/en/research_funding/programmes/nfdi/index.html

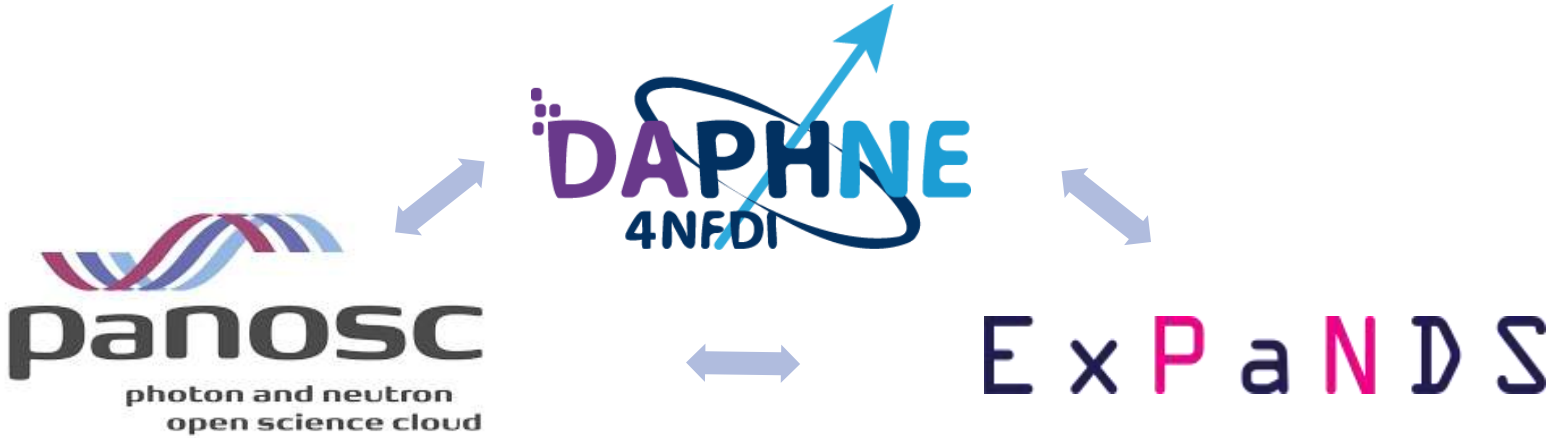


NFDI brings multiple stakeholders together in a coordinated network of consortia tasked with providing science-driven data services to research communities.

https://www.dfg.de/en/research_funding/programmes/nfdi/information_material/index.html


DAPHNE 4NFDI
Brings together:
• Universities
• Research institutions
• KFS and KFN Committees for Synchrotron- and Neutron Research
• Large-scale photon and neutron research facilities
• Wider community

- **The goal of DAPHNE** is to make the growing volume of valuable measured data FAIR for the DAPHNE community, for the whole NFDI and the scientific community.
- The key objectives to be achieved within DAPHNE are:
  - **Collection** of **metadata** so that the **measured data** is **reusable**
  - **Searchable curated databases** of raw, intermediate and processed data
  - Develop a **curated repository of managed software >> re-use** the data
  - Develop a **multidisciplinary data platform** for NFDI cross-consortia actions;
  - **Education** and **training** in research data management.



NEW in DAPHNE: Universities are on board as well!

Data Catalogues in DAPHNE

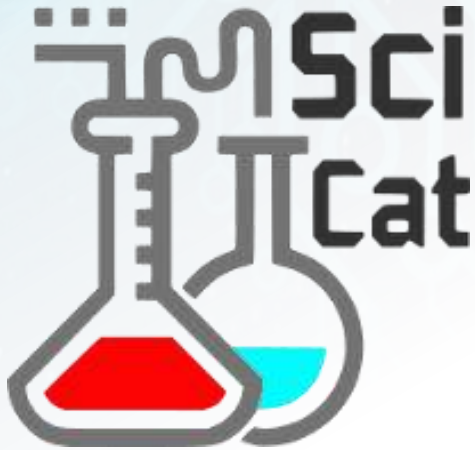
Facilities

- Either have already chosen a data catalogue solution or look at solutions that specifically fit to the institution

Universities (research groups organised in DAPHNE)

- Usually do not have data catalogues yet
- Have specific needs that differ from those of large-scale facilities
- Less IT support

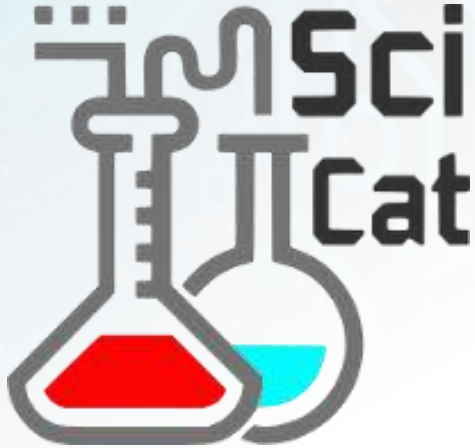
...



Statement DAPHNE4NFDI Executive Board – Test and Evaluation of SciCat

The DAPHNE4NFDI Executive board recommends that the DAPHNE participants test and evaluate SciCat at their home labs in universities and at facilities - if possible.
They should provide feedback and indicate where further collective development is required - including deployment and integration.

30/05/2022



Approaching SciCat from a University Group Perspective

Use-case specific extensions to SciCat

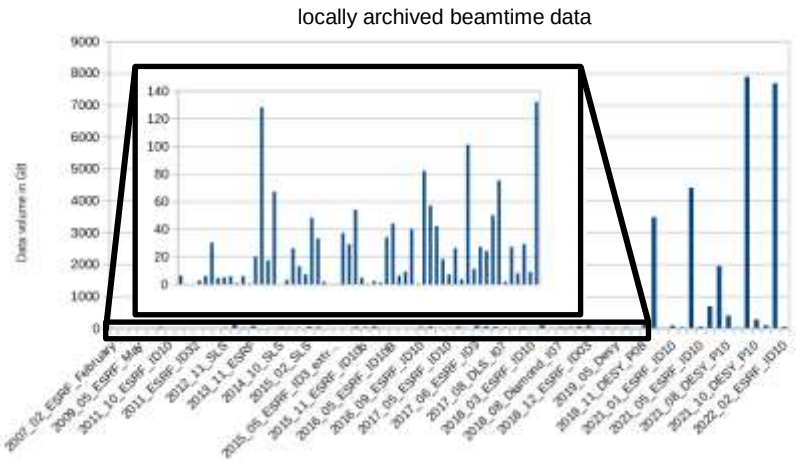
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

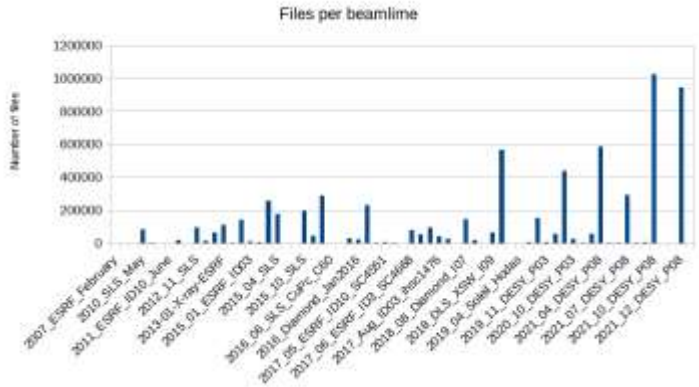
Universität Tübingen
Institute of Applied Physics
Prof. Frank Schreiber

SciCat use-case in our research group

Aims:

- Make beamtime data available for internal reuse
- Develop schema to store metadata for machine learning (ML)
- Collection of annotated datasets for training & testing of ML-codes





Obvious differences to original SciCat use-case:

- Only few users in parallel
- Manual or `semi-manual` data ingestion
- Simple user account and access right management

What did we do?

Provide prototype of an “Ingestion Frontend”

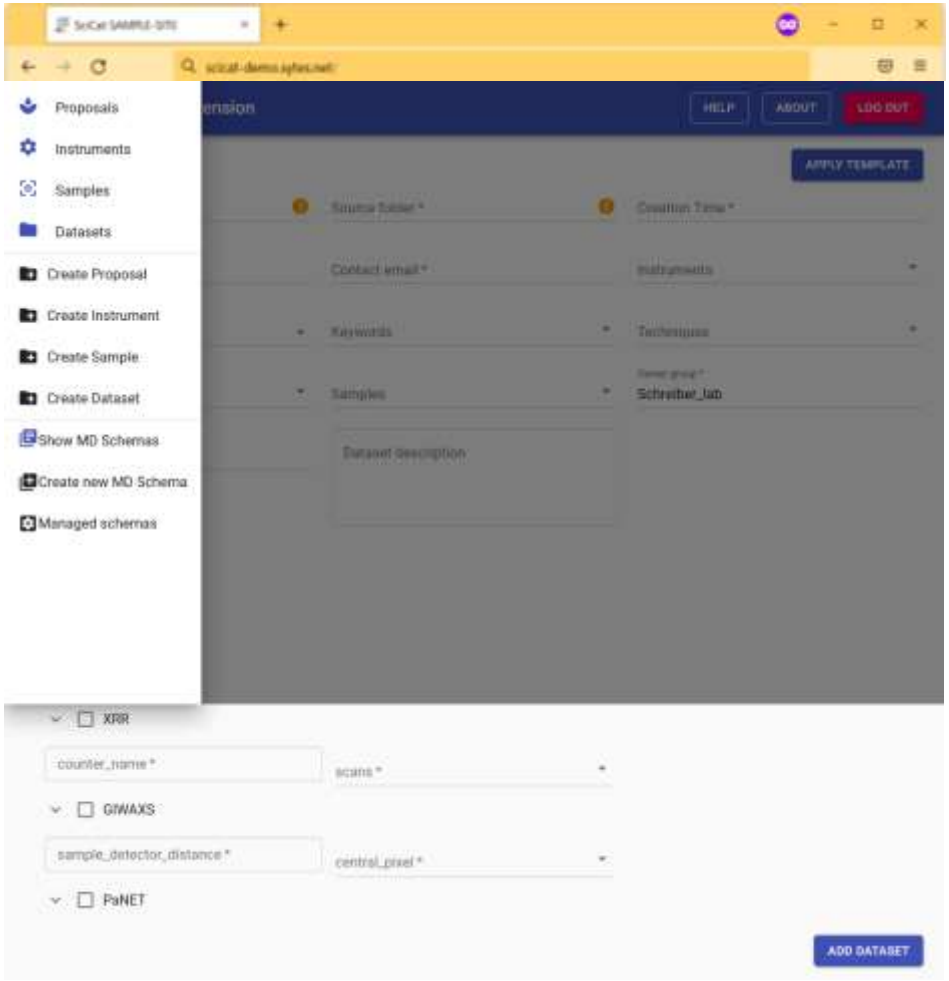
- Create datasets through web interface

Add schema management for Scientific Metadata to SciCat

- Add some validatable structure to metadata groups

Started to work on a simple, container-based production deployment

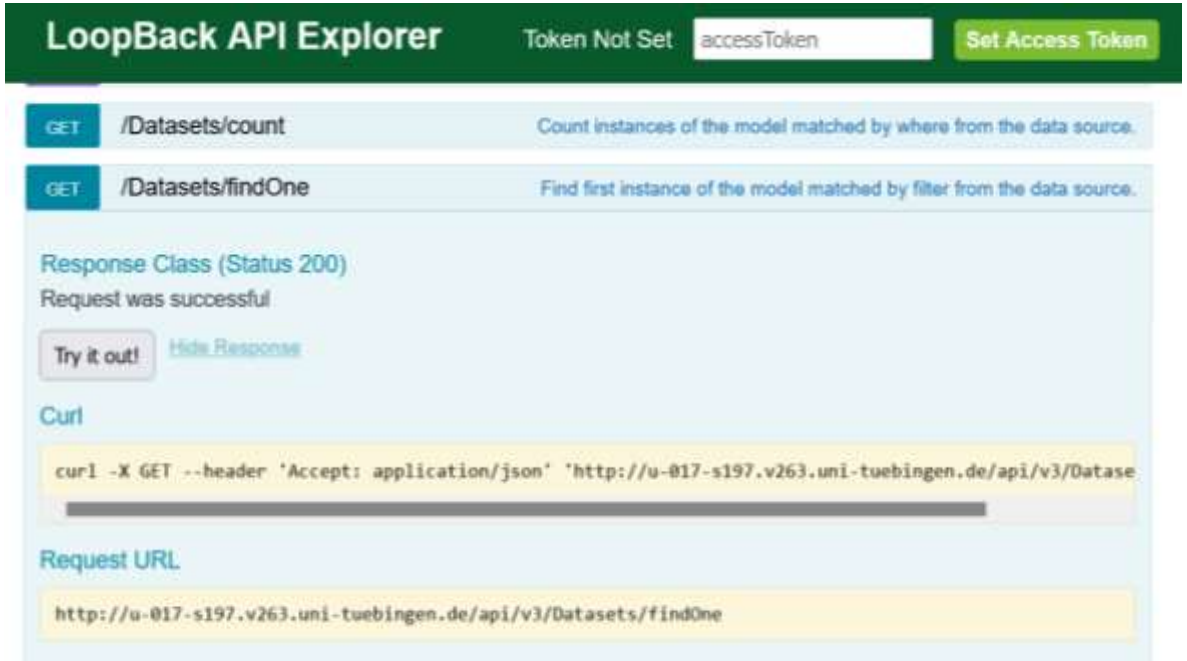
- Docker-compose based solution



SciCat features that we rely on

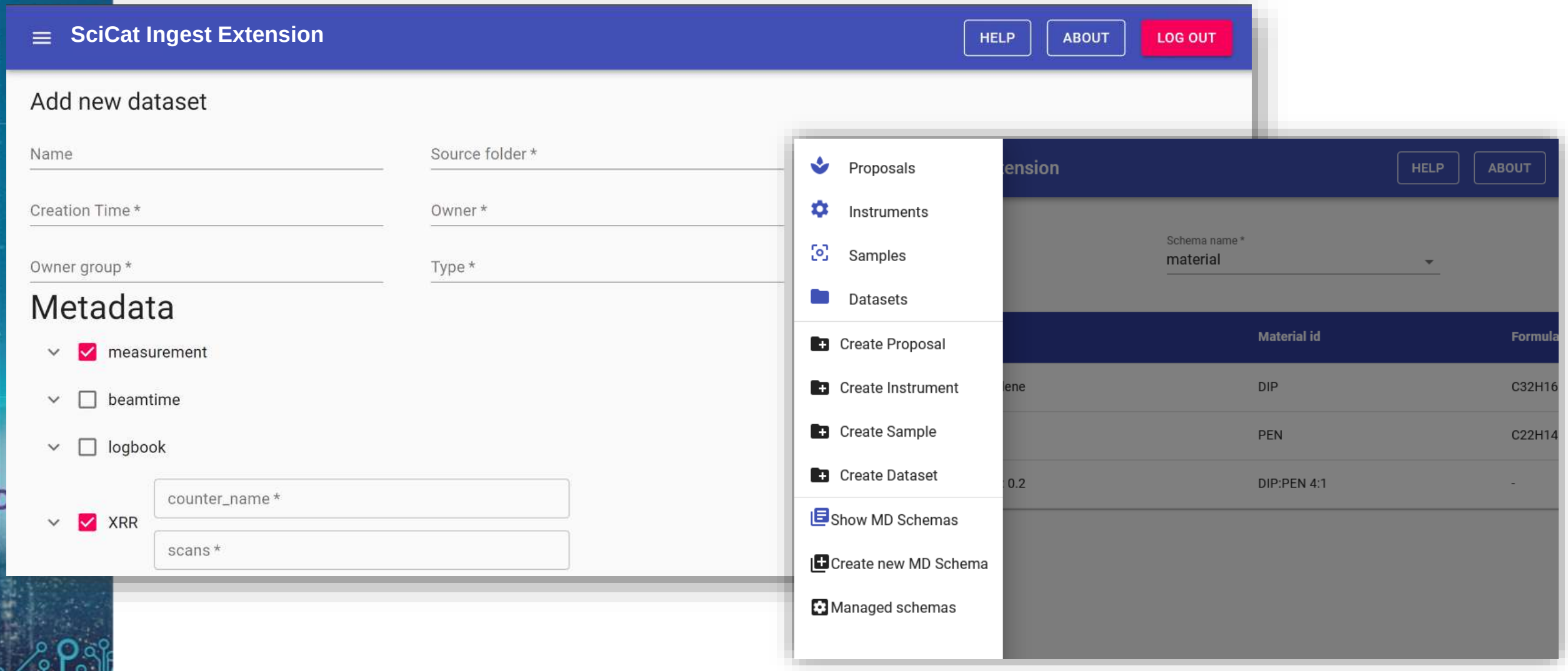
- Data structure: `Proposal` / `Sample` / `Dataset`
- The SciCat frontend to explore data
- `Scientific Metadata` management
- Handling of `raw` and `derived` datasets
- Search box
- Rest API

  (machine readable web-interface)
- Underlying database (mongodb)

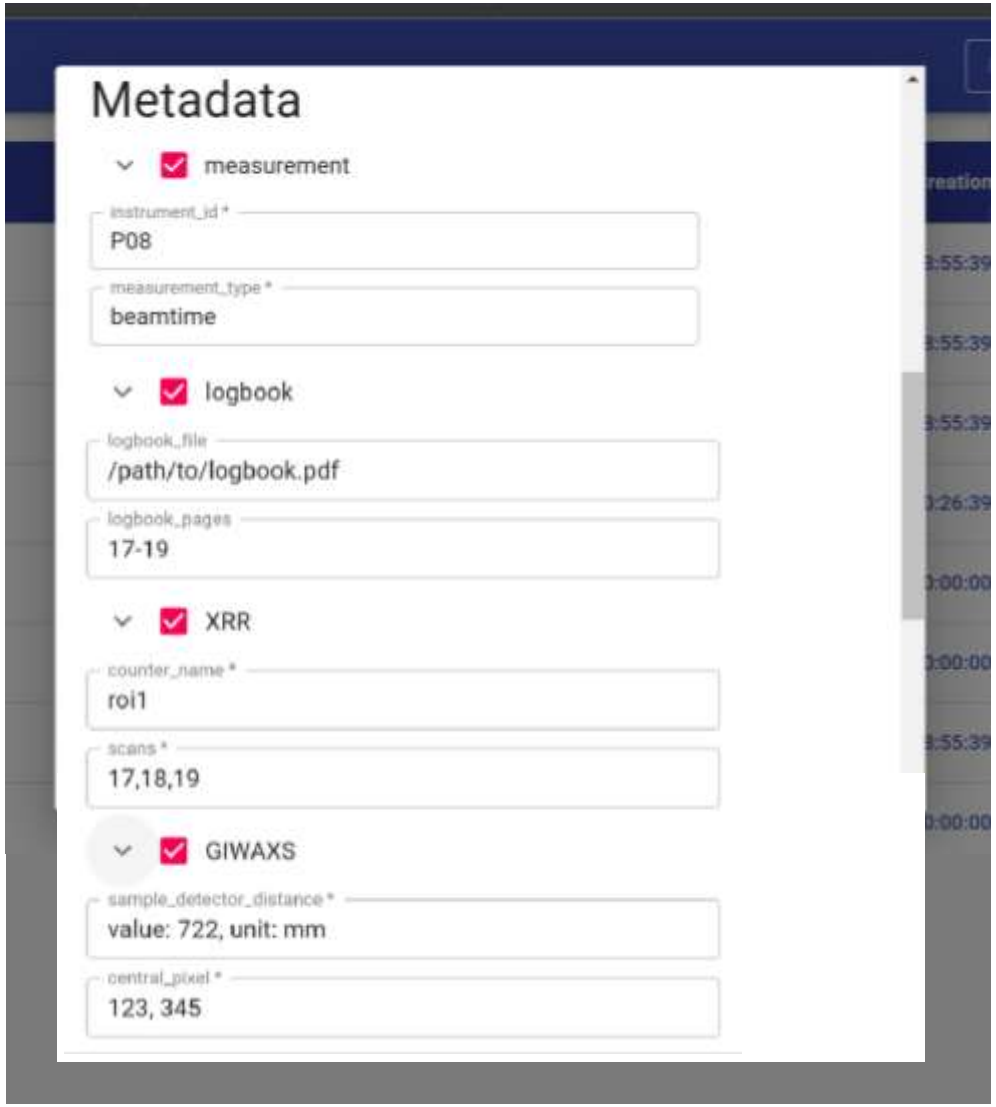


Features that SciCat does not provide out-of-the-box

- Graphical interface to ingest data
- Standardization and validation of metadata schemas
- Production-ready, easy-to-use deployment option for university hardware without dedicated IT support


SciCat Ingest Extension
HELP
ABOUT
LOG OUT
Add new dataset
Name
Source folder *
Creation Time *
Owner *
Owner group *
Type *
Metadata
measurement
beamtime
logbook
XRR
counter_name *
scans *
Proposals
Instruments
Samples
Datasets
Create Proposal
Create Instrument
Create Sample
Create Dataset
Show MD Schemas
Create new MD Schema
Managed schemas
HELP
ABOUT
Schema name *
material
Material id
Formula
DIP
C32H16
PEN
C22H14
DIP:PEN 4:1

Modular meta data schemas



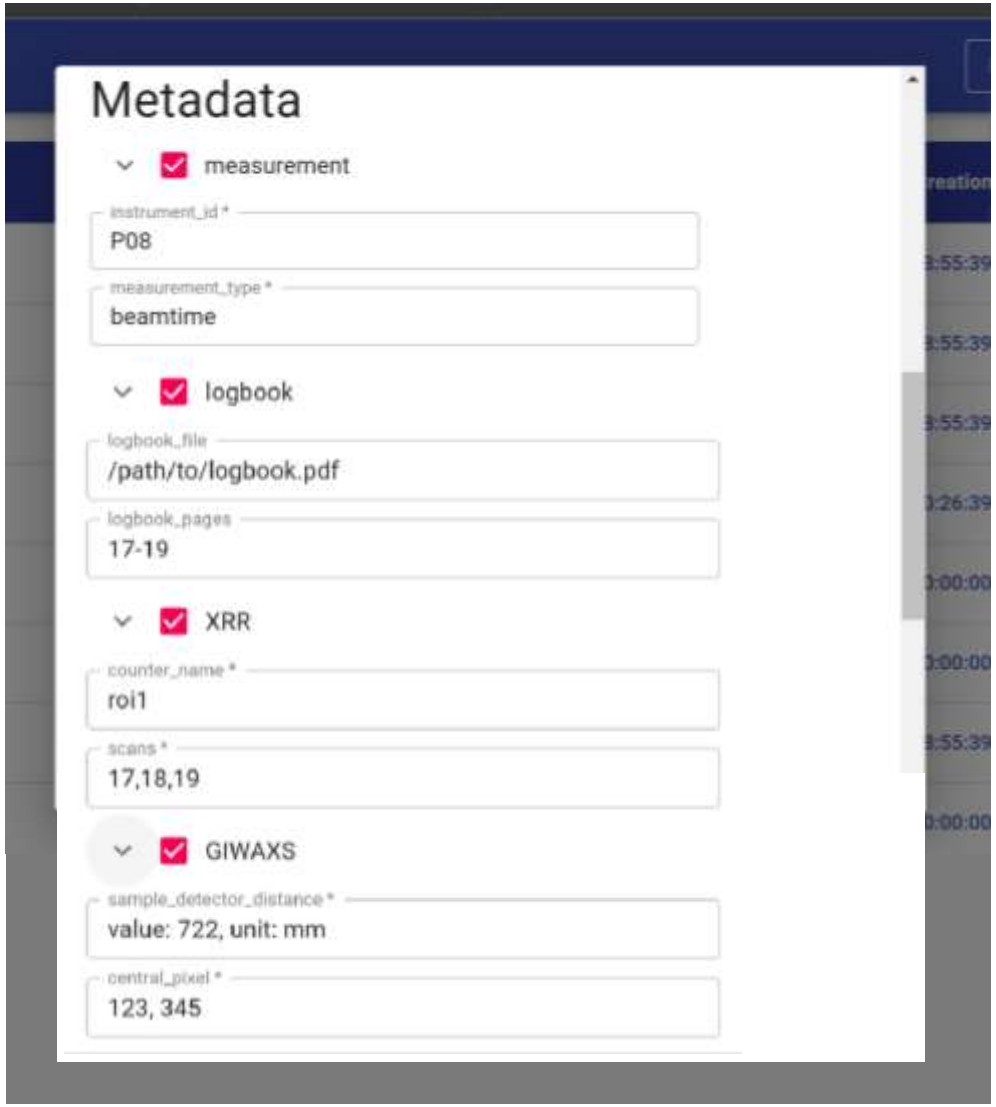
Modular meta data schemas

- Ideally: Structure metadata following a community standard e.g. an ontology or NeXus definitions
  - not directly achievable for new actors in the “meta-data game”, that can not anticipate a final, suitable metadata structure and therefore need some flexibility to find a structure that suits the needs.
  - Pragmatic approach: Organize metadata in modular categories to assure data consistency and allow modifications at run-time

→ use self-defined, small meta data schemas

→ in a way inspired by AMARCORD

→ schema approach also used in SampleDB

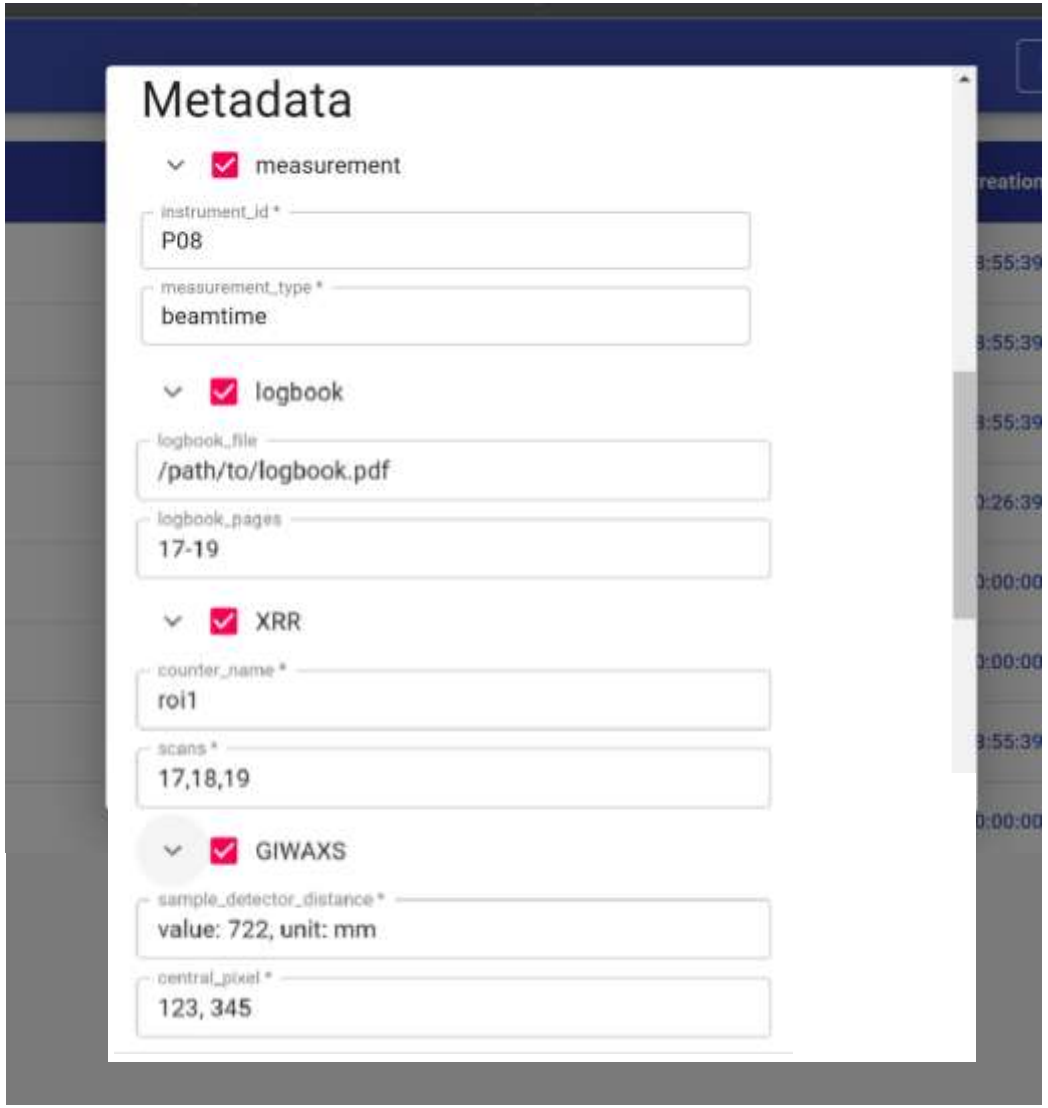


Adding Schemas to Scientific Metadata

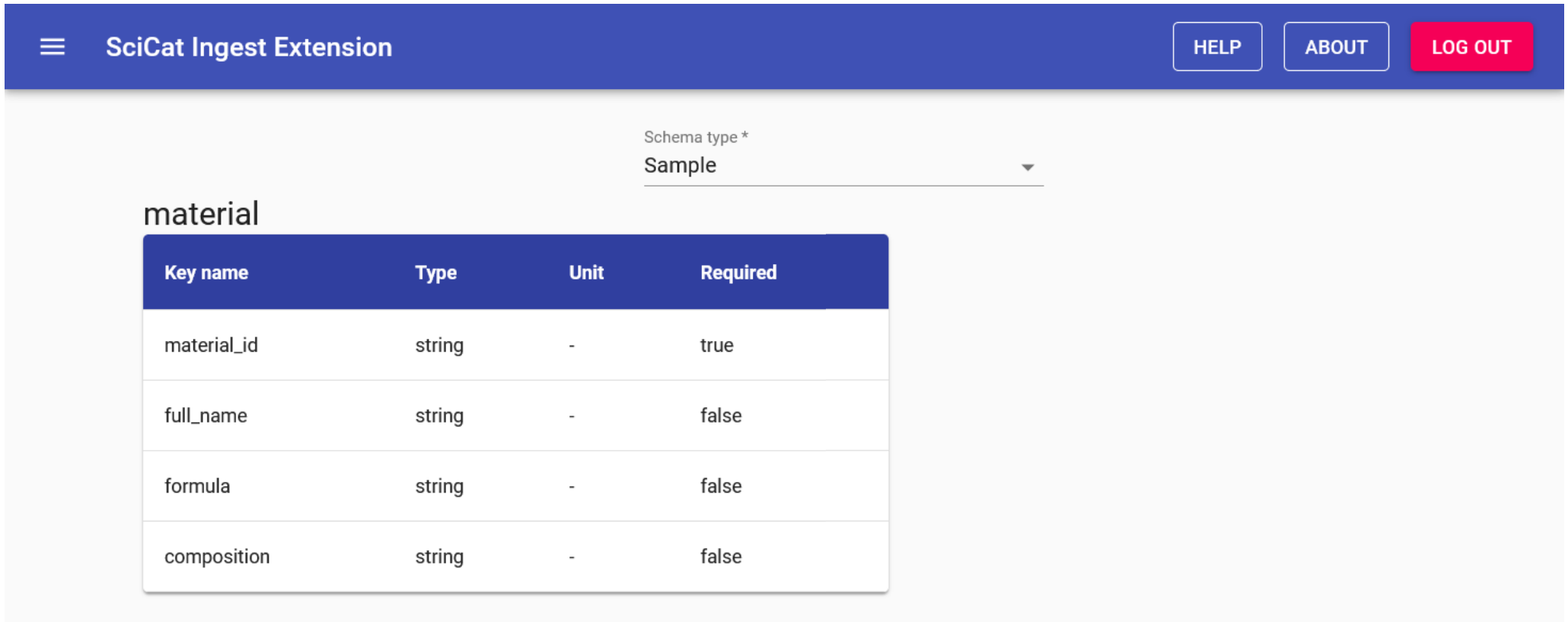
Exemplary metadata schema:

```
schema_name: GIWAXS
schema_type: dataset
keys:
   - key_name: sample_detector_distance
     type: number
     unit: "mm"
   - key_name: central_pixel
     type: list
     schema:
       - type: number
```

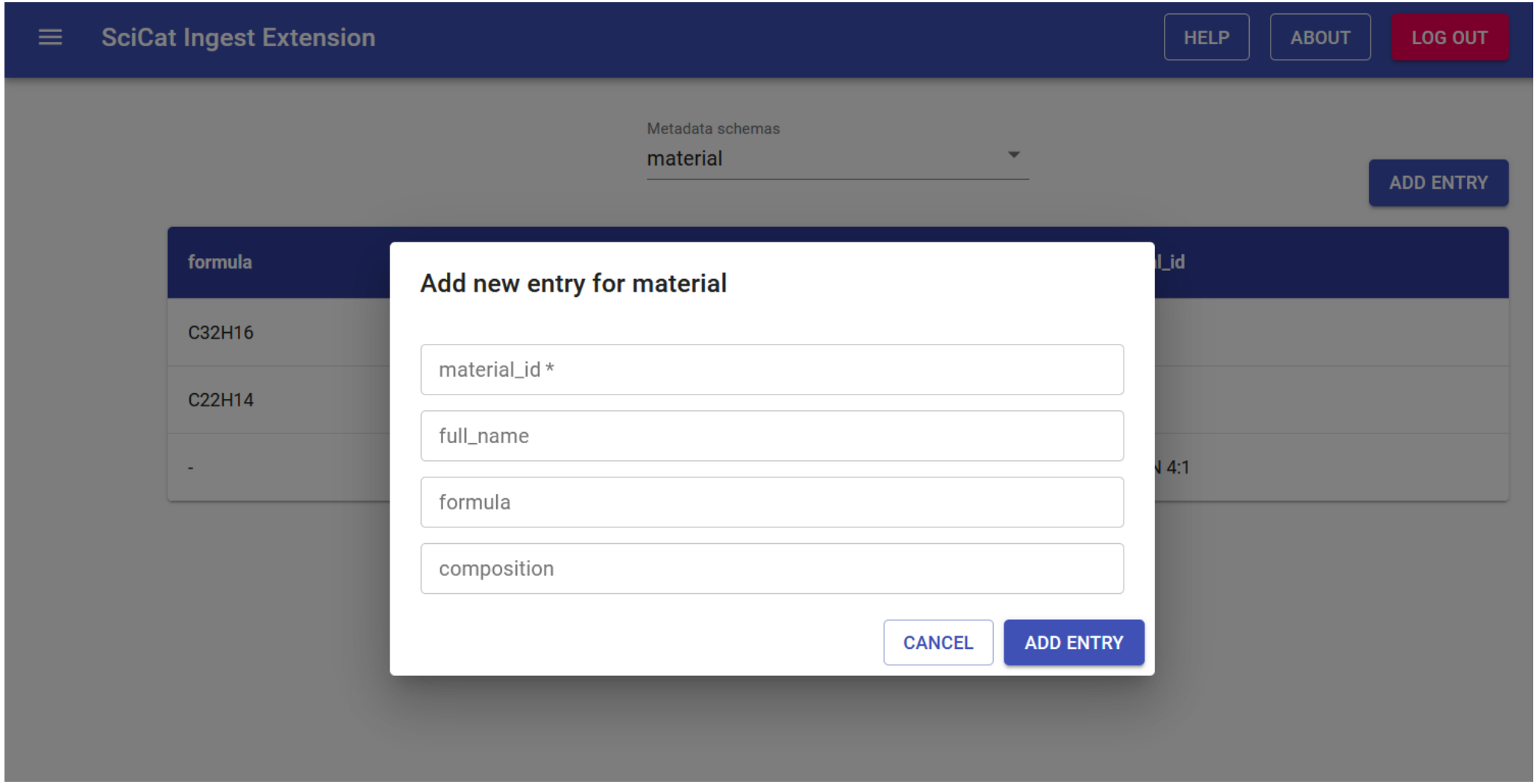
- New schemas can be defined through the web-interface
- Schema can e.g., be specific for the type of experiment or a technique)
- There is a validation step to verify that the provided metadata is in good shape



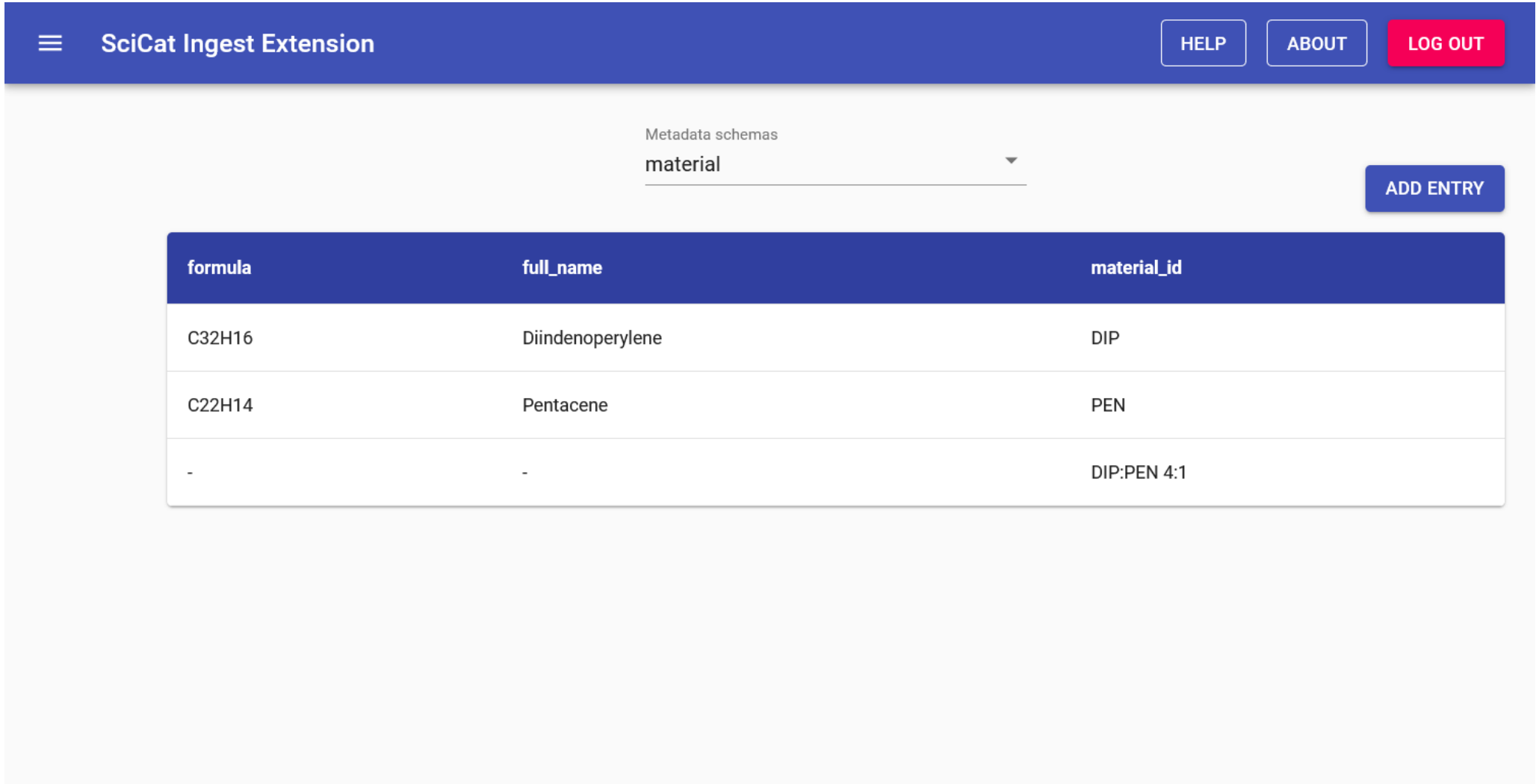
Store pre-populated meta-data blocks for better user experience



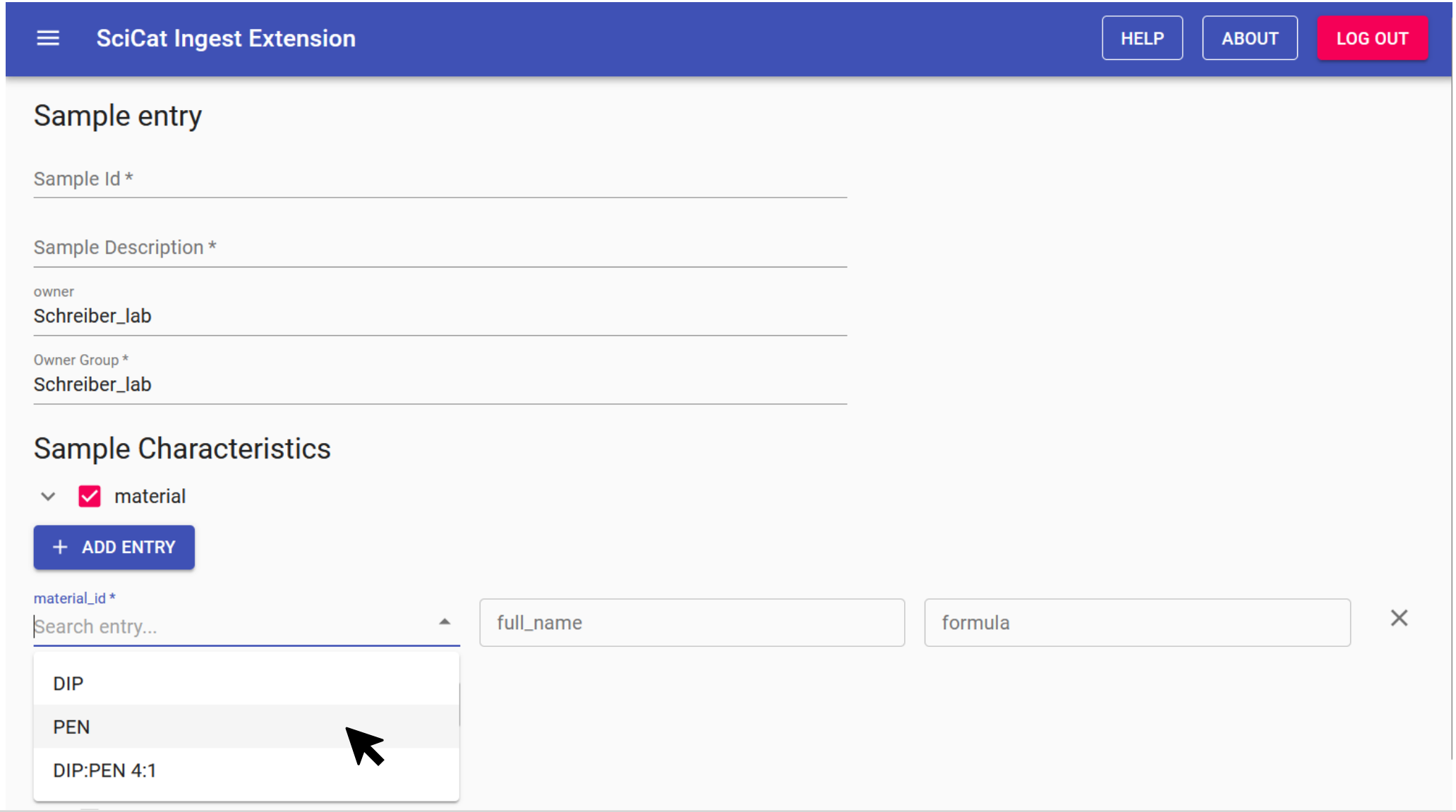
Store pre-populated meta-data blocks for better user experience



Store pre-populated meta-data blocks for better user experience



Store pre-populated meta-data blocks for better user experience



Wrap-up

Summary:

- Development of SciCat deployment strategies for **university-lab size installation**
- Prototype for data **ingestion web-interface**
- Strategies for **pragmatic** but nevertheless **systematic metadata** organisation using schemas.

Documentation
https://schreiber-lab.github.io/SciCat4daphne/

Git
https://github.com/schreiber-lab/scicat4daphne

Outlook:

- Provide `real` use cases especially around machine learning in near future
- Looking forward to share and collaborate on this project within DAPHNE

Acknowledgments

- Anastasiia Pylypenko (Uni Tübingen, frontend development)
- SciCat Project Team (Max, Carlo, Tobias, Stephan…)
- Ingo Breßler (BAM)

Illustration: https://www.ontotext.com/knowledgehub/fundamentals/metadata-fundamental/