HELMHOLTZ

RESEARCH FOR
GRAND CHALLENGES

Uniform Data Model For SciCat At DESY FS-EC

DESY. HZB Helmholtz Zentrum Berlin HZDR KIT Karlsruhe Institute of Technology

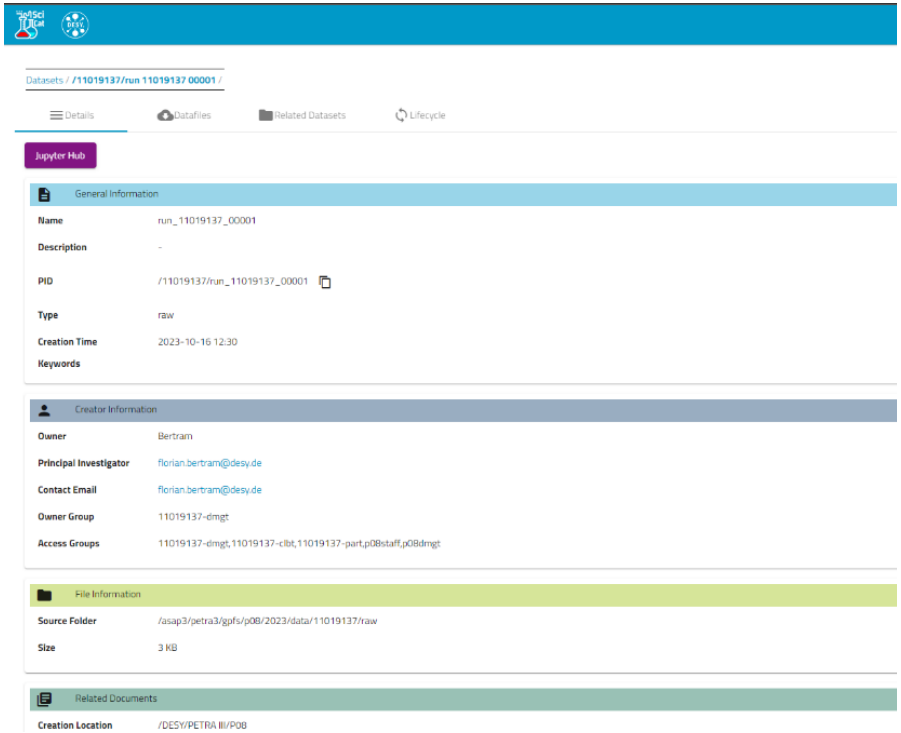
METADATA CATALOGUE - SCICAT

What is SciCat?

- A scientific data cataloging system
- Designed to manage and publish experimental data from scientific research
- Supports FAIR principles (Findable, Accessible, Interoperable, Reusable)

Purpose of SciCat:

- Facilitate data discovery and reuse
- Enhance collaboration among researchers
- Ensure long-term preservation of scientific data



LINKML

What is LinkML?

- A schema language for defining data models
- Supports creation of interoperable data structures
- Emphasizes semantic consistency and data integration

Key Features of LinkML:

- YAML-based schema definitions
- Automatic generation of documentation, JSON schemas, and more
- Support for rich metadata annotations
- Additionally, it is a framework for working with and validating data in a variety of formats (JSON, RDF, TSV), with generators for compiling LinkML schemas to other frameworks and generating documentation.

Datamodel

Development of Data Model Preparation and Validation Tool for Scientific Metadata in SciCat.

Primary Objectives:

1. Data Model Preparation
2. Validation Layer Implementation
3. Documentation Generation
4. Metadata List Management

Roadmap:

1.Centralized Git Repository:

1. Centrally managed and curated repository for schema definitions.

2.CI/CD Pipeline to Generate:

1. Documentation (based on MKDoc)
2. JSON Schema
3. Simple spreadsheet-like view for discussion

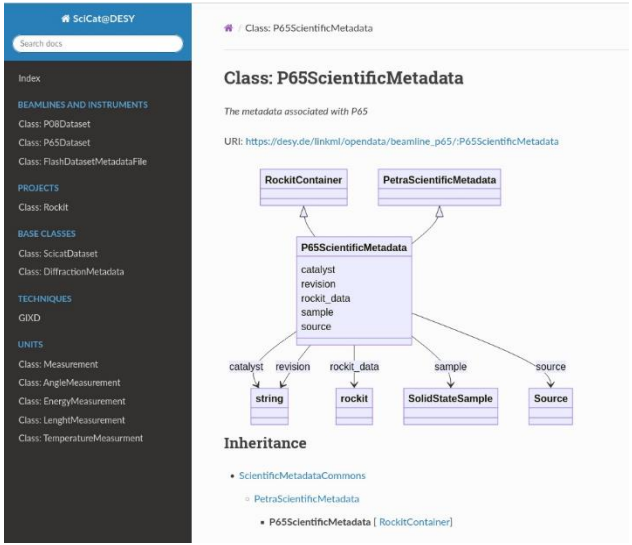
3.Foundation:

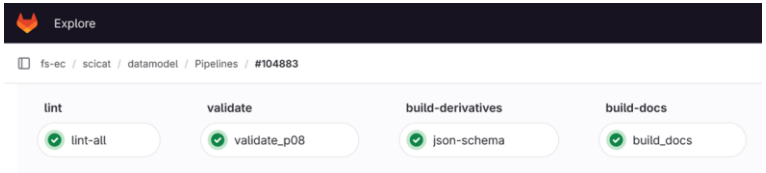
1. Use results of DAPHNE TA1 and TA2 as the foundation for schema creation.

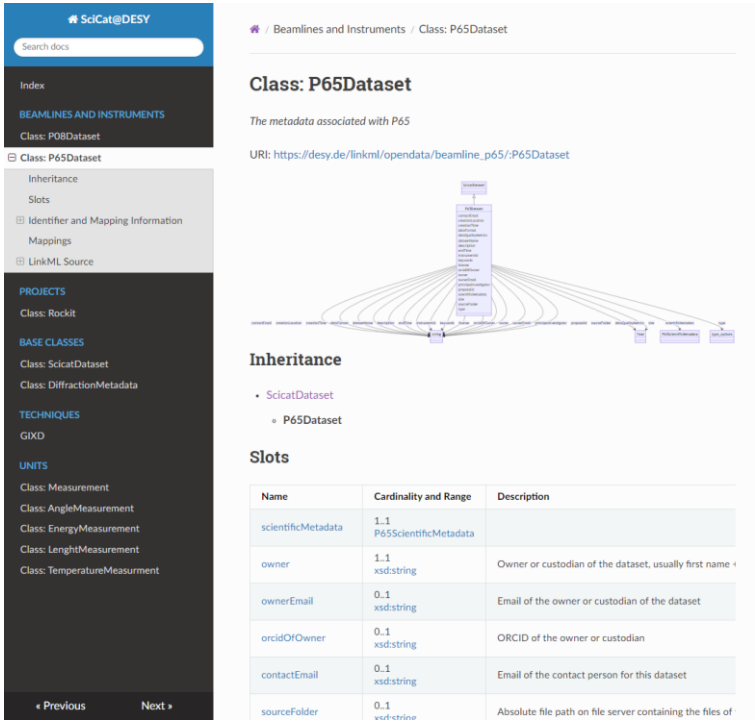
4.Schema Expansion:

1. Expand schemas to technique or beamline-specific use cases.

Data model Preliminary Results







- Gitlab datamodel link - https://gitlab.desy.de/fs-ec/scicat/datamodel
- LinkML Documentation: https://linkml.io/linkml/
- Data Model Documentation: https://fs-ec.pages.desy.de/scicat/datamodel/

Objectives

Through the use of LinkML we aim at providing:

1. The possibility to automatically convert a dataset in the DESY-SciCat into content that is ingestible by external tools through the use of ontologies (e.g. FAIRmat's NOMAD, NFDI4Cat, NFDI4Chem, ISPyB)
2. Provide validator service (ideally backed into SciCat backend) that provides information on whether or not a certain dataset complies with the proposed schema
3. Empower scientists to participate in the schema definition process

A json schema for SciCat's ScientificMetadata that is build of standardized blocks, but customized for each beamline

Your input is welcome!

These are just first steps, nothing caved in stone yet!

We are open for advice!

Thank you for your attention !!