Ontology management for the SciCat catalog using LinkML

Dylan McReynolds[1], Linus Pithan[2], Runbo Jiang[1], Anjali Aggarwal[2], Paul Millar[2], Tim Wezel[2]

[1]Advanced Light Source, [2]DESY

Schemas for Scientific Metadata in SciCat

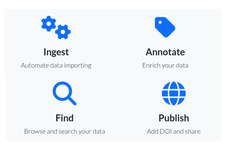
SciCat is a data catalog and web portal storing data set information in use at multiple facilities. SciCat is intentionally flexible about data schemas. A subset of fields - mainly focusing on data management and administrative needs - are required by the catalog application itself. The "Scientific Metadata" however is a freeform dictionary. On purpose, there is no built-in mandatory validation for this type of information by default when it is "ingested" into the catalog. This is to provide maximum flexibility regarding the content of “Scientific Metadata”. However for individual facilities or specific scientific communities agreeing on a standardized structure inside of this scientific metadata is of high importance. We propose using the LinkML toolset to curate and publish schemas that can be used to provide semantic meaning to fields as well as to create detailed data validations of incoming metadata.

SciCat - A MetaData Catalog for Datasets

SciCat facts:

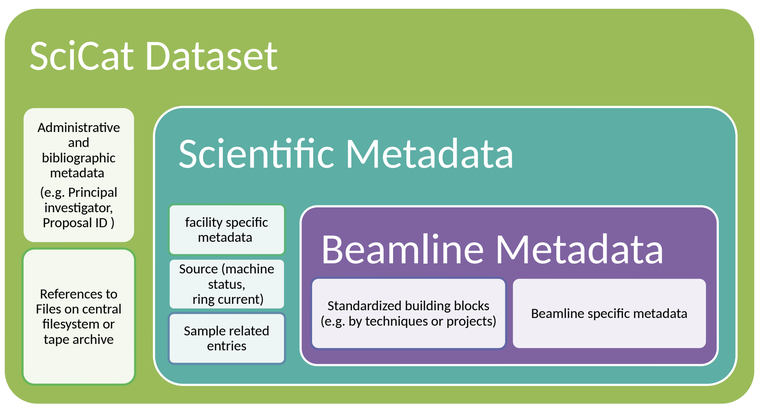
- Web Portal based on common software frameworks
- Developed at light and neutron sources
- Well defined API
- Semantics for "Raw" and "Derived" data sets
- Also stores information about proposals, instruments, samples and more.

https://scicatproject.github.io/



The “Scientific Metadata” Field in SciCat

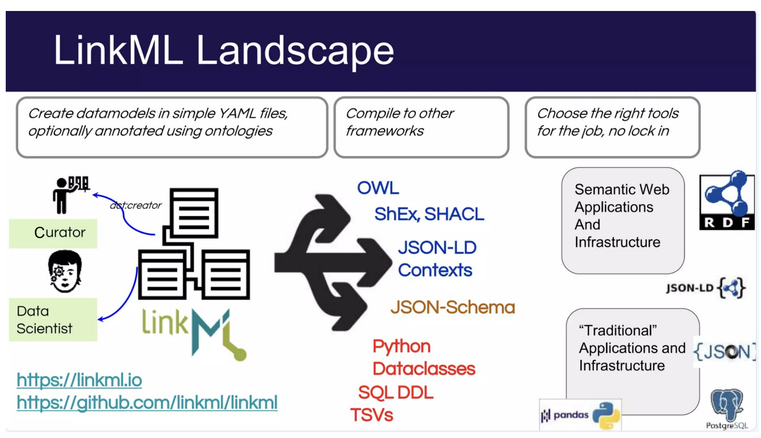
When datasets are introduced into SciCat, they co0ntain a "Scientific Metadata" section. SciCat's Scientific Metadata is completely un-opinionated, allowing for any fields and values to be extracted from datasets and added to the catalog. This flexibility has enabled adoption by a wide variety of facilities, including X-ray sources, neutron sources, and academic groups. However, this flexibility comes at the expense of standardization, documentation, and machine readability.



LinkML – A toolset around data modeling

The LinkML project provides a set of tools facilitating the definition and publication of "schemas," and allows these to be integrated into a workflow.

It enables the definition, maintenance, and interlinking of domain-specific ontologies, and expresses these in a variety of standard definition languages such as JSON Schema, JSON-LD, RDF, and OWL. Further it offers the capability to export human readable documentation and to build workflows known from software development around schema management, e.g. using git repositories, linting and CI/CD

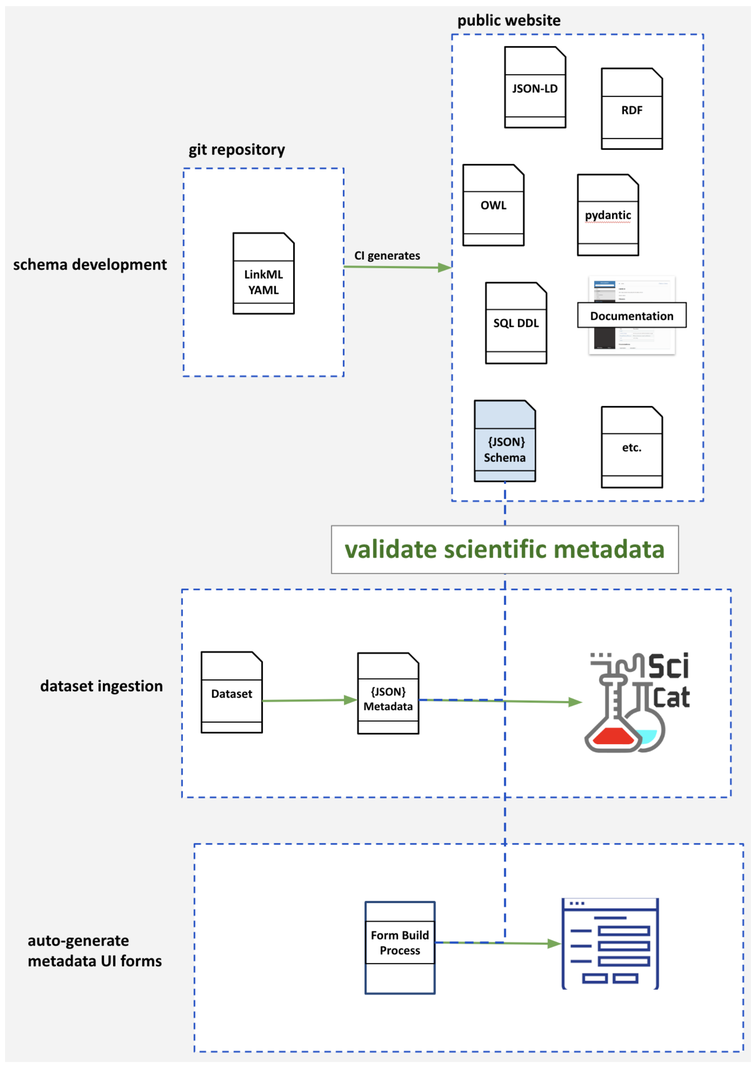


https://doi.org/10.5281/zenodo.7778641

SciCat + LinkML

Here we essentially propose to

- Maintain Schemas outside of Catalog Repository
- Enable schema versioning for scientific metadata in a dedicated Git repository
- Provide Semantic Mappings to other ontologies (e.g. NeXus or CIF/IUCr)
- Use CI/CD pipelines to auto-generate
  - user documentation
  - schema artifacts in various formats
  - Dataset ingestion UI from JSON-Schema
- Schema enforcement could live in the ingestion code and the SciCat server



LinkML, an interesting choice the data catalog context?

- brings together schemas and ontologies
- helps to generate consistent documentations
- provides tooling to handle both ‘classes’ (a.k.a. schemas) and ‘instances’ (a.k.a. datasets)

Future development plans to make metadata semantically accessible in SciCat

- Add JSON-LD context to datasets in SciCat
- Provide json schemas accompanying datasets
- Long term vision: provide access to dataset in SciCat through SPARQL endpoint or export option to triple store

CI/CD Pipeline to auto generate

- Documentation (based on MKDoc)
- Json schema/JSON-LD
- simple spreadsheet-like view schemas for discussion

Schemas and Ontologies for Scientific Metadata

Schemas as tool for standardization

There are different interest groups that look with different eyes on datasets in the PaN community (e.g. visiting scientists, beamline staff, facility data managers, data scientists). While some of the individual interests reflect in the structure of the data catalog itself others don’t. Since SciCat’s Scientific Metadata Section is completely un-opinionated, allowing for any fields and values to be extracted from datasets and added to the catalog it is easily capable to serve the individual needs of the different interest groups. However in order to be able to assure facility wide, coherent metadata structures schemas are a viable options. Schemas enable dataset validation to ensure consistent datasets throughout the full facility.

From schemas to ontologies

Even when having agreed within a facility on a certain vocabulary for it’s catalog there might be conflicting names when trying to join metadata from different facility under a common abstraction. This is specifically true for domain specific terminology (e.g. key names such as ‘sample_temperature’, ‘t_sample’ or even ‘temp’ might be used to express the same meaning in different places). Curating not only schemas but also accompanying ontologies allows to transport the meaning of specific metadata fields and thus provide options to map datasets form different sources against each other.

Semantic access to Scientific Metadata

semantic access to metadata in the catalog will allow to

- increase usability of datasets in meta-studies
- increase the level of “AI-readiness” of datasets
- enable embedding of datasets into a knowledge graph
- link and map dataset to domain specific data collections
- supports definition of maintainable metadata structure
- helps to define interfaces to upstream data-harvesting services

This work was supported as part of the Center for Materials for Water and Energy Systems MWET, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award #DE-SC0019272.

MWET combines researchers from University of Texas, UC Santa Barbara and LBNL studying membrane for separation. Many techniques are involved. Datasets are often shared across the various facilities using SciCat. We investigate defining LinkML schemas to define metadata for laboratory studies like NMR.

This work is supported by the consortium DAPHNE4NFDI in the context of the work of the NFDI e.V. - The consortium is funded by the DFG (project number 460248799)