Introduction to SampleDB

September 25th 2023 | Florian Rhiem | PGI / JCNS-TA — DMA ST1 Seminar

SampleDB

Overview

- Electronic Lab Notebook
- Metadata Database
- Originally developed as a sample database → SampleDB
- Open Source
- https://go.fzj.de/sampledb
- https://github.com/sciapp/sampledb/

SampleDB

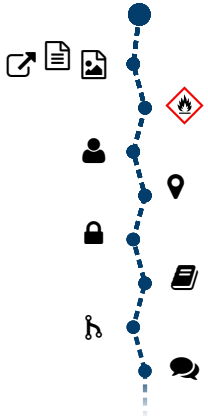
Basic Questions

- What metadata do we need?
- How can research data & metadata be . . .
  - . . . stored?
  - . . . found?
  - . . . analyzed?
  - . . . published?
  - . . . shared?

SampleDB

Metadata – What metadata do we need?

- Process-specific metadata
- Images, Log Files, Links
- Hazards
- Responsible Users
- Storage Location
- Permissions
- Publications
- Derived Objects
- Comments
- SampleDB tracks full life cycle



JÜLICH Forschungszentrum

SampleDB

Schemas – How can process-specific metadata be stored?

Plain text

- No restrictions
- No validation
- No meaning to system
- Only text search
- ▸ “lowest common denominator”

Process-specific formats

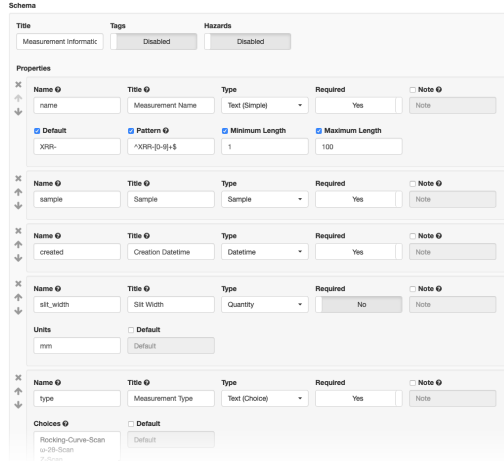
- Forms for entering data
- Validation possible
- Values can have meaning beyond text
- Search using quantities, dates, etc
- ▸ Data far more likely to become useful

How to define new processes?

SampleDB

Schemas – How can process-specific metadata be stored?

- Schemas
- Domain-specific language for formats
- Schemas can be as generic or specific as necessary
- Easy method for defining new formats
- Clear separation between formats and code
- Graphical editor for basic schemas
- Users can define schemas for new processes



JÜLICH Forschungszentrum

SampleDB

Search – How can research data be found?

- Text-based Search
  - Exact match necessary

    `0.2mm`
- Extended Search
  - Attribute Comparison

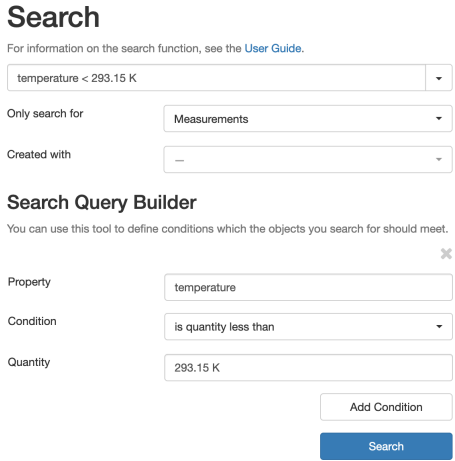
    `slit_width == 0.2mm`
  - Automatic Unit Conversion

    `slit_width == 200000nm`
  - Boolean Algebra

    `slit_width == 200000nm and step_time < 1s`
  - Query Visualization

  - Graphical Search Query Builder



JÜLICH Forschungszentrum

SampleDB

Jupyter Notebook Templates – How can research data be analyzed?

- Jupyter Notebooks for data analysis
- SampleDB
- \+ Notebook Template Server
- \+ JupyterHub
- ▸ Process-specific data analyses prepared by instrument scientists
- Ideal for new users, previews, etc.

In [ ]:
```
import glob
from ipywidgets import interact, interactive, fixed, interact_manual
import ipywidgets as widgets
import numpy as np
import gr
from gr.pygr import mlab
import fourcircle

gr.inline()
```

In [ ]:
```
measurement_id = 3166
```

In [ ]:
```
measurements = []
measurement_directory = f'/Data/instruments/FZJ/4circle/iffsamples_sync/{measurement_id}/'
spec_file_names = list(glob.glob(measurement_directory + '*.spec'))
print(f"Found {len(spec_file_names)} spec files:")
print('\n'.join(spec_file_names))
```

In [ ]:
```
for spec_file_name in spec_file_names:
    measurements.extend(fourcircle.read_spec_file(spec_file_name))
```

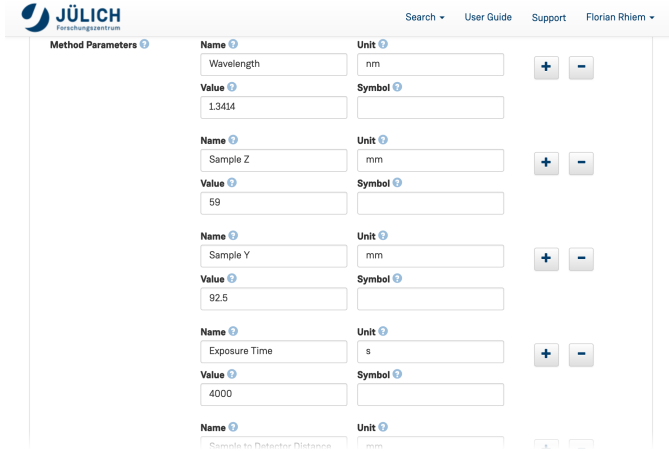
In [ ]:
```
# bounding box - either set these values or leave them as is to determine automatically
h_min = np.inf
h_max = -np.inf
k_min = np.inf
k_max = -np.inf
l_min = np.inf
l_max = -np.inf
```


JÜLICH Forschungszentrum

SampleDB

Dataverse Export – How can metadata be published?

- Export from SampleDB to Dataverse, e.g. Jülich DATA
- Metadata Blocks based on Community Standards
- *Citation Metadata* Block for common metadata
- *Process Metadata* Block for process-specific metadata
- Researcher can edit and then publish draft dataset



JÜLICH Forschungszentrum

SampleDB

SciCat Export – How can metadata be shared?

- Export from SampleDB to SciCat
- Researcher decides which properties to export
- Maps SampleDB action types to SciCat
  Sample, *Raw Dataset*, *Derived Dataset*

SampleDB

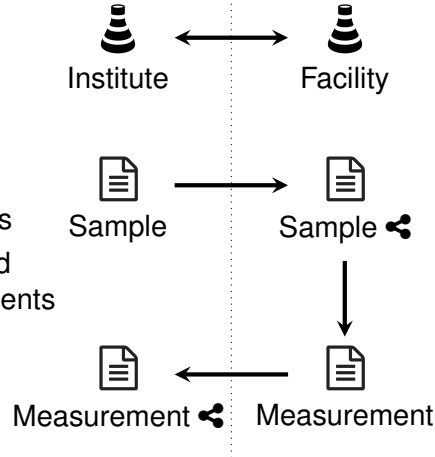
ELN Export – How can metadata be shared?

- Export from SampleDB to .eln file format
- Common file format for Electronic Lab Notebooks (e. g. elabFTW, Kadi4Mat, Juliabase, . . .)

SampleDB

Federation – How can metadata be shared?

- Sharing of data between SampleDB instances
- Administrators can configure which other instances their users may share data with
- Users share objects with users from other instances
- Enables simple collaboration between institutes and facilities, e. g. for measurements at specific instruments



JÜLICH Forschungszentrum

SampleDB

Summary

- SampleDB can track the full sample life cycle, including process-specific metadata
- Features
  - Schemas define process-specific metadata
  - Text-based and extended search
  - Jupyter Notebook templates for prepared data analyses
  - Dataverse Export for publishing metadata
  - SciCat export, .eln export and SampleDB federation for sharing metadata
- Open Source
- https://go.fzj.de/sampledb
- https://github.com/sciapp/sampledb/

Questions?

f.rhiem@fz-juelich.de
https://go.fzj.de/sampledb
https://github.com/sciapp/sampledb/