PSI Center for Scientific Computing, Theory and Data

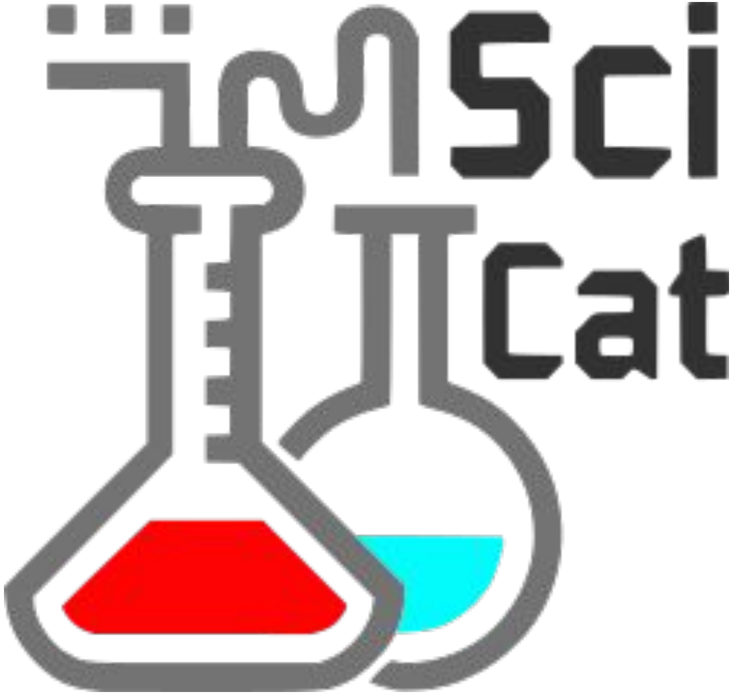
Standardising metadata between catalogues

Grenoble, 2024/09

Paul Scherrer Institut

The tradeoff

Flexibility



Searchability

The struggles

- Many fields of research, with heterogeneous metadata
  - → **no** agreement on a **global** metadata **standard**
- Resistance from scientists to catalogue data
  - → **schema** should maintain some **flexibility**
- How can we improve searchability and standardisation with these constraints?
- How to reference data?

The middleground

- Define metadata **schemas** per **domain**, discussed with domain scientists
- Metadata entities should **reference the schema**
- The schema is not enforced by the underlying database, but optionally as part of **data validation**, during ingestion
- Not all metadata will be standardised

Long term solution

- JSON-Schema or RDF with preferred encodings (XML, JSON-LD, turtle…)
- Schema definitions deposited on a public platform
- LinkML for easier schema definition and JSON-schema or RDF conversion
- Schema validation depending on domain, for related metadata
- LinkML-map for schema to data catalogue structure conversion

Maybe a start?

We already have a common high-level metadata format, and a common protocol:

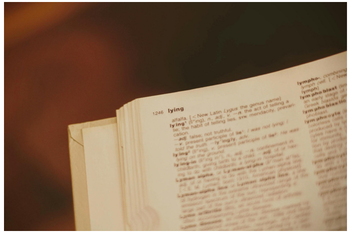
OAI-PMH with datacite

Could we build an importer from it?

Schemas and Ontologies

Ontology

Defines terms and relationships between terms

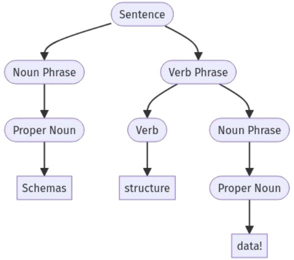


OWL

Semantic reasoners infer meaning in different contexts

Schema

Defines the data structure and validation rules



LinkML, JSON Schema

Validate data syntax

Data Serialization

The file format for the data

DOCX

JSON, YAML, RDF, XML

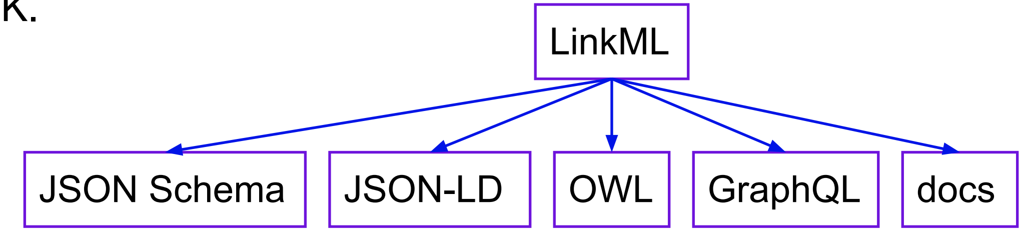
Convert between formats

Open Science Community for Electron Microscopy (OSC-EM)

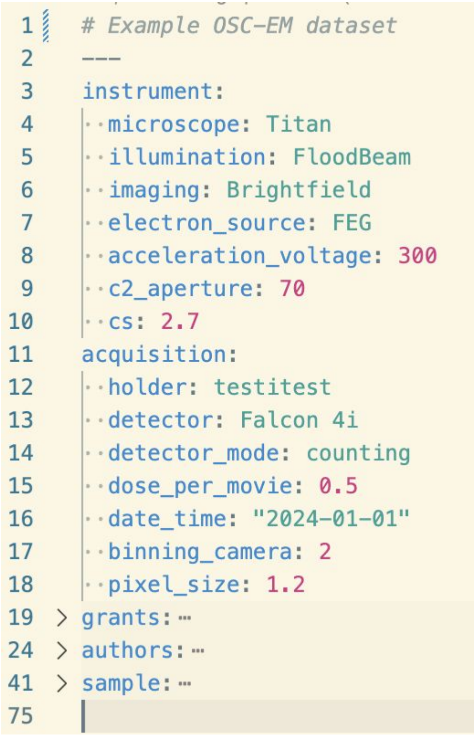
- Established in 2023 to bring together electron microscopy (EM) researchers, facilities, software developers, and data repositories to standardize EM metadata needed for data collection, processing, and deposition.
- Workshop 22-23 Feb 2024 with diverse participants
- Aims for interoperability with other ontologies and standards: CryoEM ontology, PDBx/mmCIF dictionary, Helmholz EM Glossary, NeXus-FAIRmat NXem format
- Active contributors include OpenEM facilities (swissopenem.github.io), the Instruct Image Processing Center (I2PC), and the EM Data Bank (www.ebi.ac.uk/emdb).
- Modular definition for different experimental methods and data processing stages (cryoEM, tomography, EELS, 3D reconstruction, etc).

OSC-EM Schema

- https://github.com/osc-em
- Schema in LinkML used to automatically generate JSON Schema, JSON-LD, OWL, GraphQL, etc, as well as documentation and a python SDK.



- Import from SerialEM and Thermo Fischer EPU (more coming!)
- Export to mmCIF for deposition in EMDB/PDB OneDep
- Suitable for inclusion in SciCat `scientificMetadata` field (validation coming soon).



```yaml
# Example OSC-EM dataset
---
instrument:
  microscope: Titan
  illumination: FloodBeam
  imaging: Brightfield
  electron_source: FEG
  acceleration_voltage: 300
  c2_aperture: 70
  cs: 2.7
acquisition:
  holder: testitest
  detector: Falcon 4i
  detector_mode: counting
  dose_per_movie: 0.5
  date_time: "2024-01-01"
  binning_camera: 2
  pixel_size: 1.2
grants: ⋯
authors: ⋯
sample: ⋯
```

Metadata Use Cases

-