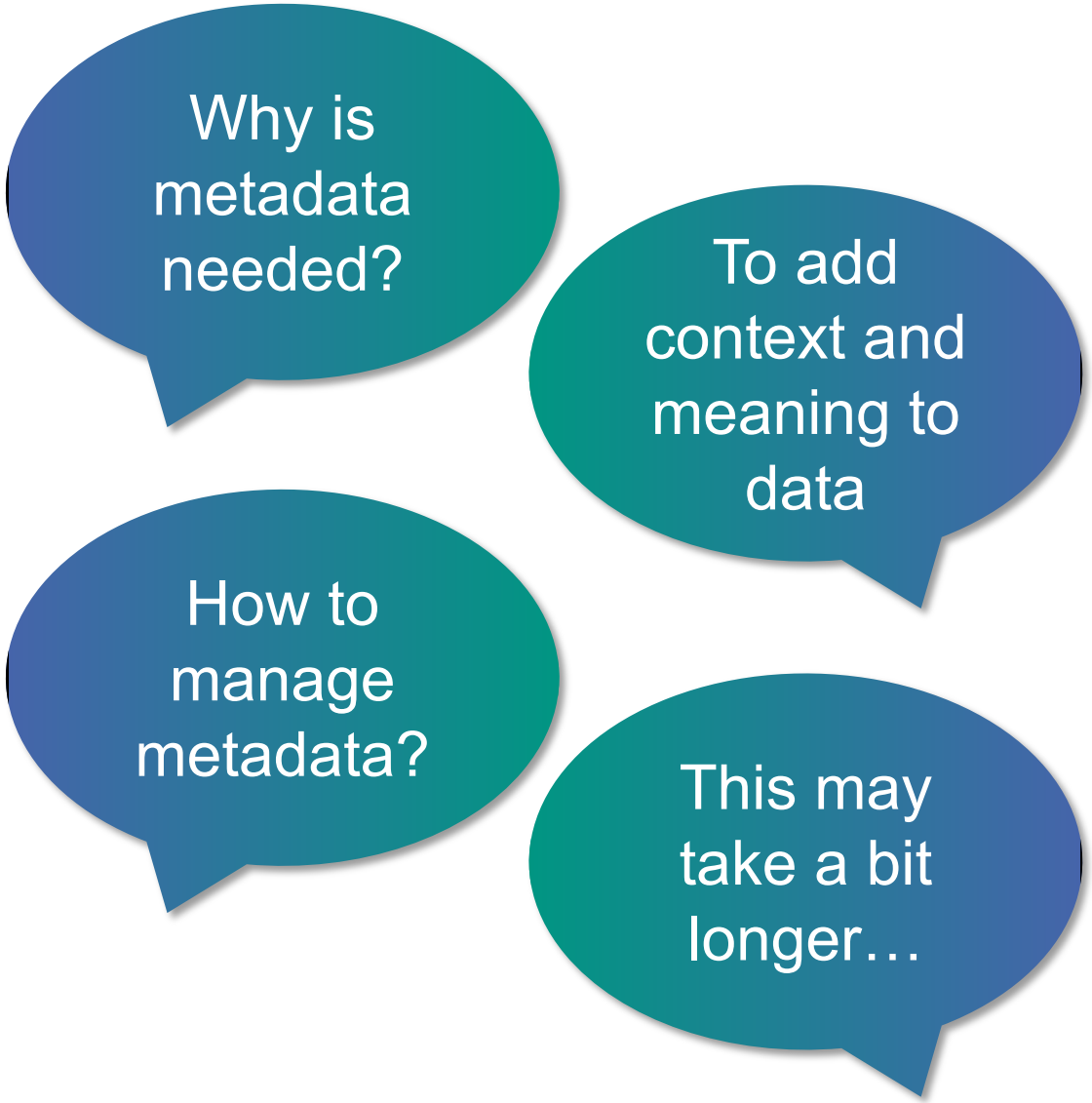
Handling and Storing Metadata

Rossella Aversa (KIT-SCC)

KNMFi User Meeting, 28.11.2023

Motivation


Why is metadata needed?
To add context and meaning to data
How to manage metadata?
This may take a bit longer…

FAIR Guiding Principles

Globally unique persistent identifiers

(Meta)data repositories

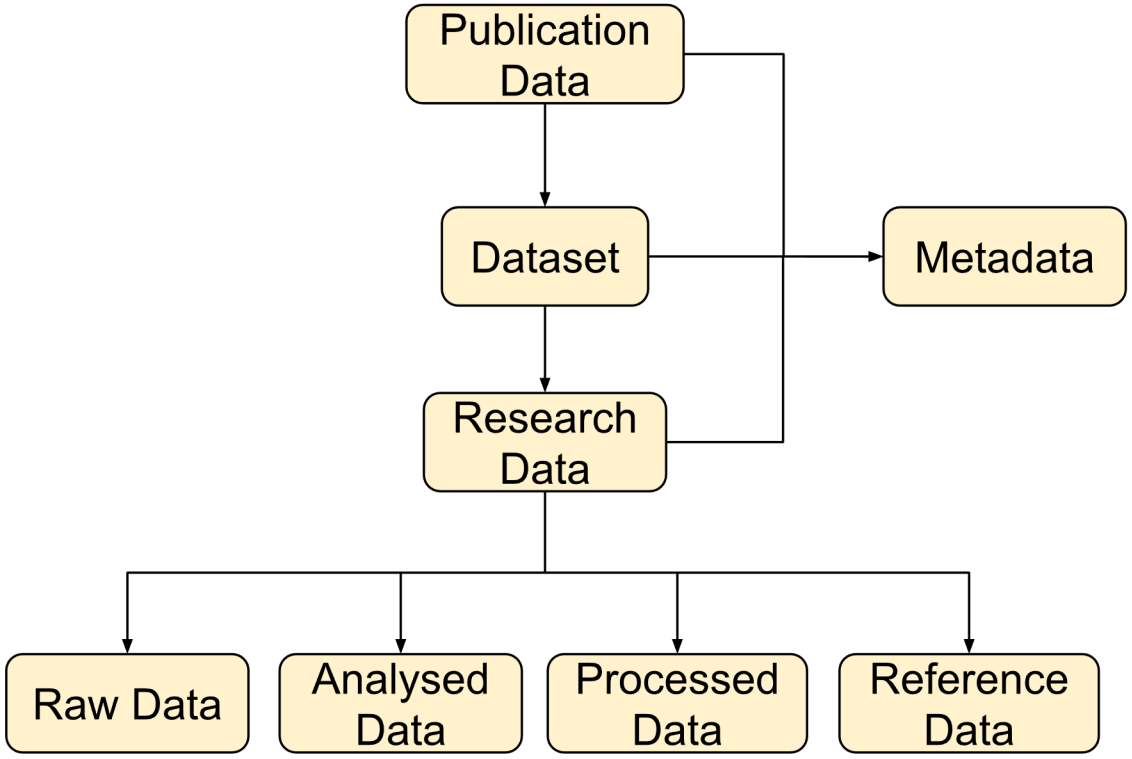
Structured metadata (schemas, vocabularies)

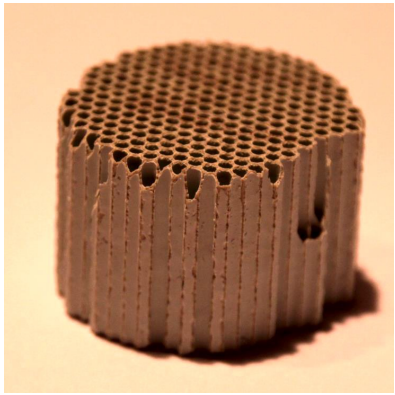
Provenance (ELNs)

https://www.go-fair.org/fair-principles/

What to describe?

- R1: *"Metadata should richly describe the data with a plurality of accurate and relevant attributes"*





Sample, courtesy of M. Mail

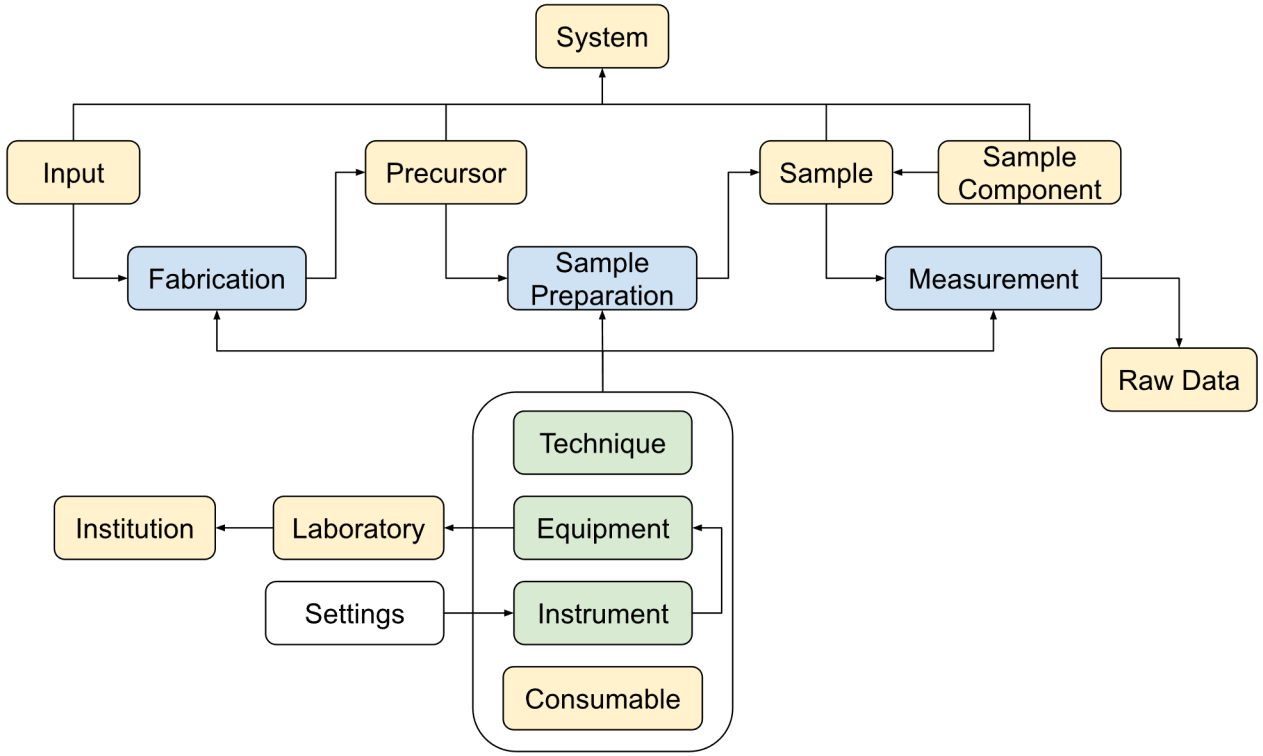
https://www.azonano.com/images/equipments/EquipmentImage_485.jpg

Atomic Force Microscope

How to structure metadata?

R1.3: *“Metadata meet domain-relevant community standards or best practices”*

- JL-MDMC Metadata WG:
  - Adoption of existing solutions
  - Common design
  - Output harmonization
  - MDMC-NEP Glossary
  - Workflows
  - Metadata schemas

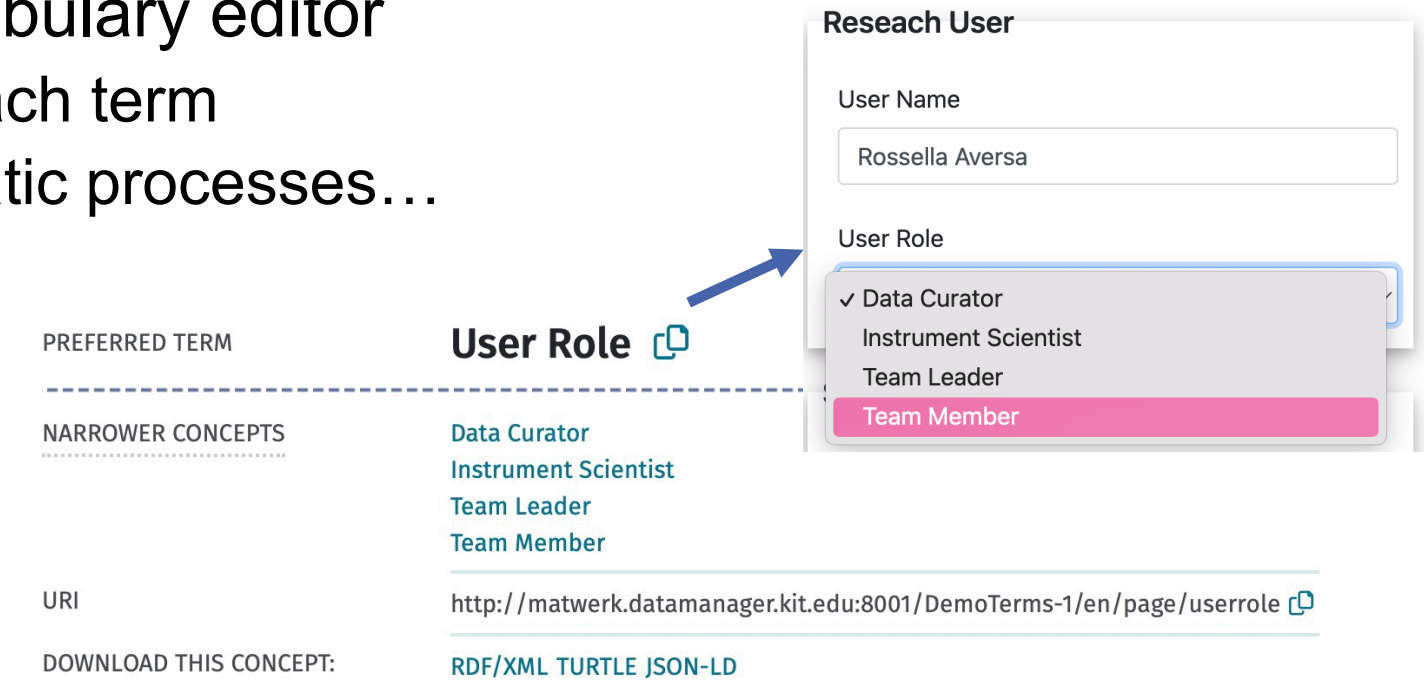


https://jl-mdmc-helmholtz.de/mdmc-activities/metadata-working-group/

How to describe data?

I2: *“Metadata use vocabularies that follow the FAIR principles”*

- EVOKS Vocabulary Service:
- Collaborative online vocabulary editor
  - Persistent identifier to each term
  - Websites, ELNs, automatic processes…
  - Centrally maintained



How to represent structured metadata?

I1: *“Metadata use a formal, accessible, shared and broadly applicable language for knowledge representation”*

JSON Metadata Schema

```
"pressureDetails":{
   "type":"object",
   "description":"(Required) - Descri
   "additionalProperties":false,
   "properties":{
      "value":{
         "type":"number",
         "default":-9999,
         "description":"(Required) -
      },
      "unit":{
         "type":"string",
         "default":"Pa",
```

Outline of the overall metadata structure

JSON Metadata Document

```
"gunPressure": {
  "value": 0.0000000373,
  "unit": "mbar"
},
"angleToEBeam": {
  "value": 54,
  "unit": "degree"
}
```

Structured information about a data resource

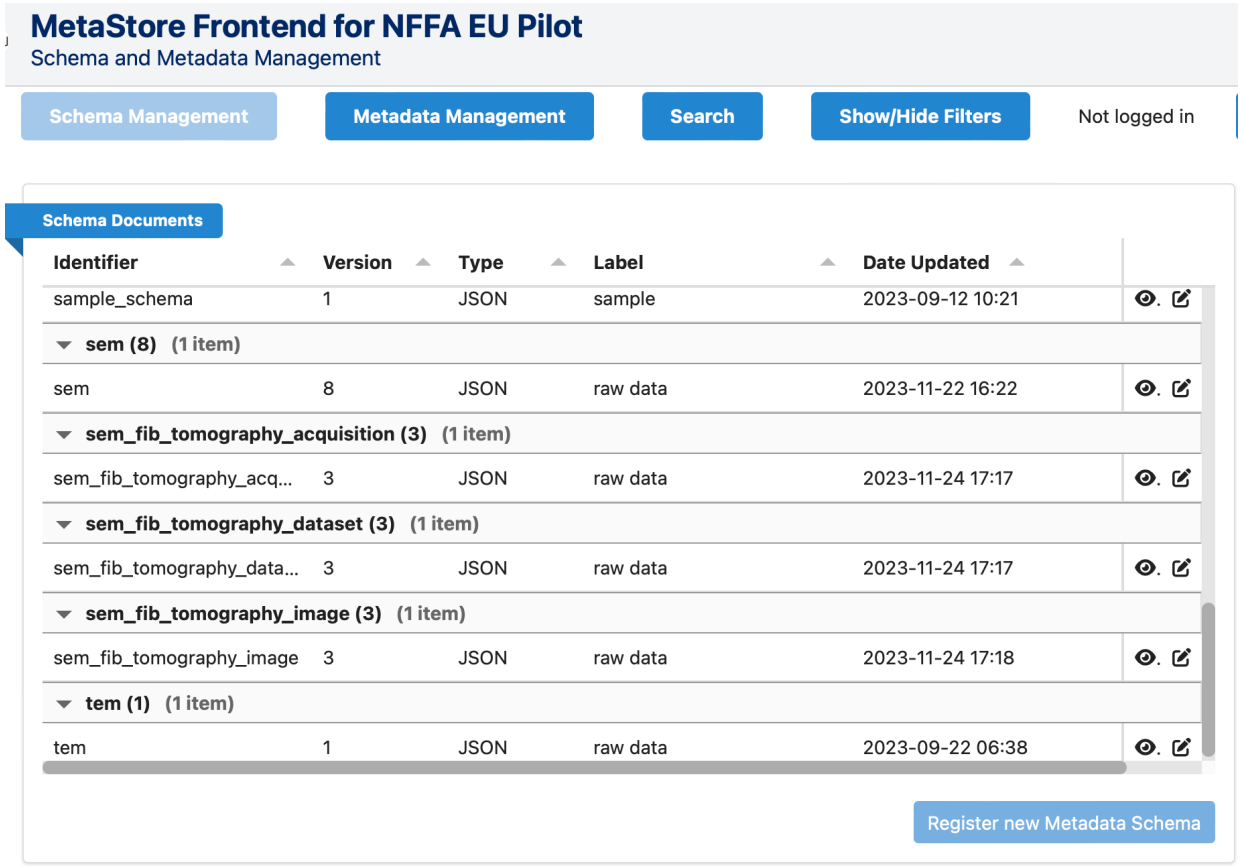
DOI: 10.5445/IR/1000141604

Where to store metadata?

F4: *“Metadata are registered or indexed in a searchable resource”*

- **Metadata repository:** store, manage and provide access to Metadata, following a policy or a set of rules that define storage and access norms. (from MDMC-NEP Glossary)
- MetaStore
  - Metadata schemas

https://metarepo.nffa.eu/

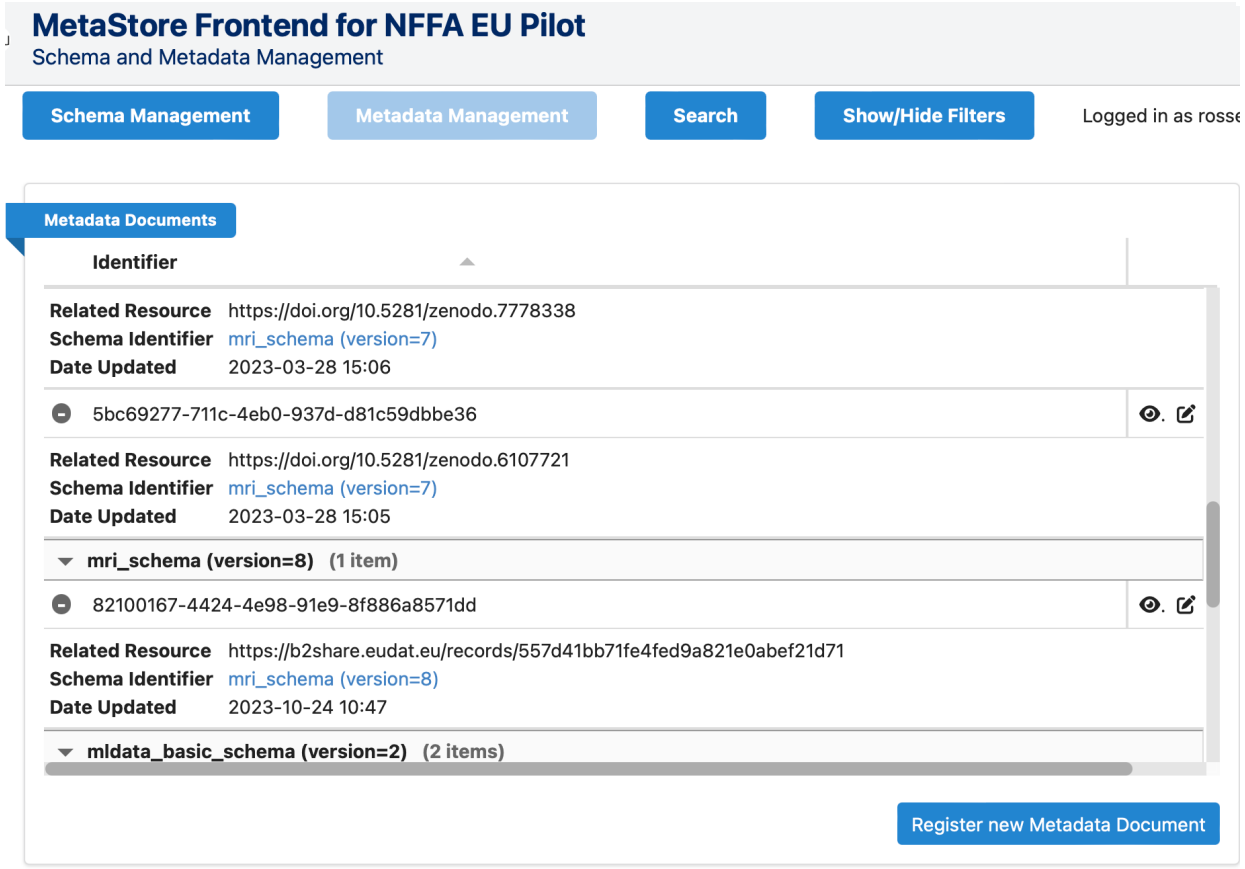


Where to store metadata?

F4: *“Metadata are registered or indexed in a searchable resource”*

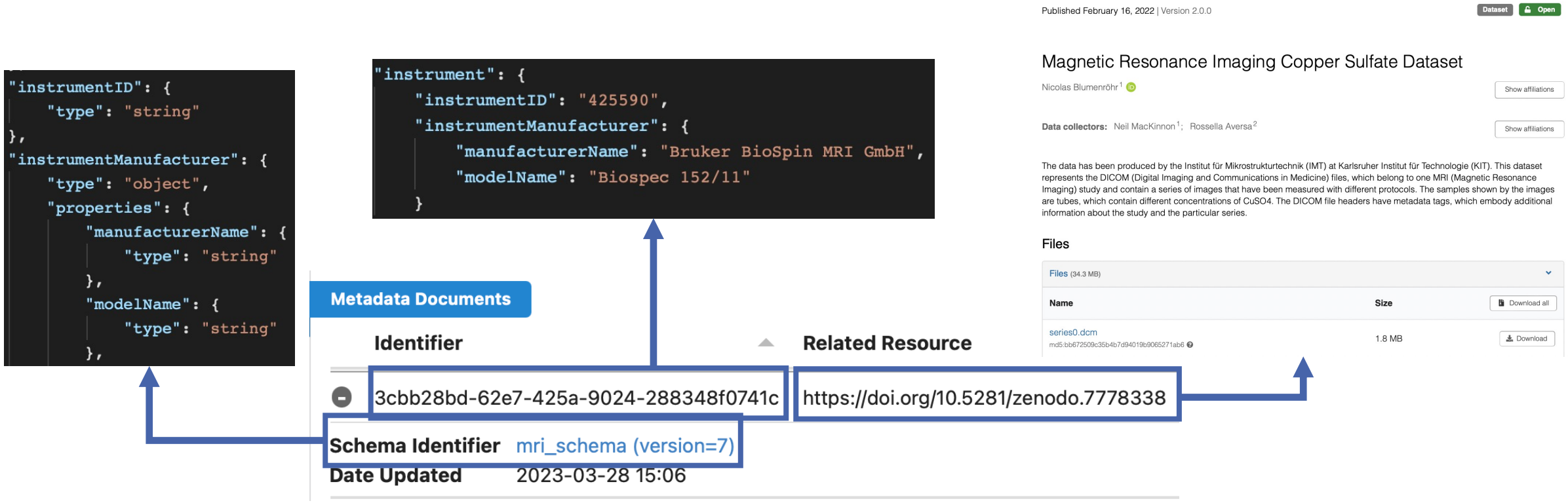
- **Metadata repository:** store, manage and provide access to Metadata, following a policy or a set of rules that define storage and access norms. (from MDMC-NEP Glossary)
- MetaStore
  - Metadata schemas
  - Metadata documents

    https://metarepo.nffa.eu/



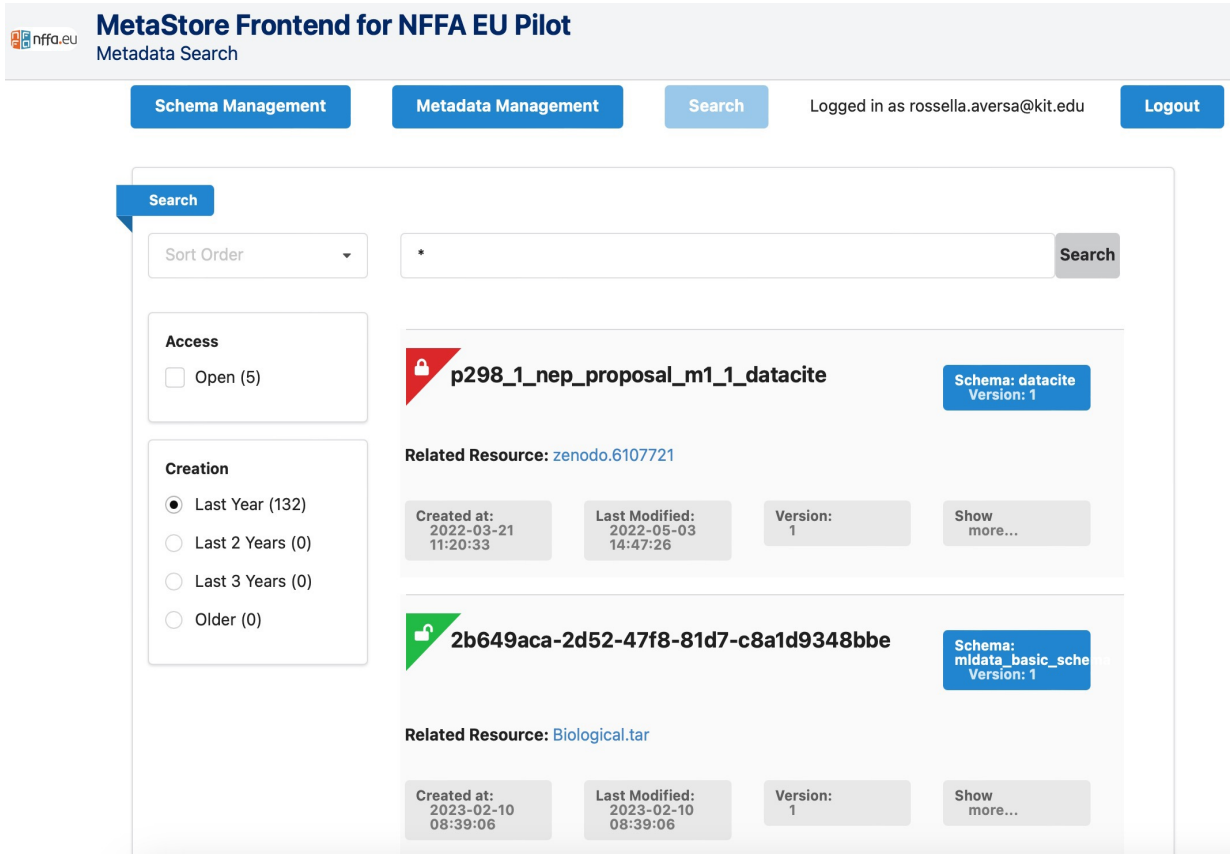
How to link data and metadata?

F1: *“(Meta)data are assigned globally unique and persistent identifiers”*
F3: *“Metadata include the identifier of the data they describe”*



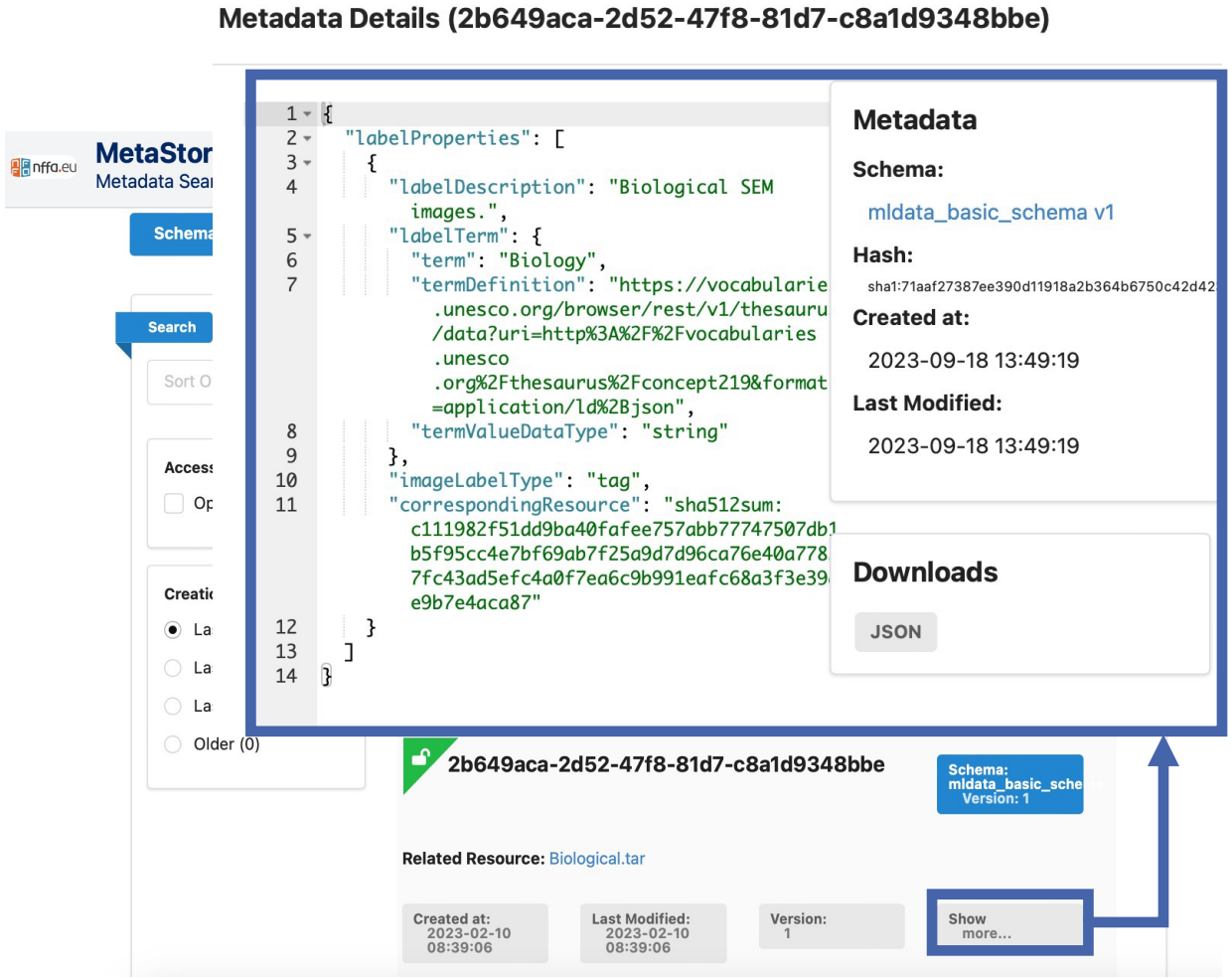
How to find data from metadata?

- Use the content of metadata documents to search for relevant data
- What is the data about? Is it useful for my needs?
- Full-text search
- (basic, customizable) faceted search
- Private vs Public resources



How to find data from metadata?

- Use the content of metadata documents to search for relevant data
- What is the data about? Is it useful for my needs?
- Full-text search
- (basic, customizable) faceted search
- Private vs Public resources

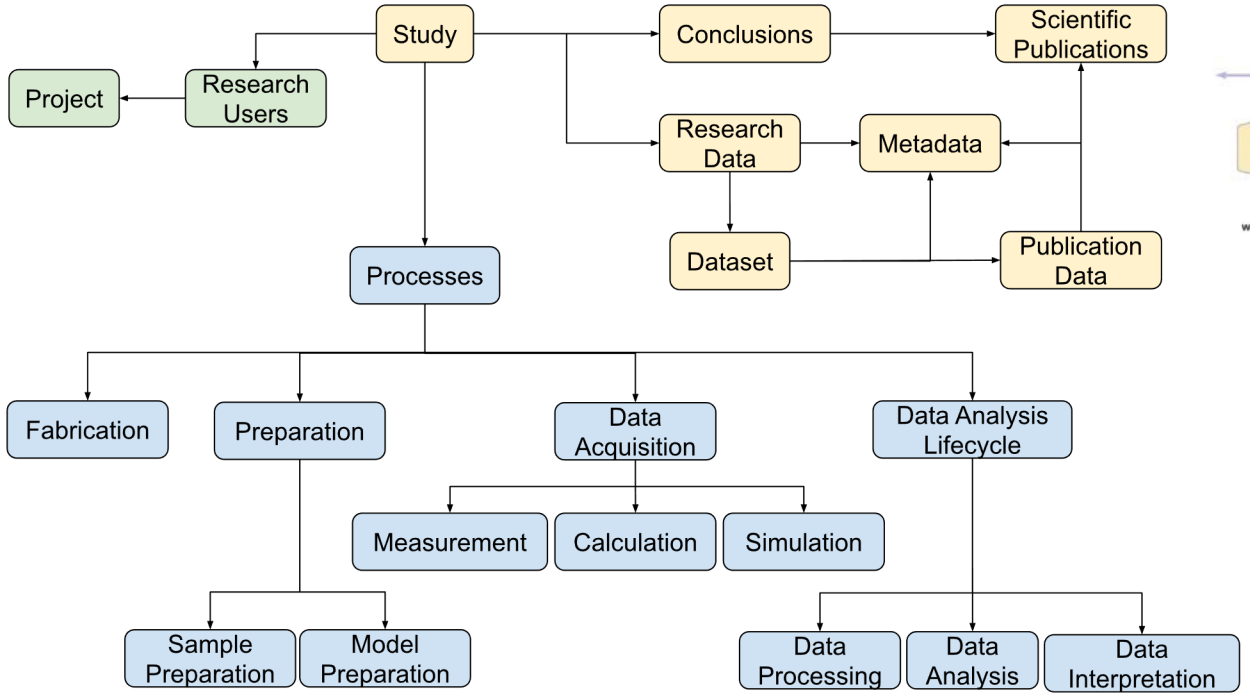


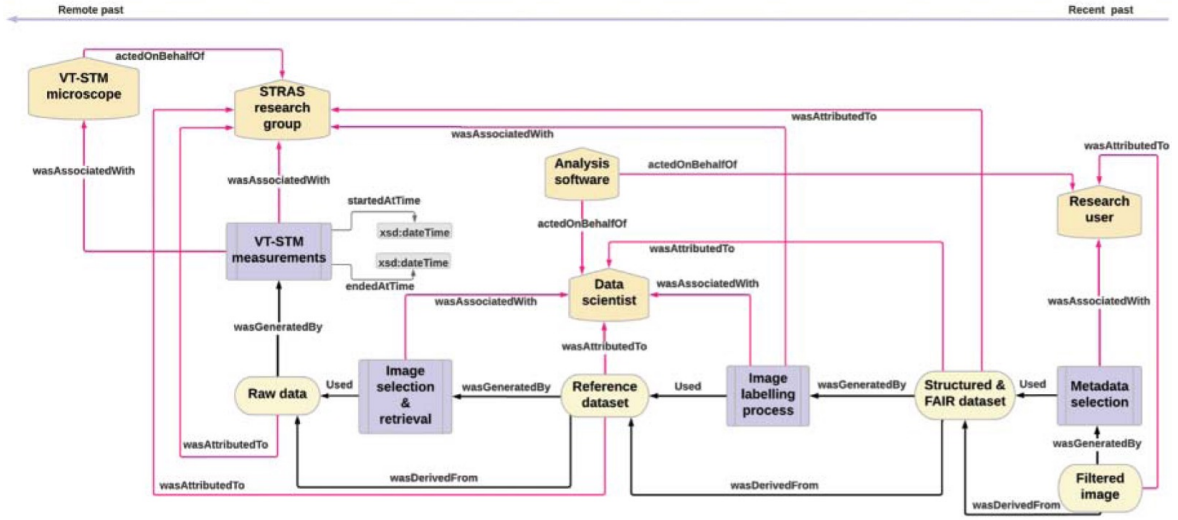
What are the features of MetaStore?

- Versioning
- Automatic validation of records
- Access control management of each record
- Search
- Supports an arbitrary number of XML and JSON schemas
- Structure of the records based on DataCite standard
- Both GUI and API (customizable)
- Possibility to install local instances (e.g., for high performance storage)
- Open source https://github.com/kit-data-manager/metastore2

How to reproduce a research item?

R1.2: *“Metadata are associated with detailed provenance”*





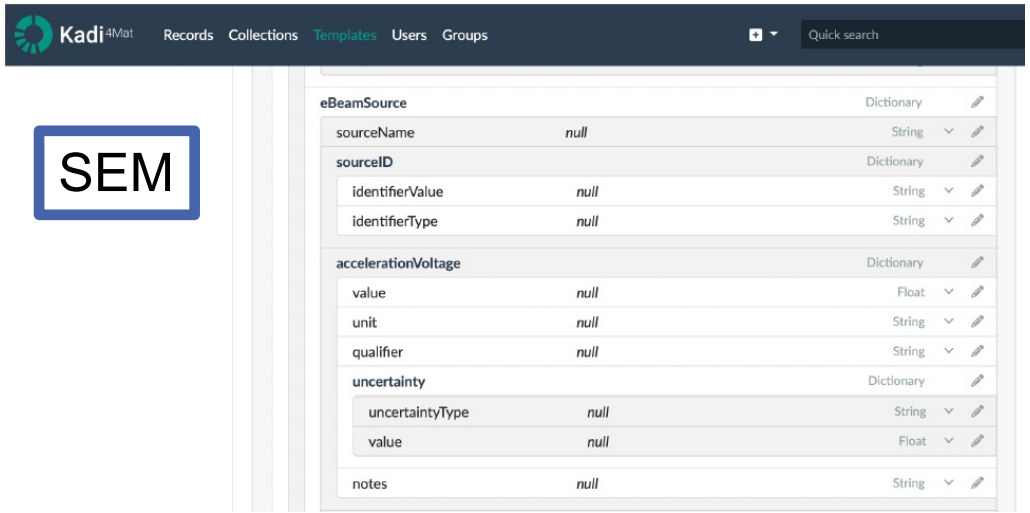
FAIRification of STM Images

DOI: 10.1162/dint_a_00164

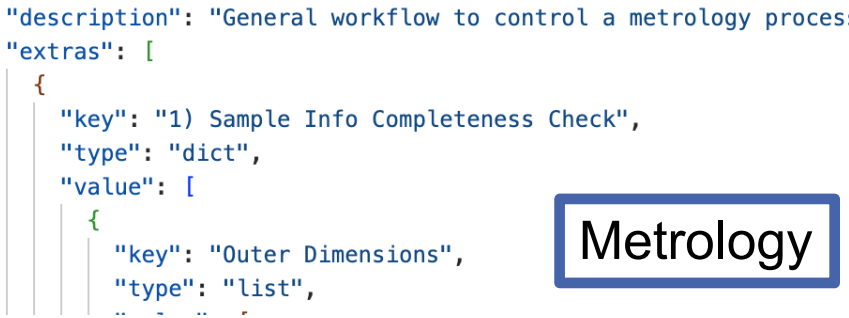
How to achieve it in practice?

Metadata schemas as ELN templates

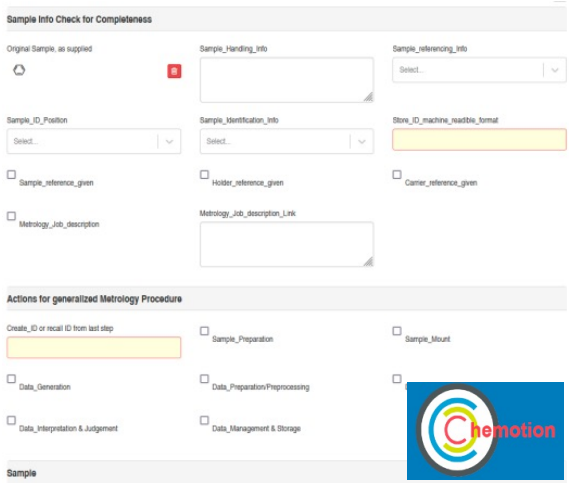
- Collaborations ongoing with:
  - Kadi4Mat (M. Selzer)
  - Chemotion (N. Jung)
  - IAM (S. Schlabach, D.-V. Szabo)
  - INT (C. Kübel, A. Boubnov)
  - IMT (R. Thelen)



SEM



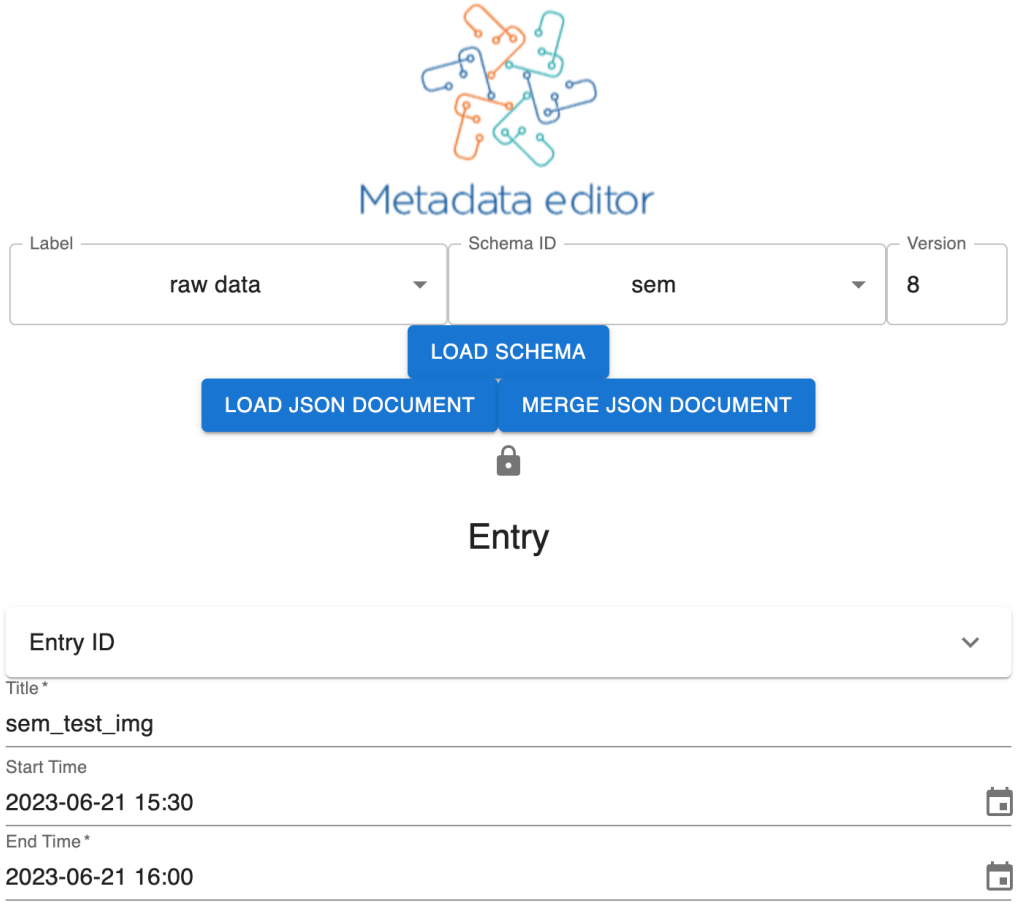
Metrology



How to achieve it in practice?

Metadata editor

- Edit metadata documents according to schemas
- Register metadata documents to MetaStore via the GUI
- Local application (exposes API for instruments/ELNs)

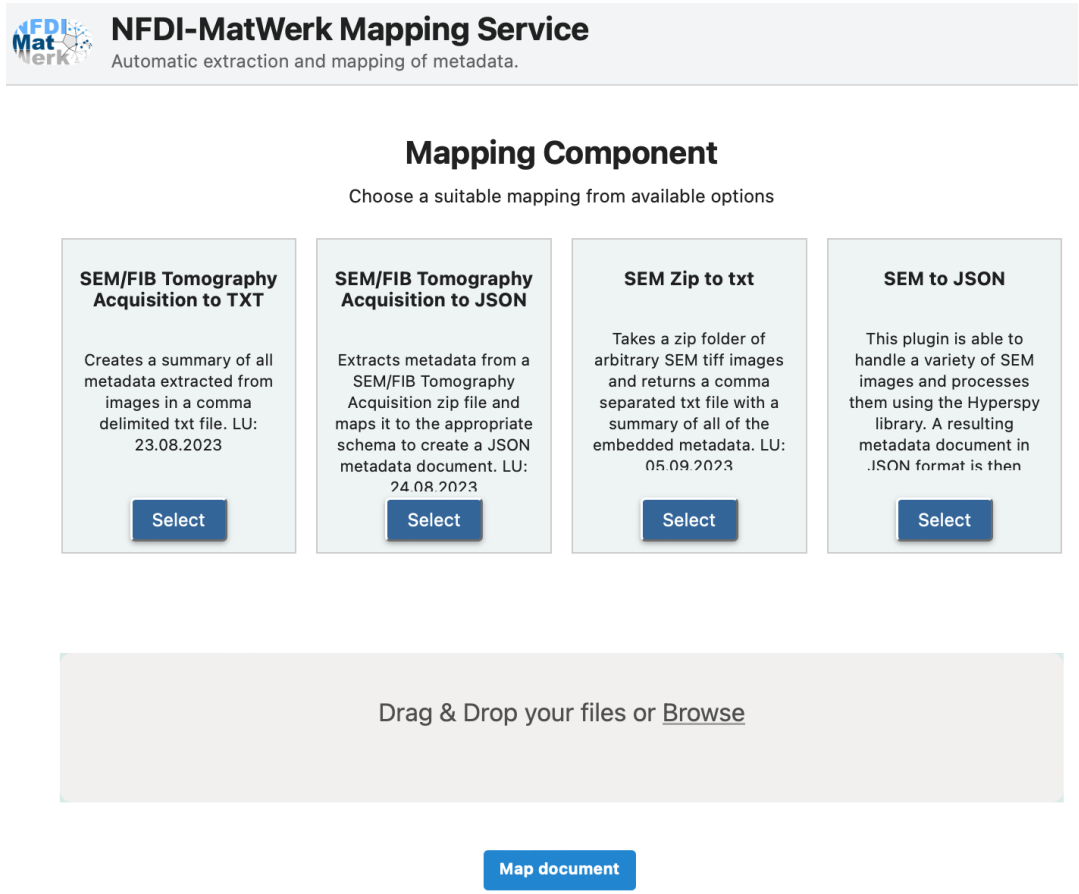


https://metadata-editor.gitlab.io/documentation/

How to automate the process?

Mapping Service

- Input: image, zip file, …
- Output: (incomplete) metadata document
  - Load into Metadata Editor
  - Load into ELN
- Collaboration ongoing with Chemotion (N. Jung)
- New plugins continuously under development



https://matwerk.datamanager.kit.edu/mapping-service-ui.html

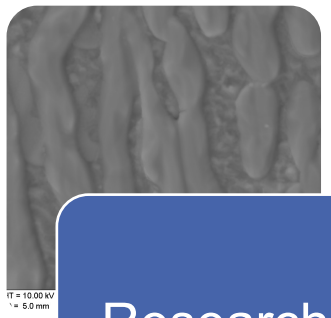
General Metadata Management Workflow

Metadata extraction

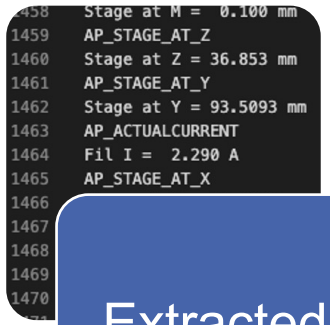
Mapping service

Metadata editor / ELN

MetaStore



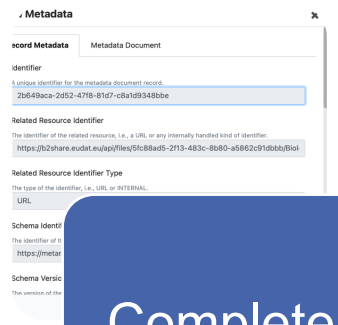
Research Data



Extracted Metadata

Incomplete JSON md document

Complete JSON md document



Complete md record

Related to the data it describes via Persistent Identifier

Summary and conclusions

- **Vocabularies:** meaning/context of metadata is unambiguously described
- **Vocabulary service:** metadata can be referenced elsewhere while centrally maintained
- **Metadata schemas:** structured metadata can be interpreted (also by machines), data can be compared
- **Provenance metadata:** data can be assessed/reproduced
- **ELNs/metadata editors:** metadata can be easily tracked
- **Mapping tool:** (some) metadata can be automatically mapped to schema
- **(Meta)data repositories:** data can be searched, accessed, reused and referenced from outside.

- **Contacts:** rossella.aversa@kit.edu
- **Acknowledgements to:** G. Abdildina, N. Blumenröhr, F. Ernst, V. Hartmann, M. Inkmann, T. Jejkal, A. Kirar, E. Vitali, NEP JA6 members, JL-MDMC Metadata WG members, and all the colleagues who contributed with tests, feedbacks and consultancy
- **Funded by:** the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460247524; the Joint Laboratory Model and Data-driven Materials Characterization (JL MDMC), a cross-centre platform of the Helmholtz Association; NFFA-Europe-Pilot (EU H2020 - n. 101007417); the research program 'Engineering Digital Futures' of the Helmholtz Association of German Research Centers, the Helmholtz Metadata Collaboration Platform.