

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



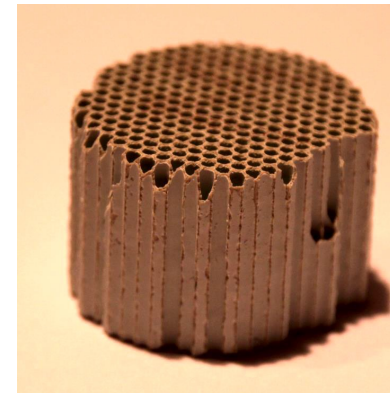
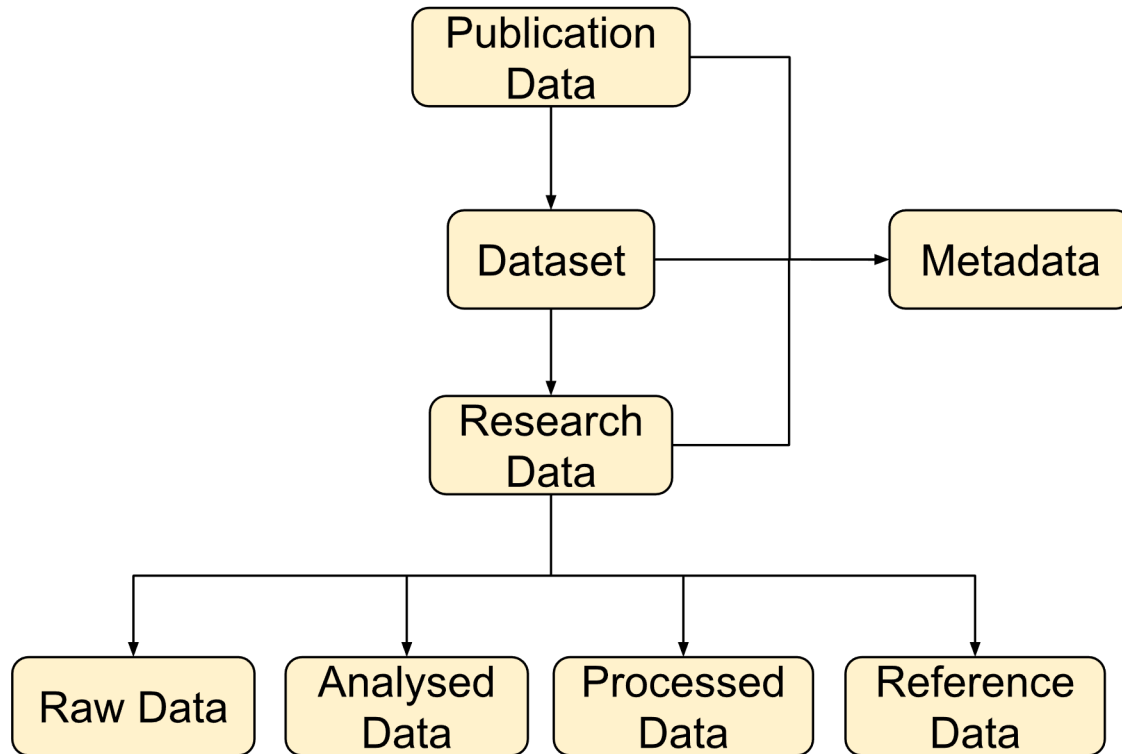
Structured metadata
(schemas, vocabularies)

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



Atomic Force
Microscope

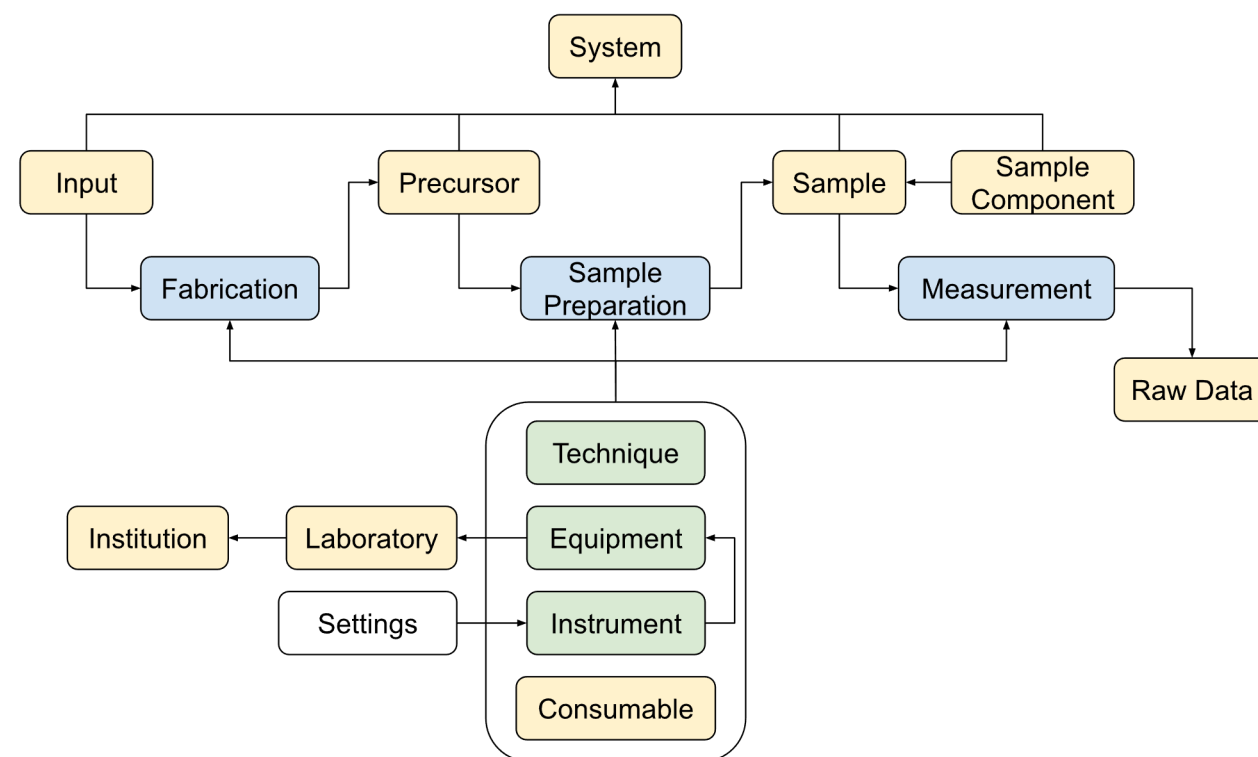
https://www.azonano.com/images/equipments/EquipmentImage_485.jpg

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

PREFERRED TERM

NARROWER CONCEPTS

URI

DOWNLOAD THIS CONCEPT:

User Role 

Data Curator
Instrument Scientist
Team Leader
Team Member

<http://matwerk.datamanager.kit.edu:8001/DemoTerms-1/en/page/userrole> 

[RDF/XML](#) [TURTLE](#) [JSON-LD](#)

Research User

User Name

User Role

- Data Curator
- Instrument Scientist
- Team Leader
- Team Member



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) - Descrip
  "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](https://doi.org/10.5445/IR/1000141604)

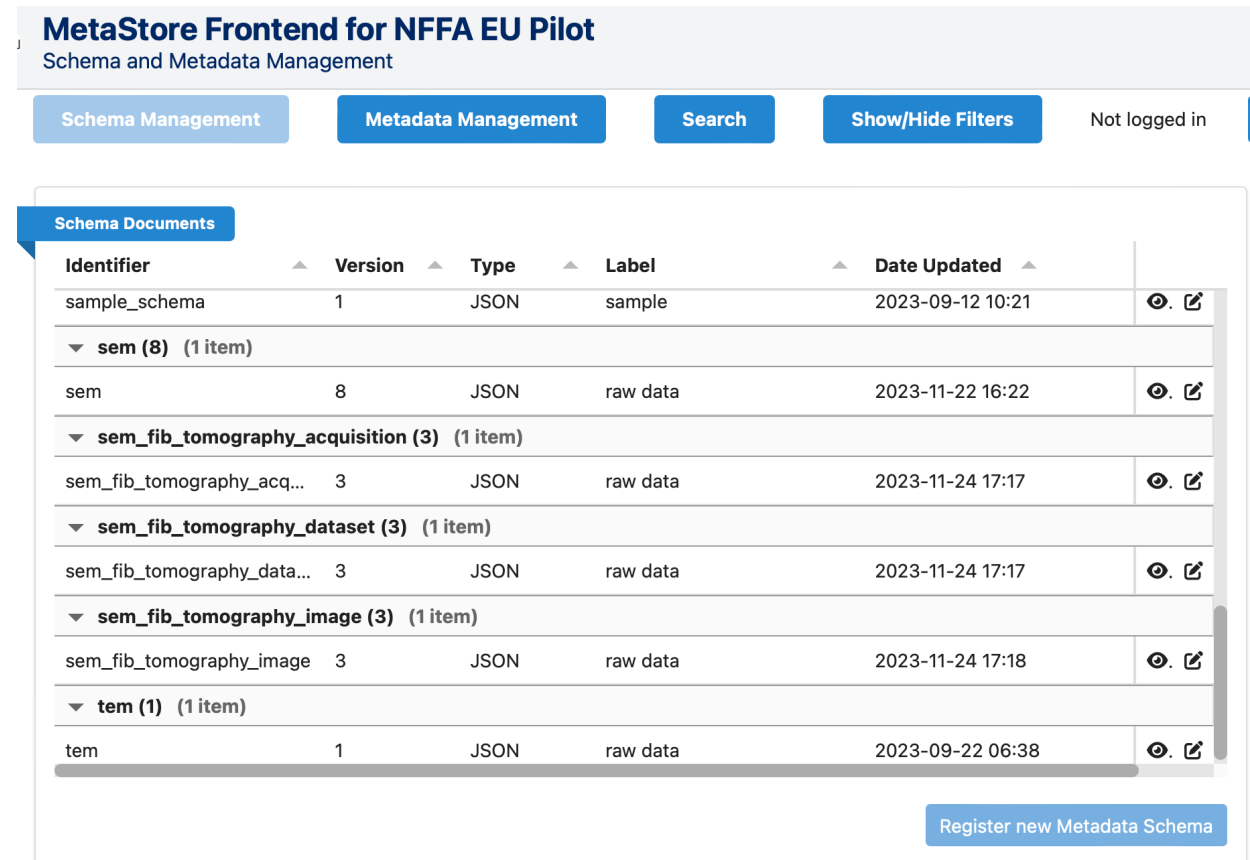
Where to store metadata?

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/>



MetaStore Frontend for NFFA EU Pilot
Schema and Metadata Management

Schema Management Metadata Management Search Show/Hide Filters Not logged in

Schema Documents

Identifier	Version	Type	Label	Date Updated	
sample_schema	1	JSON	sample	2023-09-12 10:21	👁️ ✎
▼ sem (8) (1 item)					
sem	8	JSON	raw data	2023-11-22 16:22	👁️ ✎
▼ sem_fib_tomography_acquisition (3) (1 item)					
sem_fib_tomography_acq...	3	JSON	raw data	2023-11-24 17:17	👁️ ✎
▼ sem_fib_tomography_dataset (3) (1 item)					
sem_fib_tomography_data...	3	JSON	raw data	2023-11-24 17:17	👁️ ✎
▼ sem_fib_tomography_image (3) (1 item)					
sem_fib_tomography_image	3	JSON	raw data	2023-11-24 17:18	👁️ ✎
▼ tem (1) (1 item)					
tem	1	JSON	raw data	2023-09-22 06:38	👁️ ✎

Register new Metadata Schema

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/>

The screenshot displays the 'MetaStore Frontend for NFFA EU Pilot' interface. At the top, there are navigation buttons for 'Schema Management', 'Metadata Management', 'Search', and 'Show/Hide Filters'. The user is logged in as 'rosse'. The main content area is titled 'Metadata Documents' and shows a list of documents with the following details:

Identifier	Related Resource	Schema Identifier	Date Updated
5bc69277-711c-4eb0-937d-d81c59dbbe36	https://doi.org/10.5281/zenodo.7778338	mri_schema (version=7)	2023-03-28 15:06
82100167-4424-4e98-91e9-8f886a8571dd	https://doi.org/10.5281/zenodo.6107721	mri_schema (version=7)	2023-03-28 15:05
▼ mri_schema (version=8) (1 item)			
82100167-4424-4e98-91e9-8f886a8571dd	https://b2share.eudat.eu/records/557d41bb71fe4fed9a821e0abef21d71	mri_schema (version=8)	2023-10-24 10:47
▼ mldata_basic_schema (version=2) (2 items)			

A 'Register new Metadata Document' button is located at the bottom right of the interface.

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”

```
"instrumentID": {
  "type": "string"
},
"instrumentManufacturer": {
  "type": "object",
  "properties": {
    "manufacturerName": {
      "type": "string"
    },
    "modelName": {
      "type": "string"
    }
  }
}
```

```
"instrument": {
  "instrumentID": "425590",
  "instrumentManufacturer": {
    "manufacturerName": "Bruker BioSpin MRI GmbH",
    "modelName": "Biospec 152/11"
  }
}
```

Metadata Documents

Identifier	Related Resource
3cbb28bd-62e7-425a-9024-288348f0741c	https://doi.org/10.5281/zenodo.7778338
Schema Identifier mri_schema (version=7)	
Date Updated	2023-03-28 15:06

Published February 16, 2022 | Version 2.0.0

Dataset Open

Magnetic Resonance Imaging Copper Sulfate Dataset

Nicolas Blumenröhr¹

Data collectors: Neil MacKinnon¹; Rossella Aversa²

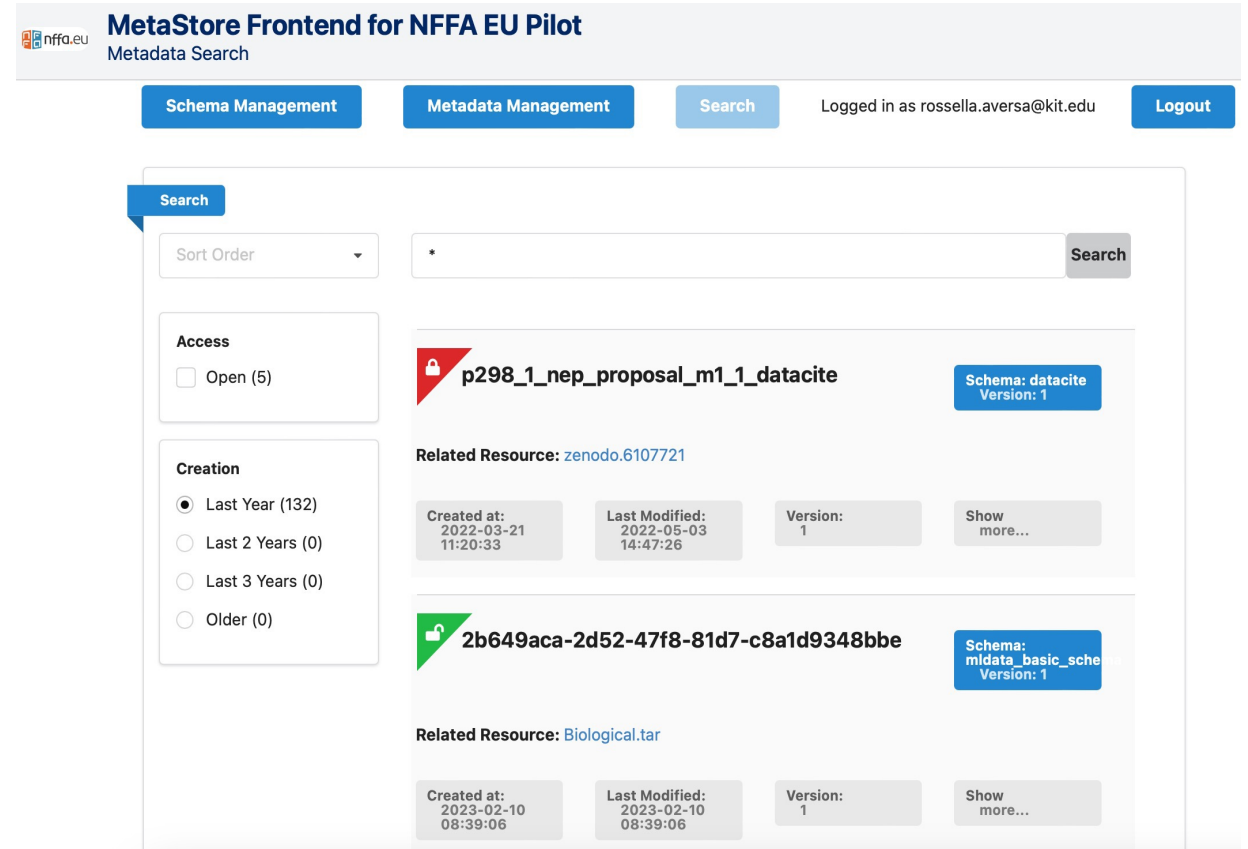
The data has been produced by the Institut für Mikrostrukturtechnik (IMT) at Karlsruher Institut für Technologie (KIT). This dataset represents the DICOM (Digital Imaging and Communications in Medicine) files, which belong to one MRI (Magnetic Resonance Imaging) study and contain a series of images that have been measured with different protocols. The samples shown by the images are tubes, which contain different concentrations of CuSO₄. The DICOM file headers have metadata tags, which embody additional information about the study and the particular series.

Files

Name	Size
series0.dcm md5:bb672509c35b4b7d94019c9065271ab6	1.8 MB

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



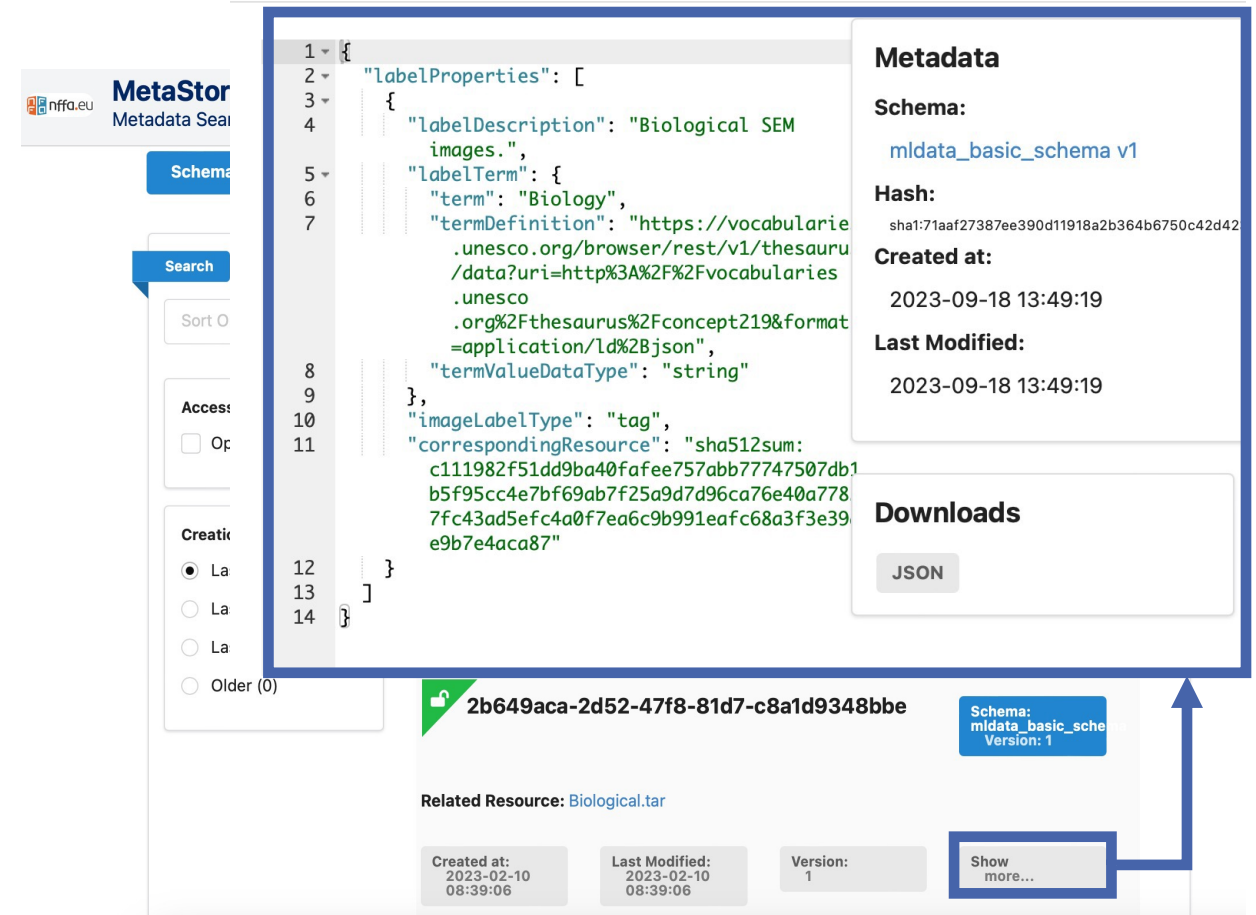
The screenshot shows the 'MetaStore Frontend for NFFA EU Pilot' interface. At the top, there is a navigation bar with 'Schema Management', 'Metadata Management', and 'Search' buttons. The user is logged in as 'rossella.aversa@kit.edu'. The main search area includes a search input field with a search button, a 'Sort Order' dropdown, and a 'Search' button. On the left, there are faceted search options for 'Access' (Open (5)) and 'Creation' (Last Year (132), Last 2 Years (0), Last 3 Years (0), Older (0)). The search results display two entries:

- Entry 1:** p298_1_nep_proposal_m1_1_datacite (Schema: datacite, Version: 1). Related Resource: zenodo.6107721. Created at: 2022-03-21 11:20:33, Last Modified: 2022-05-03 14:47:26, Version: 1.
- Entry 2:** 2b649aca-2d52-47f8-81d7-c8a1d9348bbe (Schema: mldata_basic_sche, Version: 1). Related Resource: Biological.tar. Created at: 2023-02-10 08:39:06, Last Modified: 2023-02-10 08:39:06, Version: 1.

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

Metadata Details (2b649aca-2d52-47f8-81d7-c8a1d9348bbe)



The screenshot shows the MetaStor Metadata Search interface. The main content area displays the metadata details for a resource with ID 2b649aca-2d52-47f8-81d7-c8a1d9348bbe. The metadata is shown in a JSON format, and a 'JSON' button is available for downloading the metadata. The interface also includes a search bar, a schema selector, and a 'Show more...' button.

```

1 {
2   "labelProperties": [
3     {
4       "labelDescription": "Biological SEM
5         images.",
6       "labelTerm": {
7         "term": "Biology",
8         "termDefinition": "https://vocabularie
9           .unesco.org/browser/rest/v1/thesauru
10          /data?uri=http%3A%2F%2Fvocabularies
11          .unesco
12          .org%2Fthesaurus%2Fconcept219&format
13          =application/ld%2Bjson",
14         "termValueDataType": "string"
15       },
16       "imageLabelType": "tag",
17       "correspondingResource": "sha512sum:
18         c111982f51dd9ba40fafee757abb77747507db1
19         b5f95cc4e7bf69ab7f25a9d7d96ca76e40a778
20         7fc43ad5efc4a0f7ea6c9b991eafcf68a3f3e39
21         e9b7e4aca87"
22     }
23   ]
24 }
  
```

Metadata

Schema:
mldata_basic_schema v1

Hash:
sha1:71aaf27387ee390d11918a2b364b6750c42d42

Created at:
2023-09-18 13:49:19

Last Modified:
2023-09-18 13:49:19

Downloads
JSON

2b649aca-2d52-47f8-81d7-c8a1d9348bbe

Schema:
mldata_basic_sche
Version: 1

Related Resource: Biological.tar

Created at: 2023-02-10 08:39:06

Last Modified: 2023-02-10 08:39:06

Version: 1

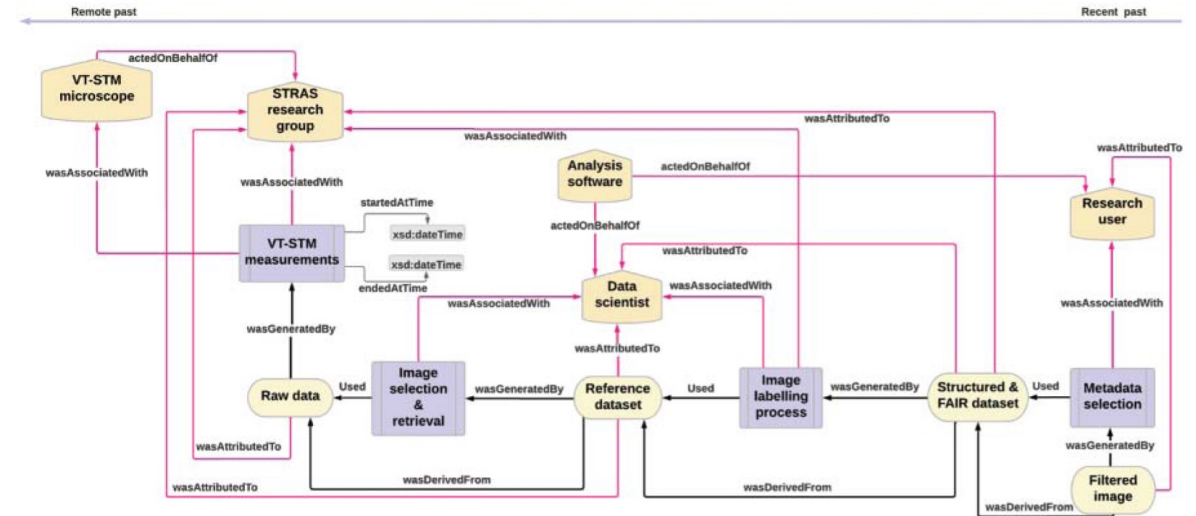
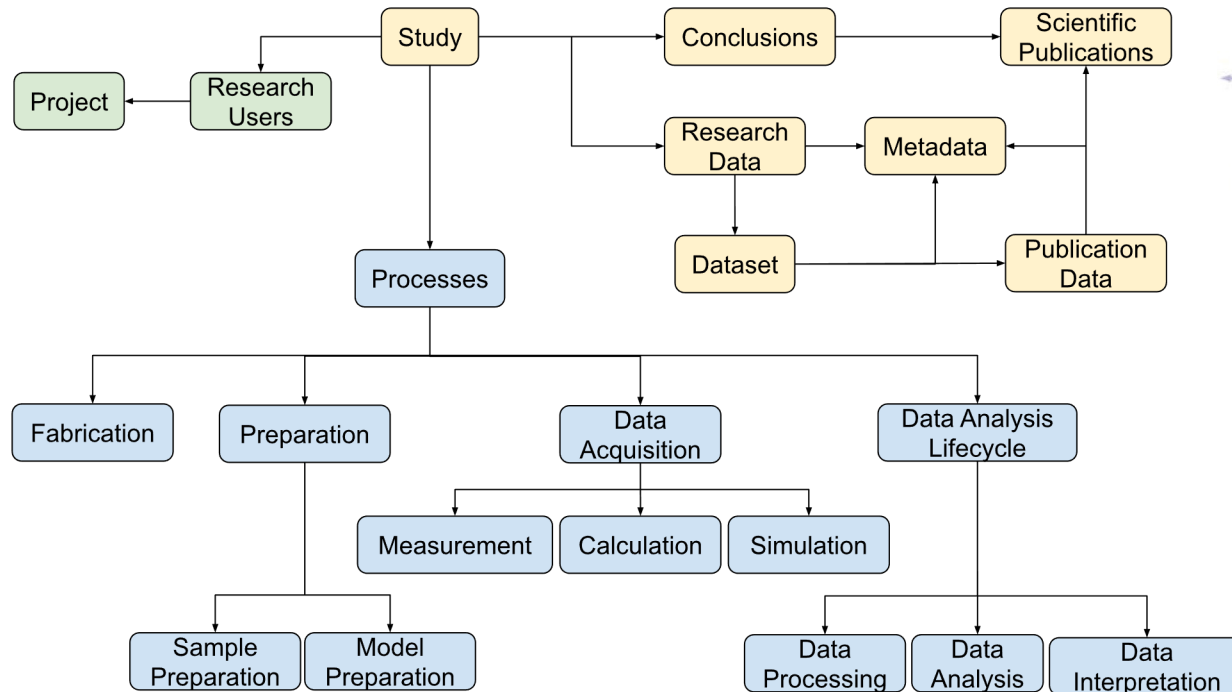
Show more...

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](https://doi.org/10.1162/dint_a_00164)

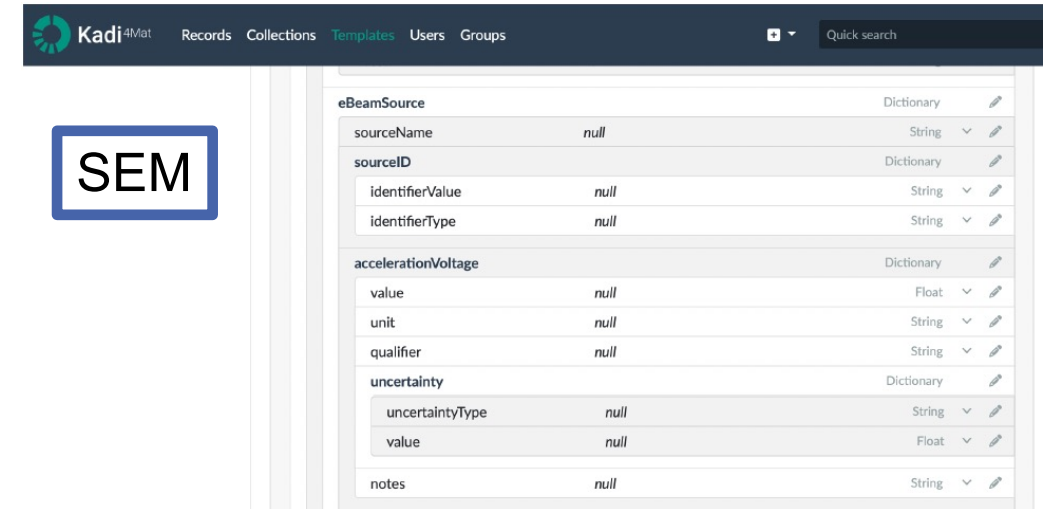
How to achieve it in practice?

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

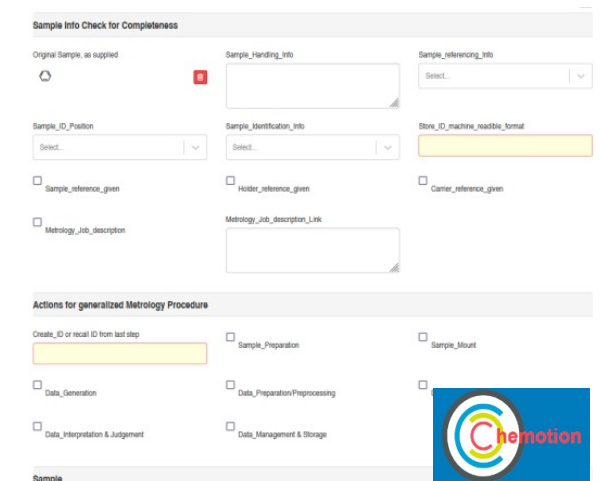


```

"description": "General workflow to control a metrology process."
"extras": [
  {
    "key": "1) Sample Info Completeness Check",
    "type": "dict",
    "value": [
      {
        "key": "Outer Dimensions",
        "type": "list",
        ..
      }
    ]
  }
]

```

Metrology




How to achieve it in practice?

Metadata editor

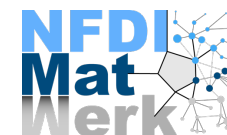
- Edit metadata documents according to schemas
- Register metadata documents to MetaStore via the GUI
- Local application (exposes API for instruments/ELNs)



The screenshot shows the Metadata editor interface. At the top, there is a logo consisting of several interlocking shapes in orange and blue. Below the logo, the text "Metadata editor" is displayed. The main interface features three dropdown menus: "Label" with the value "raw data", "Schema ID" with the value "sem", and "Version" with the value "8". Below these menus are three blue buttons: "LOAD SCHEMA", "LOAD JSON DOCUMENT", and "MERGE JSON DOCUMENT". A lock icon is positioned below the buttons, and the word "Entry" is centered below the lock. At the bottom, there is a form with an "Entry ID" dropdown menu. Below the dropdown, the "Title*" field contains the text "sem_test_img". The "Start Time" field contains "2023-06-21 15:30" and the "End Time*" field contains "2023-06-21 16:00". Both time fields have a calendar icon to their right.

<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

NFDI-MatWerk Mapping Service
Automatic extraction and mapping of metadata.

Mapping Component

Choose a suitable mapping from available options

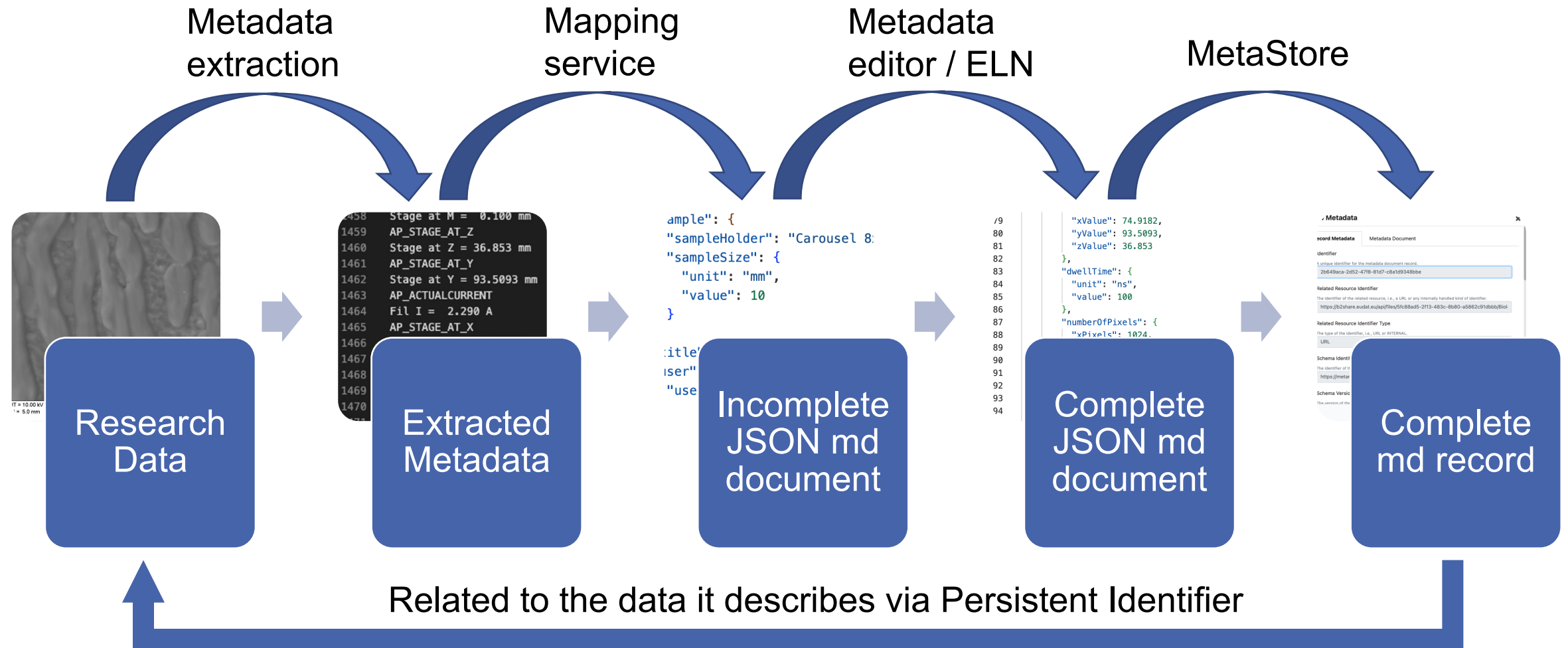
SEM/FIB Tomography Acquisition to TXT	SEM/FIB Tomography Acquisition to JSON	SEM Zip to txt	SEM to JSON
Creates a summary of all metadata extracted from images in a comma delimited txt file. LU: 23.08.2023	Extracts metadata from a SEM/FIB Tomography Acquisition zip file and maps it to the appropriate schema to create a JSON metadata document. LU: 24.08.2023	Takes a zip folder of arbitrary SEM tiff images and returns a comma separated txt file with a summary of all of the embedded metadata. LU: 05.09.2023	This plugin is able to handle a variety of SEM images and processes them using the Hyperspy library. A resulting metadata document in JSON format is then
Select	Select	Select	Select

Drag & Drop your files or [Browse](#)

Map document

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

General Metadata Management Workflow



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.