

Metadata Management in Scientific Research: an overview

Rossella Aversa (KIT-SCC)

iEntrance Advanced School, 22.02.2024



Outline

1.

Introduction

Motivation and recap of the FAIR principles

2.

Effective Metadata Management

Best practices and tools to implement them

3.

Metadata Management in practice

Top 10 management steps in real projects

4.

Conclusions

Summary and main takeaways

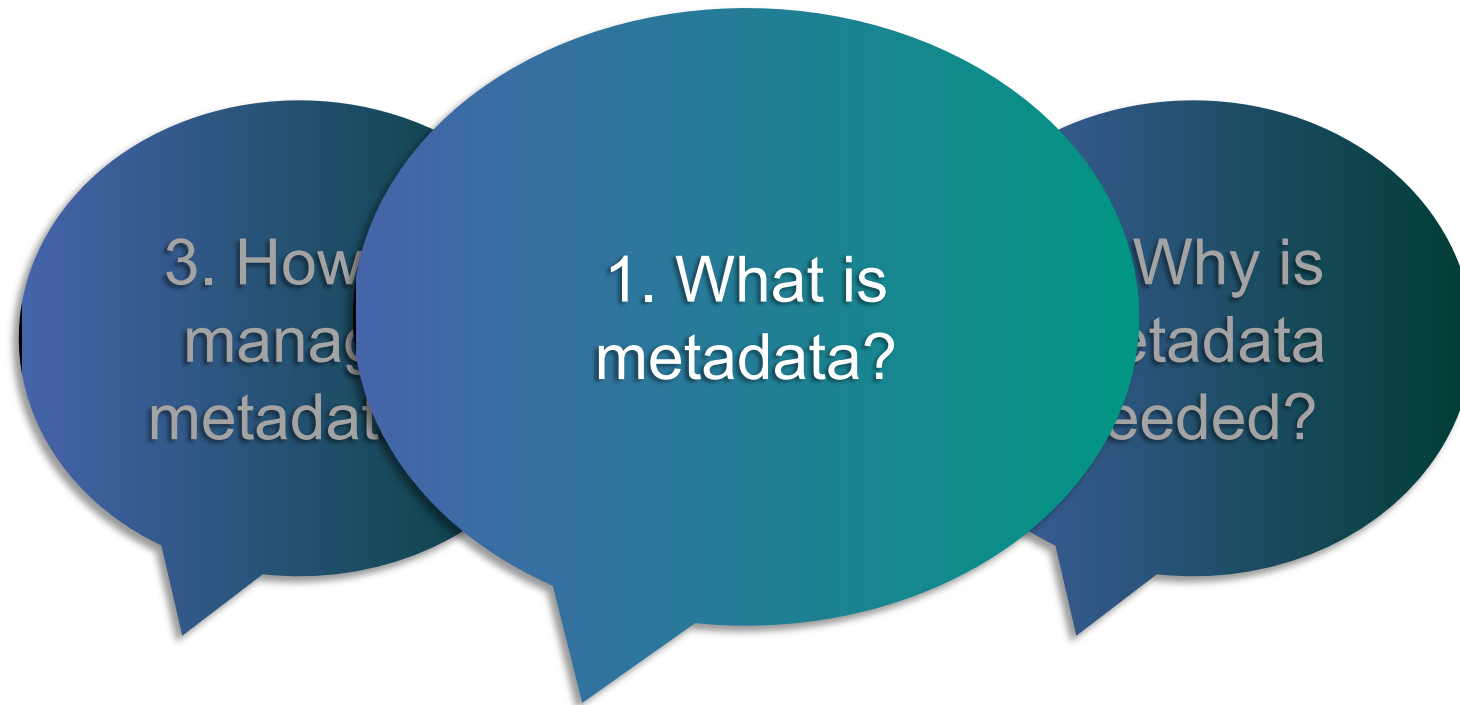
1. Introduction

Motivation



Powered by Bing Image Creator

Metadata



Data describing other data

Metadata



To add context and meaning to data

Metadata

2. Why
metadata
needed?

3. How to
manage
metadata?

What is
metadata?

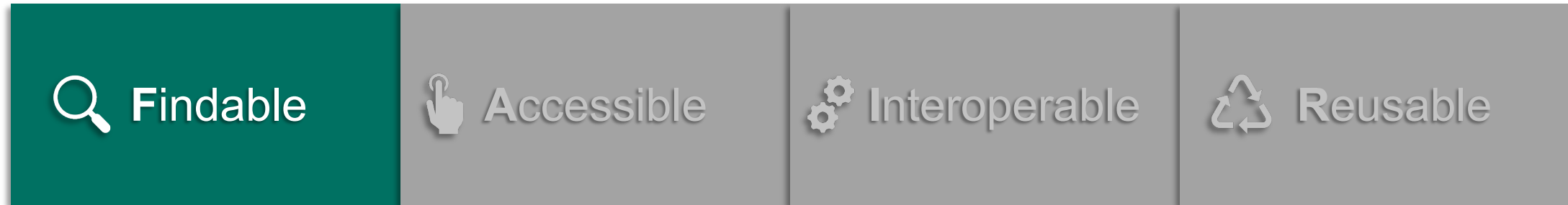
We will see it throughout the next slides...

The FAIR Guiding Principles



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

The FAIR Guiding Principles



(Meta)data should be easy to find for both humans and computers



Globally unique persistent identifiers (PID)

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

The FAIR Guiding Principles



It should be known how (meta)data can be accessed



(Meta)data repositories, authorization & authentication

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

The FAIR Guiding Principles



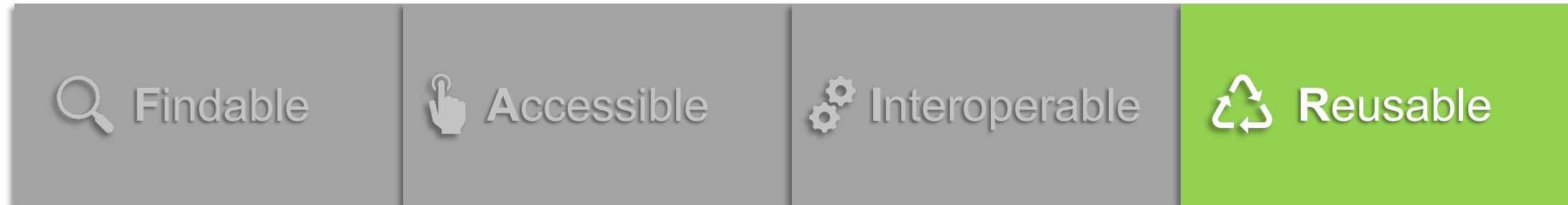
Data should be exchanged and interpreted by humans and computers



Structured metadata (schemas, vocabularies)

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

The FAIR Guiding Principles



It should be clear how data can be reused and/or replicated

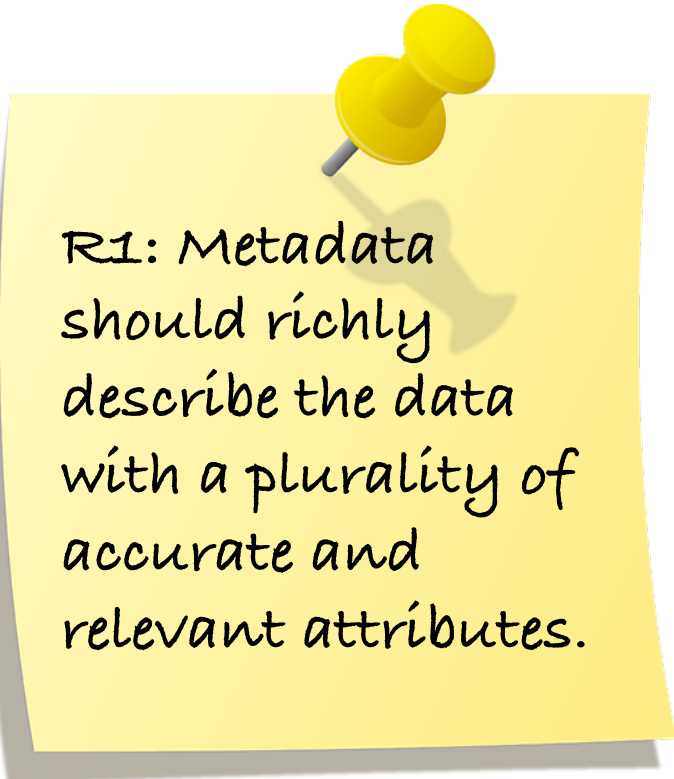


Licences, rich (provenance) metadata

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

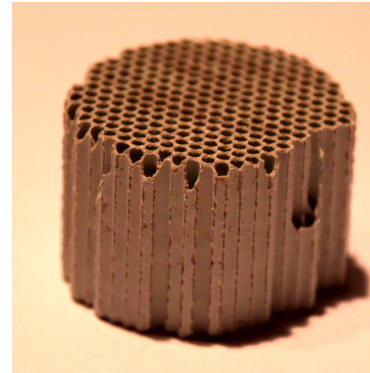
2. Effective Metadata Management

What to describe?



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

Sample



Sample Holder



Instrument

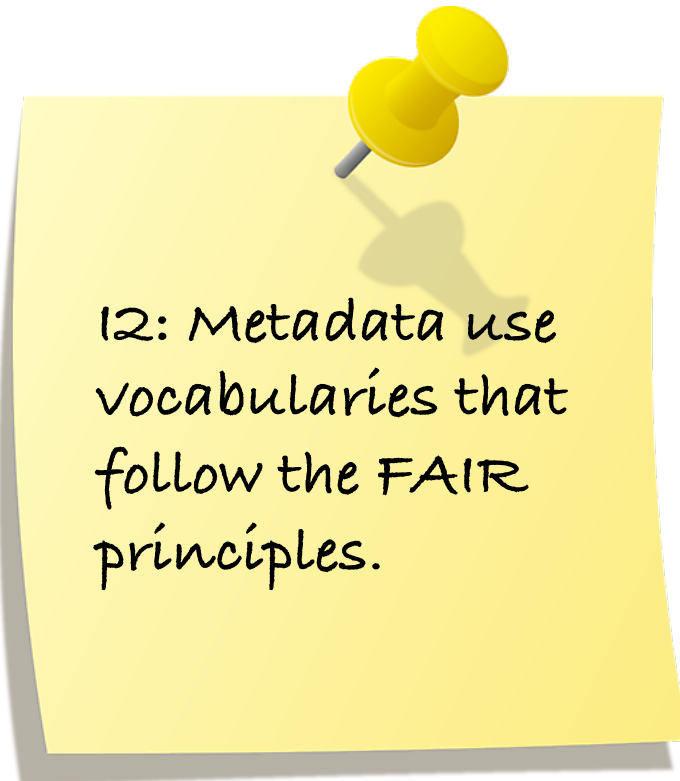


Research data



Images: courtesy of R. Thelen and M. Mail. Powered by Bing Image Creator

How to describe data?



Minimal requirements:

- The vocabulary and its terms have globally unique PIDs
- The vocabulary and its terms are documented
- The documentation is findable and accessible by users

Skosmos Vocabularies About Feedback Help | Interface language: English ▾

UNESCO Thesaurus Content language: English ▾ Search

Alphabetical **Hierarchy** Groups

A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z

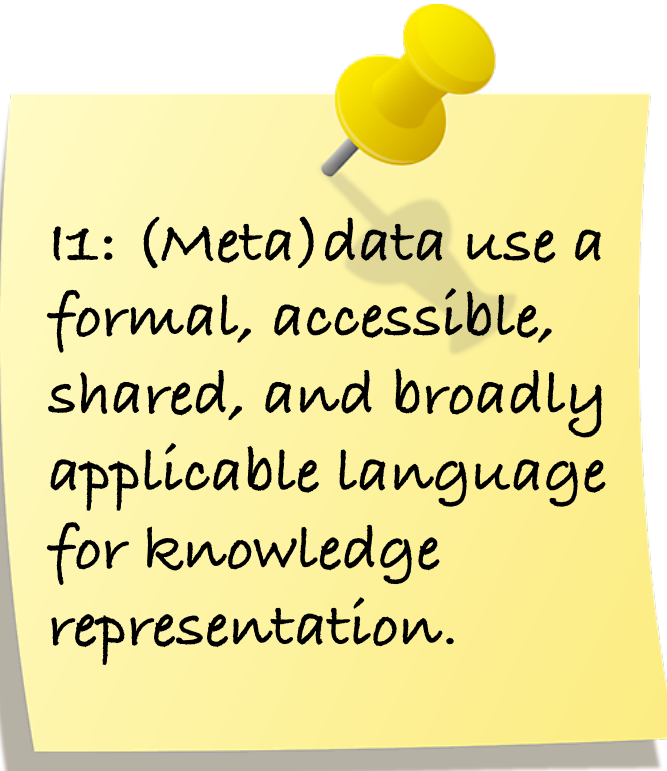
- Abadi → Awadhi
- Abandoned children
- Abbreviations
- Ability
- Ability grouping
- Abohi → Awadhi
- Aboriginals → Indigenous peoples
- Abortion
- Absenteeism → Leave
- Abstract journals → Abstracts
- Abstract reasoning → Reasoning
- Abstracting
- Abstracting and indexing services → Bibliographic services
- Abstracts
- Abuse of human rights → Human rights violations
- Abuse of power → Oppression
- Academic achievement
- Academic admission → Admission requirements
- Academic buildings
- Academic degrees → Degrees
- Academic equipment → Educational equipment
- Academic facilities → Educational facilities
- Academic failure
- Academic fraud
- Academic freedom
- Academic grouping → Educational grouping
- Academic laboratories → University laboratories
- Academic libraries
- Academic management → Educational management
- Academic misconduct → Academic fraud
- Academic performance → Academic achievement

Vocabulary information

TITLE	UNESCO Thesaurus
CREATOR	UNESCO
LICENSE	http://creativecommons.org/licenses/by-sa/3.0/igo/
RIGHTS	CC-BY-SA
LANGUAGE	http://id.loc.gov/vocabulary/iso639-2/eng http://id.loc.gov/vocabulary/iso639-2/fre http://id.loc.gov/vocabulary/iso639-2/rus http://id.loc.gov/vocabulary/iso639-2/spa
SOURCE	http://databases.unesco.org/thesaurus/ http://www2.ulcc.ac.uk/unesco/
CREATED	Saturday, January 1, 1977 00:00:00
CONFORMS TO	ISO 25964
IDENTIFIER	http://skos.um.es/unescothes
TYPE	thesaurus
TYPE	http://www.w3.org/2004/02/skos/core#ConceptScheme
URI	http://skos.um.es/unescothes/CS000

<https://skosmos.org>

How to represent metadata?



11: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.



Resource Description Framework: metadata model to represent interconnected data.

<https://www.w3.org/RDF/>



Simple Knowledge Organization System: standard to represent knowledge organization systems using RDF


<https://www.w3.org/2004/02/skos/>



Web Ontology Language: computational logic-based language to represent complex knowledge.

<https://www.w3.org/OWL/>

How to structure metadata?



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

General purpose

 **DublinCore** <http://dublincore.org/schemas/>

 **DataCite** <http://schema.datacite.org>

 **Schema.org** <https://schema.org>

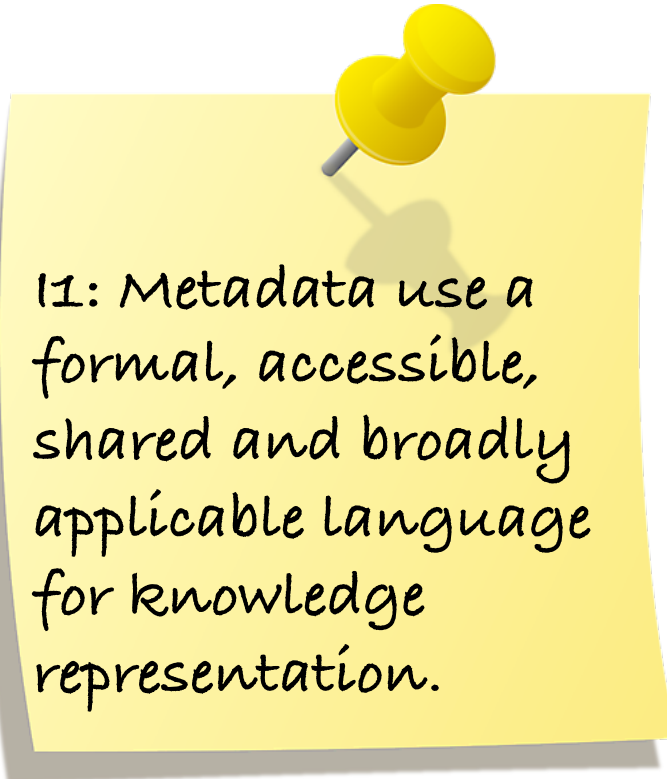
Neutron, x-ray, muon

 **NeXus** <http://www.nexusformat.org>

Crystallography

 **CIF** <https://www.iucr.org/resources/cif>

How to represent structured metadata?



11: Metadata use a formal, accessible, shared and broadly applicable language for knowledge representation.


 XML (eXtensible Markup Language)

 JSON (JavaScript Object Notation)

 JSON Schema <https://json-schema.org>

 JSON-LD (JSON for Linked Data)

How to publish (meta)data?



F1: (Meta)data are assigned globally unique and persistent identifiers.

Cite this article

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).

<https://doi.org/10.1038/sdata.2016.18>

Details

DOI

DOI [10.5281/zenodo.7778338](https://doi.org/10.5281/zenodo.7778338)

Resource type

Dataset

Publisher

Zenodo

Languages

English



ORCID
Connecting research and researchers

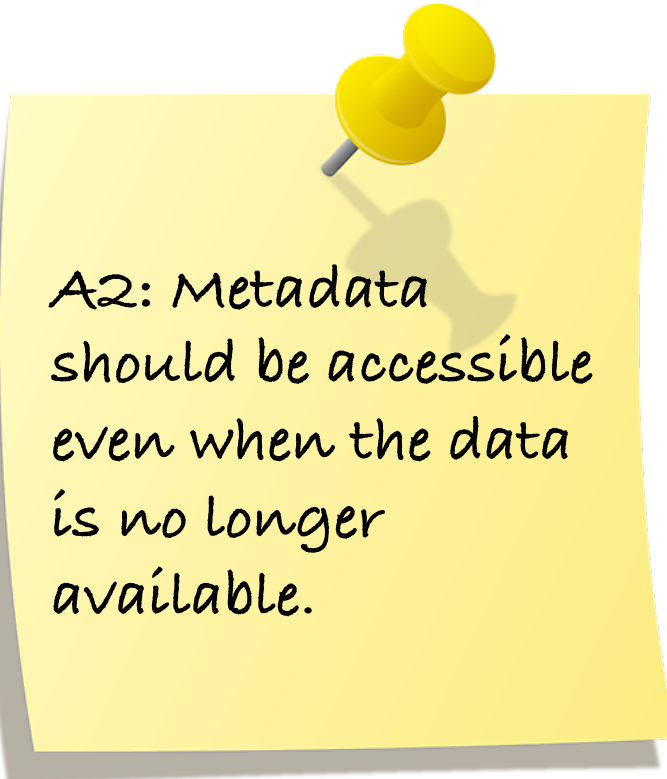


<https://orcid.org/>

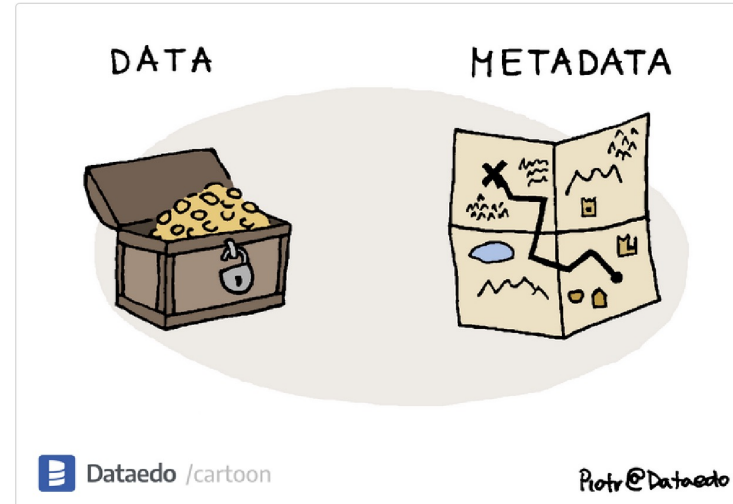
0000-0003-2534-0063

[Preview public record](#)

How to publish metadata?

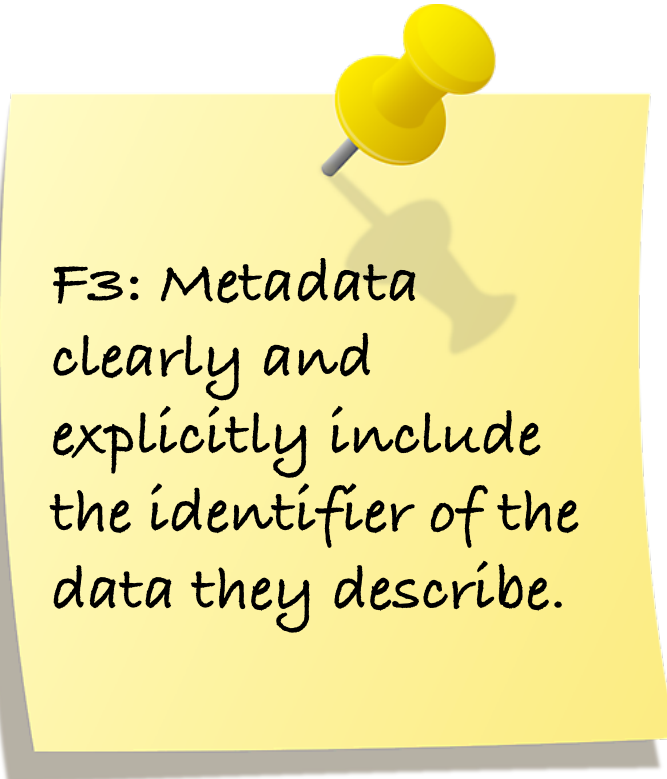


A2: Metadata should be accessible even when the data is no longer available.



Metadata
Repositories!

How to find data from metadata?



F3: Metadata clearly and explicitly include the identifier of the data they describe.



Details

DOI

DOI [10.5281/zenodo.7778338](https://doi.org/10.5281/zenodo.7778338)

Resource type

Dataset

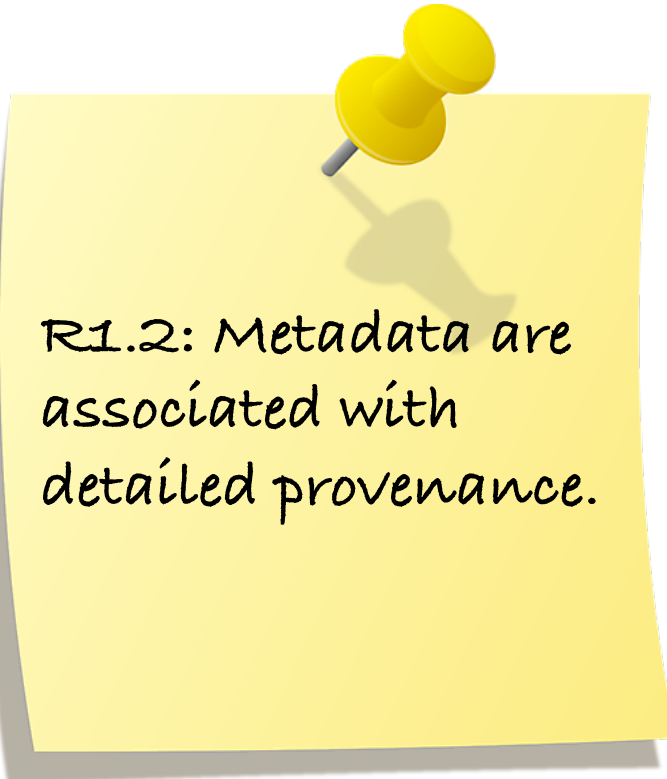
Publisher

Zenodo

Languages

English

How to find data from metadata?



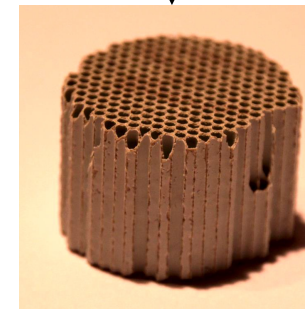
R1.2: Metadata are associated with detailed provenance.



Data acquired from measurement



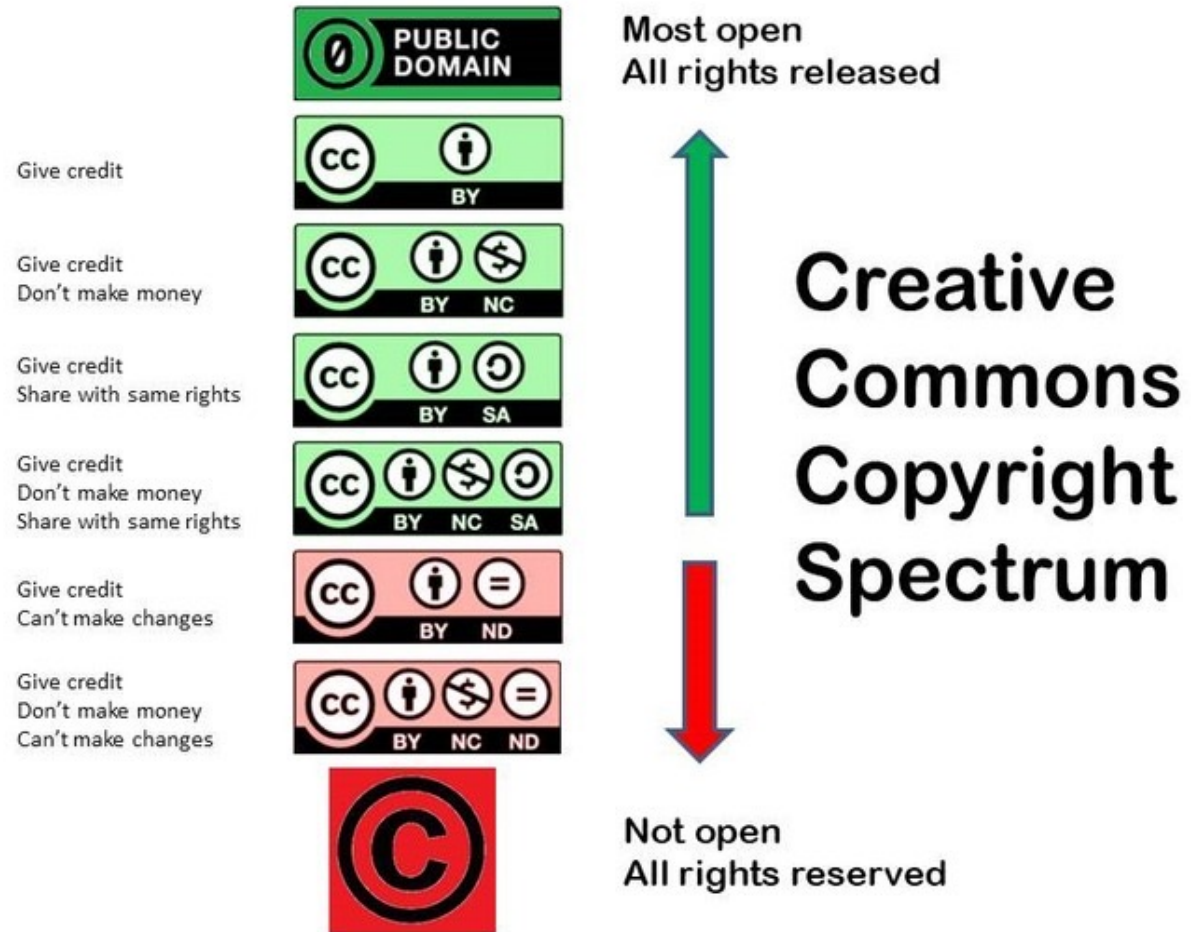
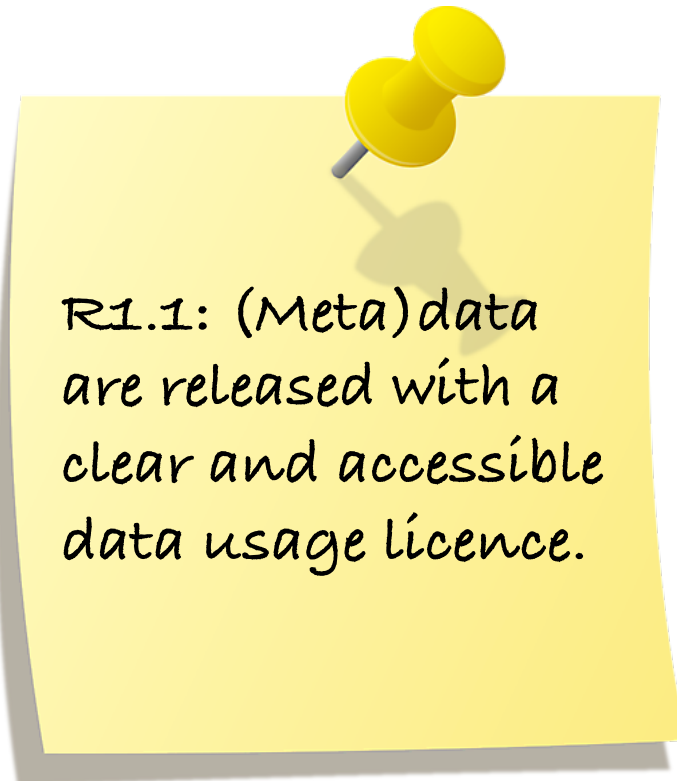
Measurement performed on sample



Sample placed on holder

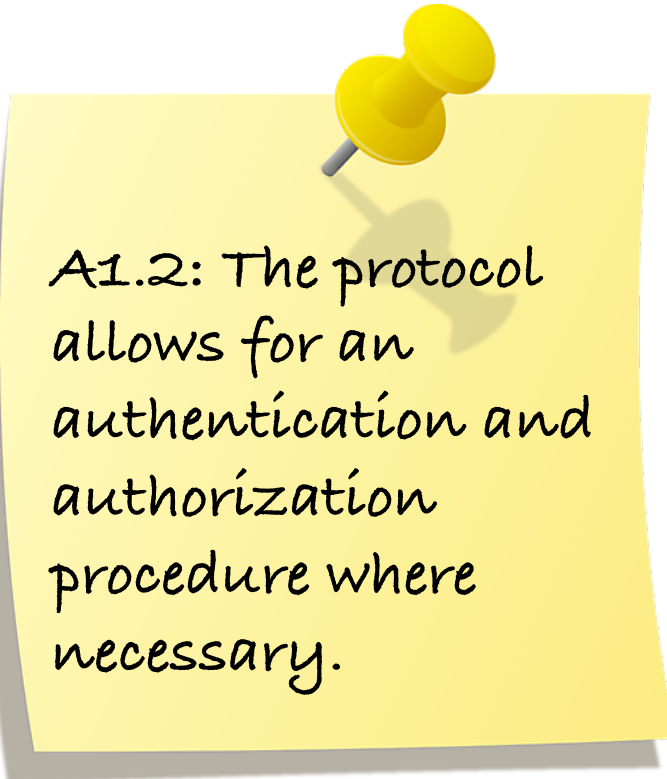


How to reuse the data?



<https://creativecommons.org/share-your-work/cclicenses/>

Should FAIR data be open?

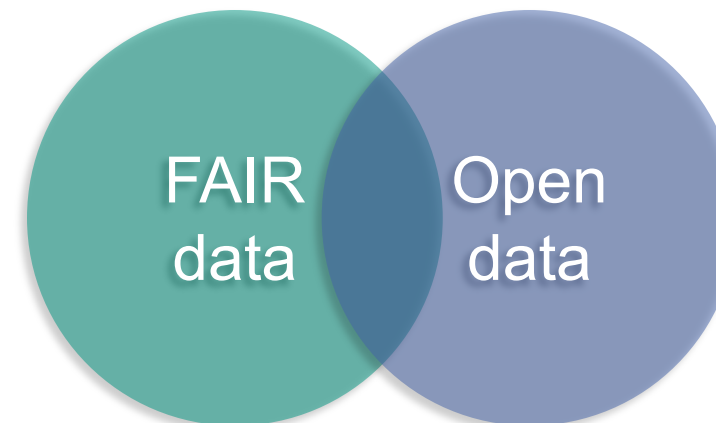


A1.2: The protocol allows for an authentication and authorization procedure where necessary.

Open data: “can be freely used, modified, and shared by anyone for any purpose”

<https://opendefinition.org>

FAIR data: “as open as possible, as closed as necessary”



FAIR or open?

My data is copyright protected	FAIR	
My dataset can be used only by a specific group of scientists	FAIR	
An image is shared on a public website	Open	
A dataset is published on Zenodo with an open licence	FAIR	Open
A data file is on my Dropbox	None	

3. Metadata Management in practice

The projects



Nanoscience Foundries and Fine Analysis – Europe Pilot (NEP)
<https://nffa.eu>

Access to nanoscience research infrastructure
Synthesis, growth of nanostructures, fine analysis, theory and simulation

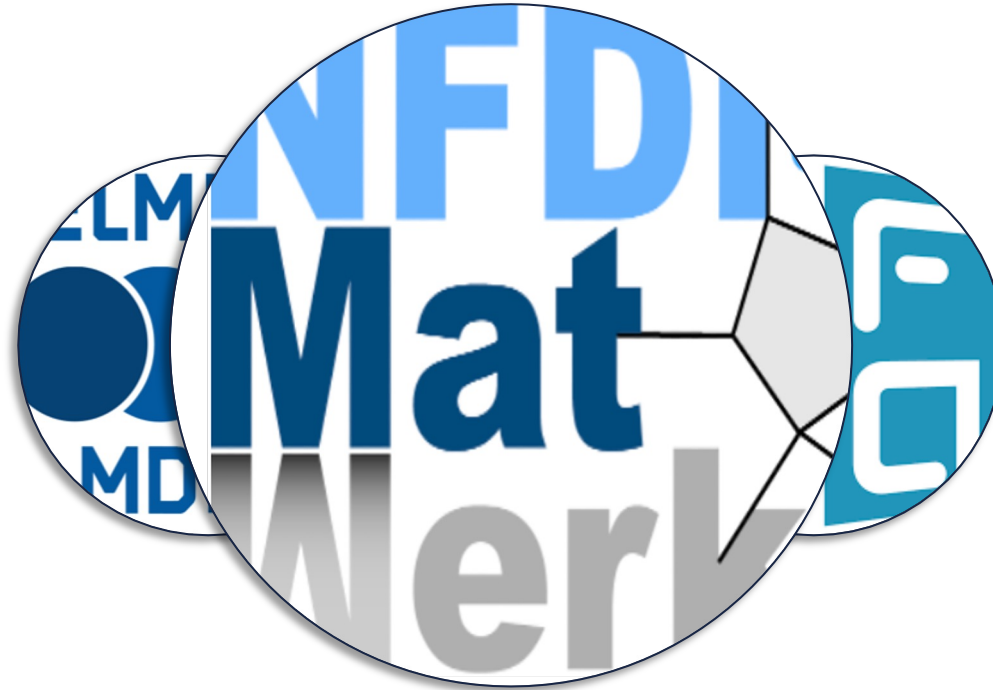
The projects



Joint-Lab Model and Data driven
Materials Characterization (MDMC)
of the Helmholtz Association
<https://jl-mdmc-helmholtz.de>

Platform for multiscale and
multidimensional characterization,
analytics and simulation methods

The projects



National Research Data Infrastructure
for Materials Science and Engineering
(NFDI-MatWerk)
<https://nfdi-matwerk.de>

Infrastructure for the digital
representation of materials and
their relevant process

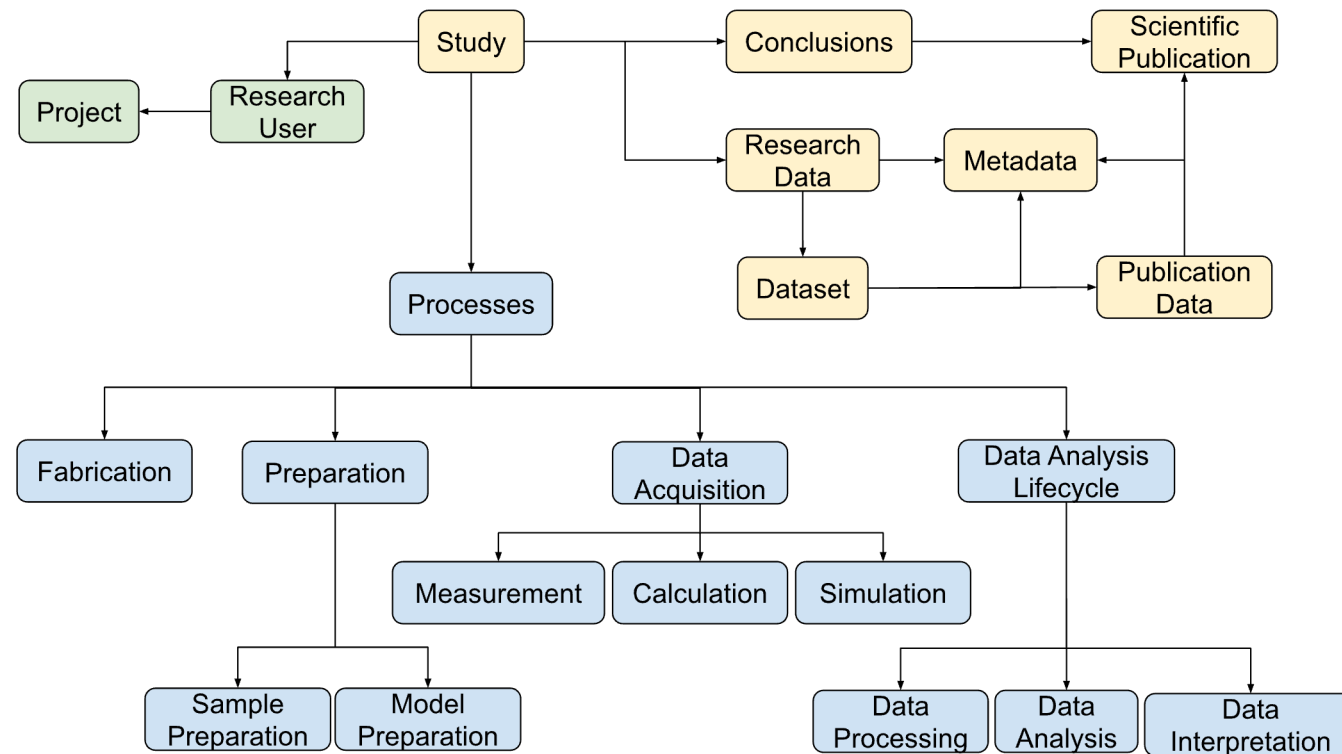
The projects



Common aims:

- Implement (meta)data management practices following the FAIR principles
- Develop tools and infrastructure solutions guided by community requirements
- Agree on common descriptions
- Collaborate on interoperable results

1. Definition of terms

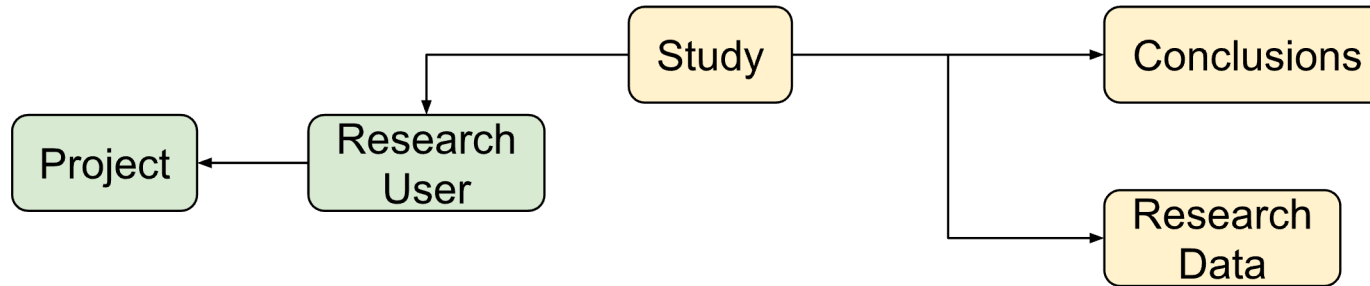


Glossary of Terms

- High-level description of experimental and computational workflows
- Framed in the management infrastructure of the projects
- Allows to track the provenance information
- Adopts existing terms

MDMC-NEP Glossary of Terms. DOI: [10.5281/zenodo.10663833](https://doi.org/10.5281/zenodo.10663833)

1. Definition of terms



Research User

Person, usually member of a **Project**, who conducts any part of the **Study**, in order to collect and/or analyse **Research Data** or is interested in reusing **Research Data** by a third party (e.g., **Reference Data**) with the final aim to extract insights that support the answer to some specific research question (i.e., **Conclusions**). **Research Users** may be assigned with a role (data curator, instrument scientist, team leader, team member).

2. Terms in a Vocabulary Service

PREFERRED TERM	User Role
<hr style="border-top: 1px dashed #ccc;"/>	
NARROWER CONCEPTS	Data Curator Instrument Scientist Team Leader Team Member
URI	http://matwerk.datamanager.kit.edu:8001/DemoTerms-1/en/page/userrole
DOWNLOAD THIS CONCEPT:	RDF/XML TURTLE JSON-LD

Research User

User Name

User Role

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

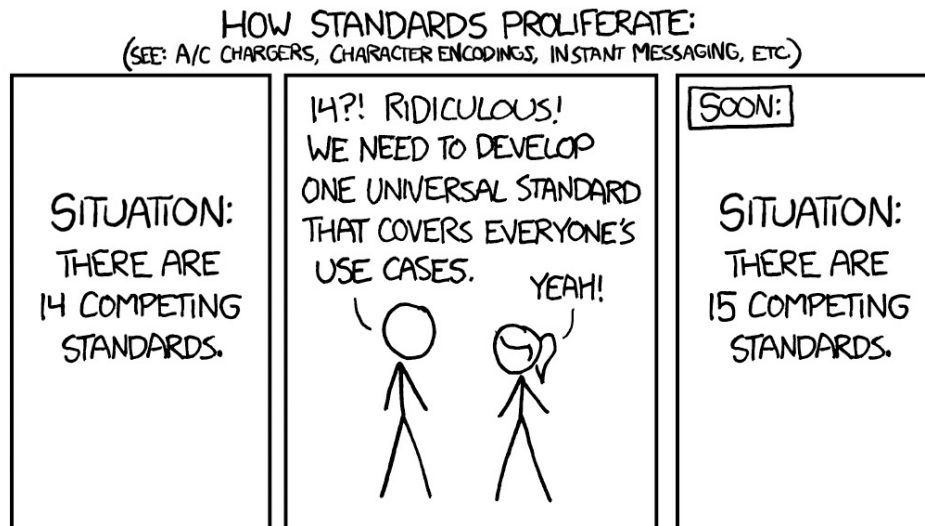
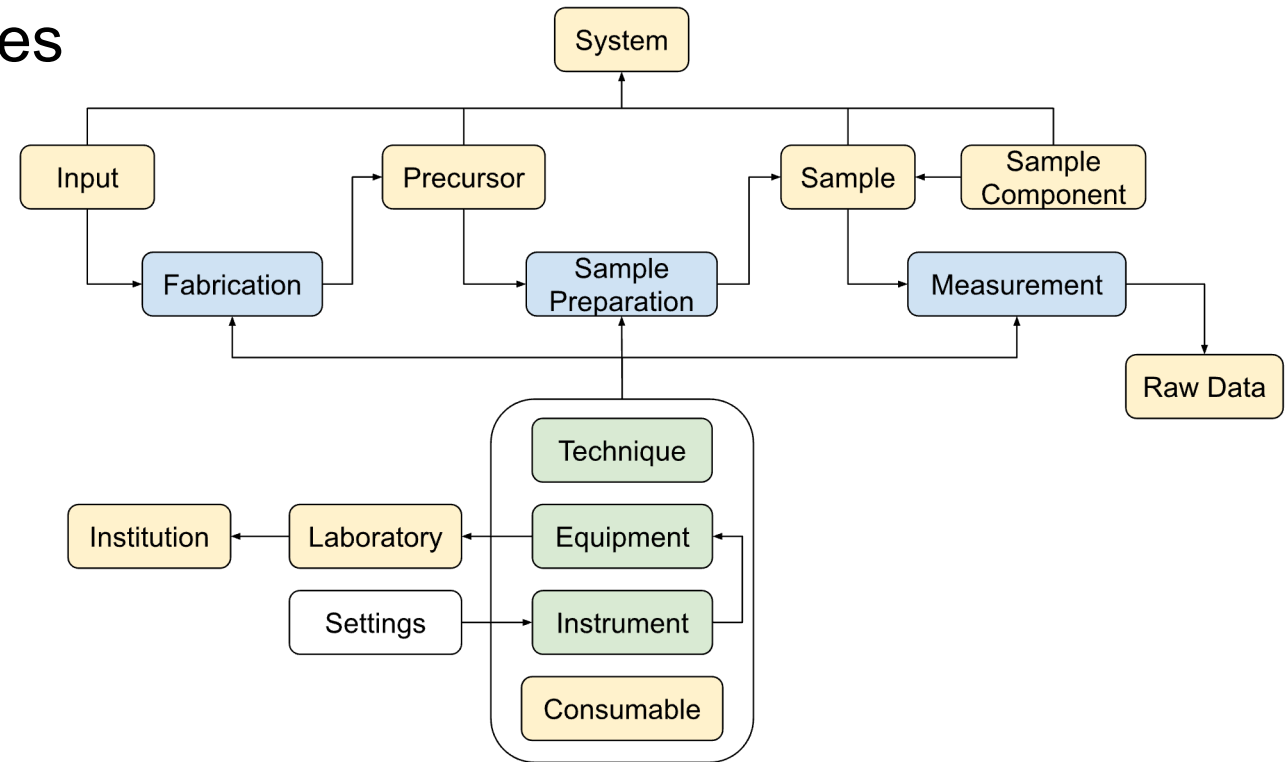
EVOKS Vocabulary Service: Collaborative online vocabulary editor

- RDF model + SKOS model
- Persistent identifier to each term
- Can be resolved in interfaces, websites, automatic processes...
- Centrally maintained
- Public read-only Skosmos instance



3. Metadata schemas

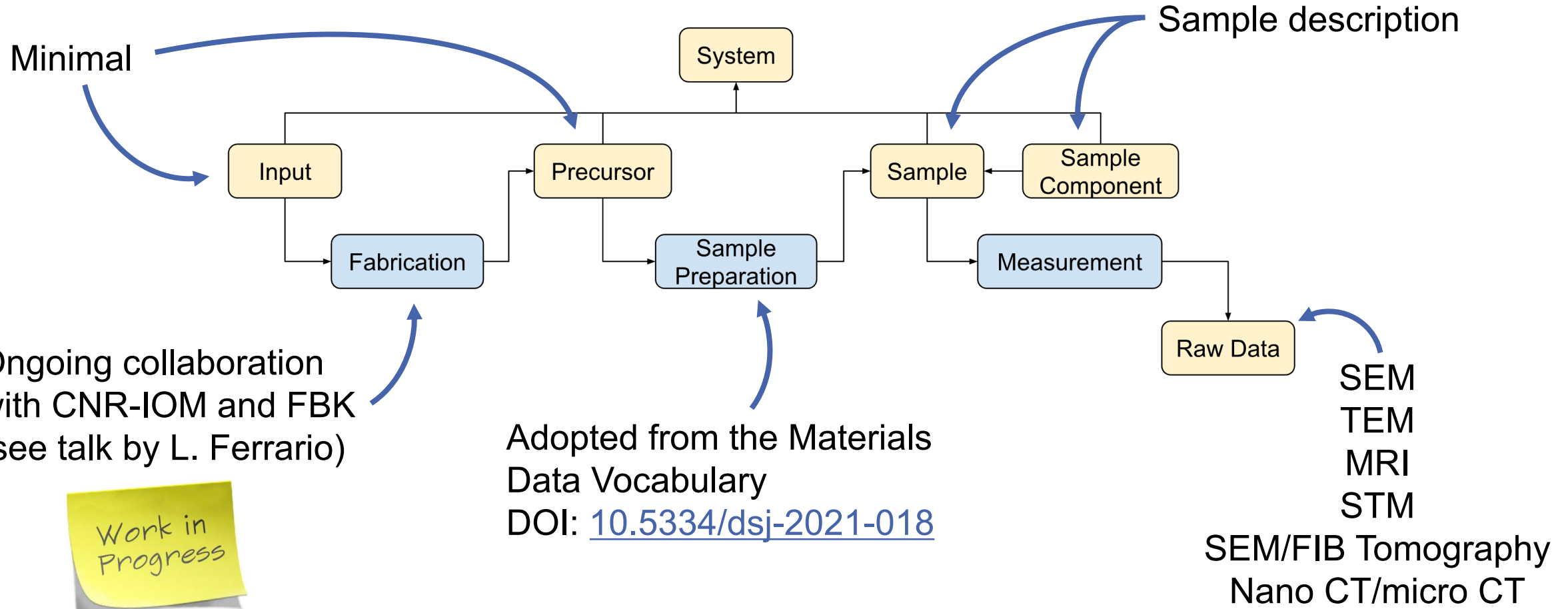
- Describe inputs/outputs of processes
- JSON schema
- Adopt existing solutions
- Avoid proliferation of schemas



<https://xkcd.com/927/>

MDMC-NEP Glossary of Terms. DOI: [10.5281/zenodo.10663833](https://doi.org/10.5281/zenodo.10663833)

3. Metadata schemas

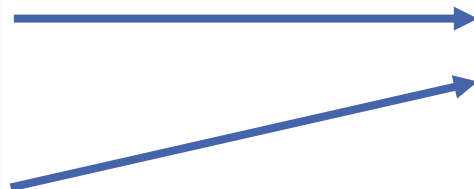


4. Metadata schemas and documents

Metadata Schema: outline of the overall structure of the metadata (elements, value types, rules, ...)

Metadata Document: structured information about a data resource

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

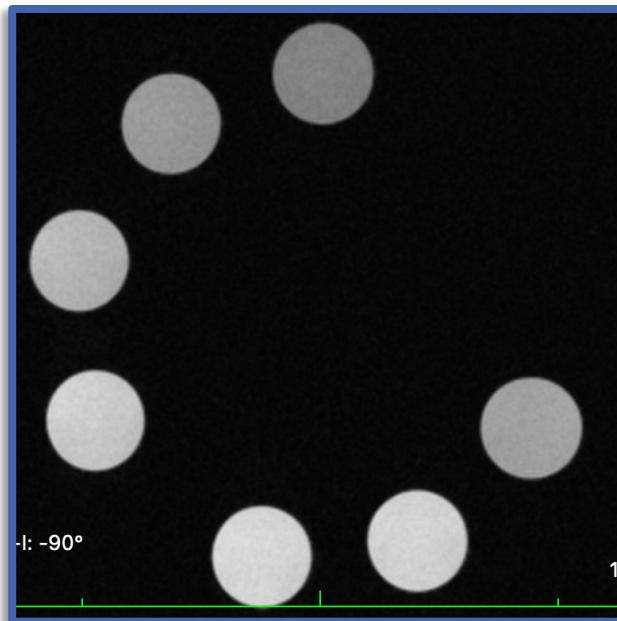


```
ent": {  
  "instrumentID": "425590",  
  "instrumentManufacturer": {  
    "manufacturerName": "Bruker B  
    "modelName": "Biospec 152/11"
```

MRI schema, DOI: [10.5445/IR/1000159552](https://doi.org/10.5445/IR/1000159552)

From data to metadata

Raw Data



Metadata Schema

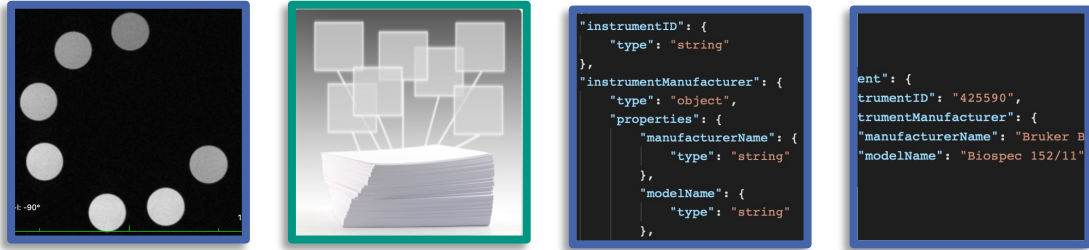
```
"instrumentID": {  
  "type": "string"  
},  
"instrumentManufacturer": {  
  "type": "object",  
  "properties": {  
    "manufacturerName": {  
      "type": "string"  
    },  
    "modelName": {  
      "type": "string"  
    }  
  }  
}
```

Metadata Document

```
ent": {  
  trumentID": "425590",  
  trumentManufacturer": {  
    "manufacturerName": "Bruker B  
    "modelName": "Biospec 152/11"
```

Image from Magnetic Resonance Imaging Copper Sulfate Dataset. DOI: [10.5281/zenodo.6107720](https://doi.org/10.5281/zenodo.6107720)

5. Mapping service



- Online service
- Input: data file(s)
- Extract unstructured metadata
- Map them to the metadata schema
- Output: structured metadata

Mapping Service UI

Extract metadata and map it to json

The Mapping Service is a tool designed to extract metadata from different kinds of data produced by instruments, and map this metadata to published metadata schemas. [Show More](#)

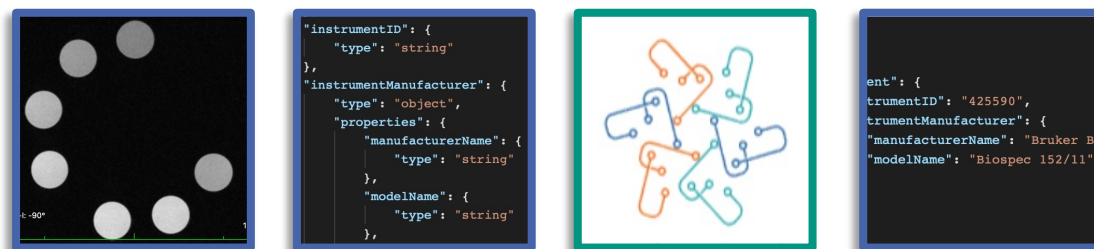
Choose a suitable mapping from available options

SEM to TXT	MRI to JSON	SEM to JSON
JSON based tool extracts from machine generated microscopy images in the format and generates a TXT file containing a summary of the metadata. Last edited: 10.12.2022	Takes a single .dcm or zipped directory of .dcm files and maps to the MRI schema returning a JSON metadata document. LU: 01.02.2024	This plugin is able to handle a variety of SEM images and processes them using the Hype library. A resulting metadata document in JSON format is then created. LU: 06.02.2024
Select	Select	Select

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

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

6. Metadata Editor



- Local service connected to the metadata repository
- Load schema from registered ones
- Load existing metadata documents
- Manually edit metadata documents
- Download metadata documents
- Register metadata documents
- Create the provenance file



Metadata editor

Label: raw data | Schema ID: mri_schema | Version: 7

LOAD SCHEMA

LOAD JSON DOCUMENT | MERGE JSON DOCUMENT

Study

Study ID*
1.2.840.10008.5.1.4.1.1.4.1

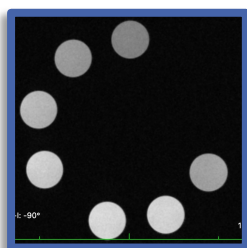
Study Title
7 samplesCuSO4 0 - 100 mM

Study Date Time*
2021-10-15 15:48

Program
['Acquisition PV-360.3.1', 'ParaVision 360.3.1']

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

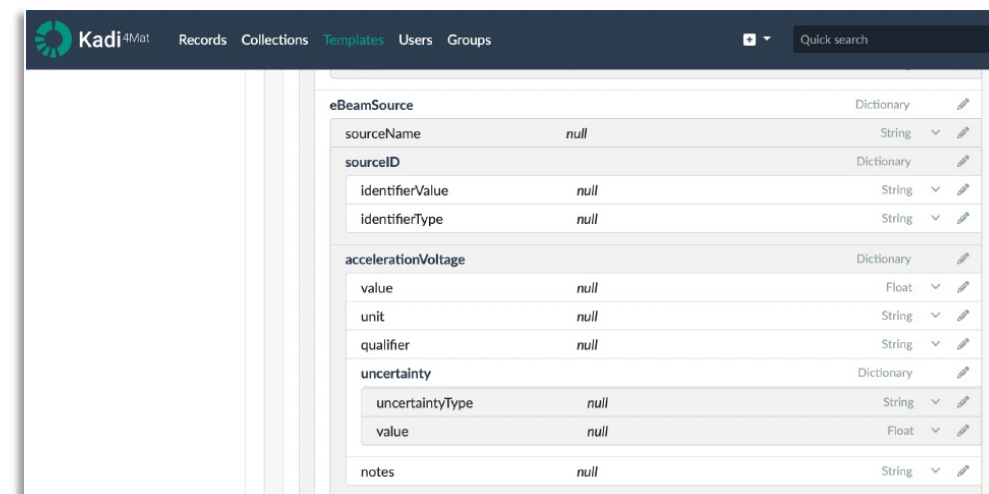
7. ELN and LIMS



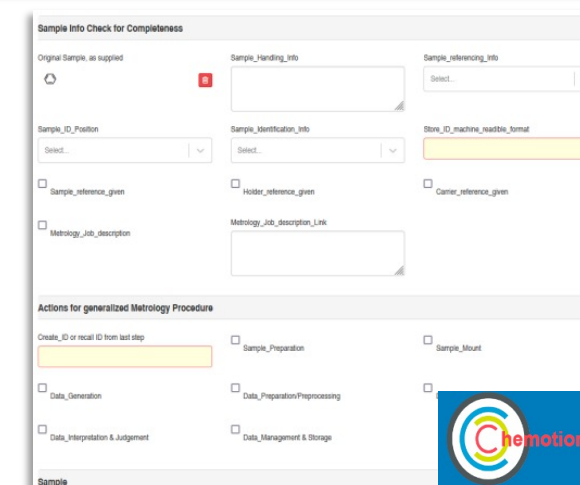
```
"instrumentID": {  
  "type": "string"  
},  
"instrumentManufacturer": {  
  "type": "object",  
  "properties": {  
    "manufacturerName": {  
      "type": "string"  
    },  
    "modelName": {  
      "type": "string"  
    }  
  }  
}
```



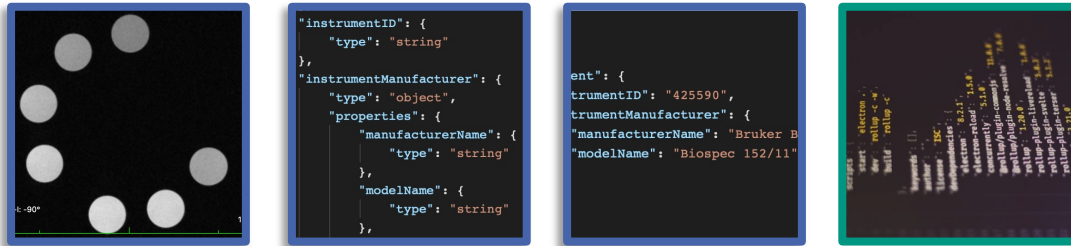
```
ent": {  
  "instrumentID": "425590",  
  "instrumentManufacturer": {  
    "manufacturerName": "Bruker B",  
    "modelName": "Biospec 152/11"  
  }  
}
```



- Electronic Lab Notebooks
- Lab Information Management Systems
- Metadata schemas as templates
- Ongoing collaborations:
 - KIT
 - CNR-IOM
 - FBK (see talk L. Ferrario)



8. MetaStore



- Metadata repository
- Register/find metadata schemas
- Register/find metadata documents
- Validate metadata documents
- Versioning
- Access control management
- User authentication

MetaStore Frontend for NFFA EU Pilot
Schema and Metadata Management

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

Schema Documents

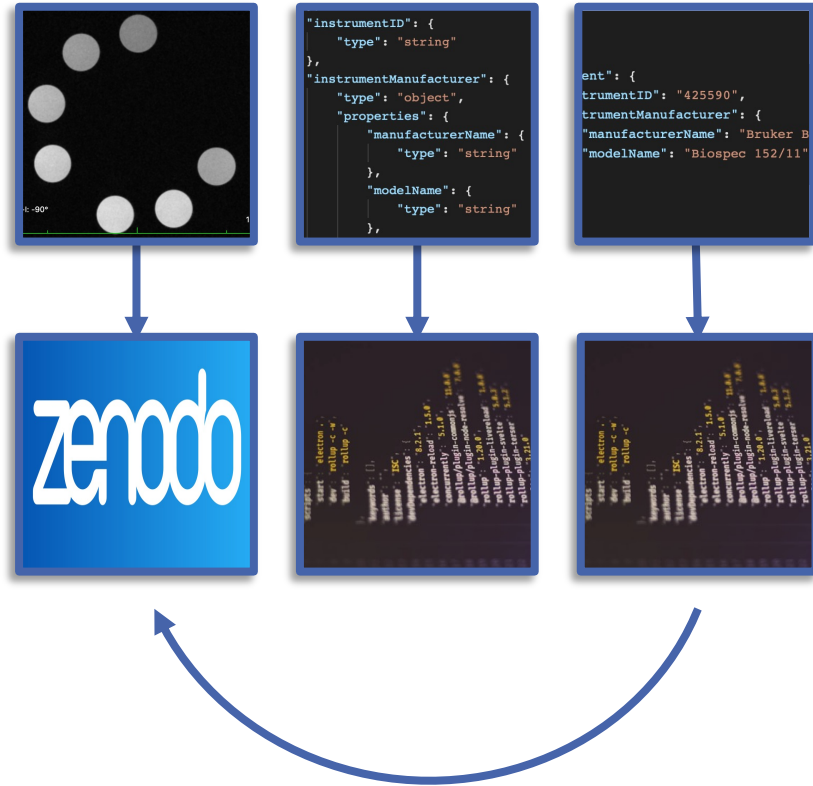
Identifier	Version	Type	Label	Date Updated	
sample_schema	1	JSON	sample	2023-09-12 10:21	👁️ 📄
▼ sem (8) (1 item)					
sem	8	JSON	raw data	2023-11-22 16:22	👁️ 📄
▼ sem_fib_tomography_acquisition (3) (1 item)					
sem_fib_tomography_acq...	3	JSON	raw data	2023-11-24 17:17	👁️ 📄
▼ sem_fib_tomography_dataset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

<https://metarepo.nffa.eu/>

<https://github.com/kat-data-manager/metastore2>

9. Link metadata to data







MetaStore Frontend for NFFA EU Pilot

Schema and Metadata Management

Schema Management | Metadata Management | Search | Show/Hide Filters | Logged in as rossella

Metadata Documents

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

[Register new Metadata Document](#)

9. Link metadata to data

5bc69277-711c-4eb0-937d-d81c59dbbe36

Related Resource <https://doi.org/10.5281/zenodo.6107721>

Schema Identifier [mri_schema \(version=7\)](#)

Date Updated 2023-03-28 15:05

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

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

Published February 16, 2022 | Version 2.0.0

Magnetic Resonance Imaging Copper Sulfate Dataset

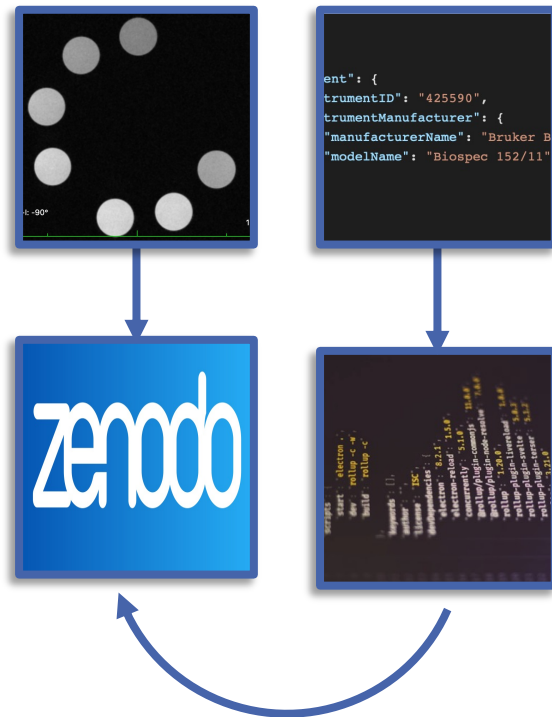
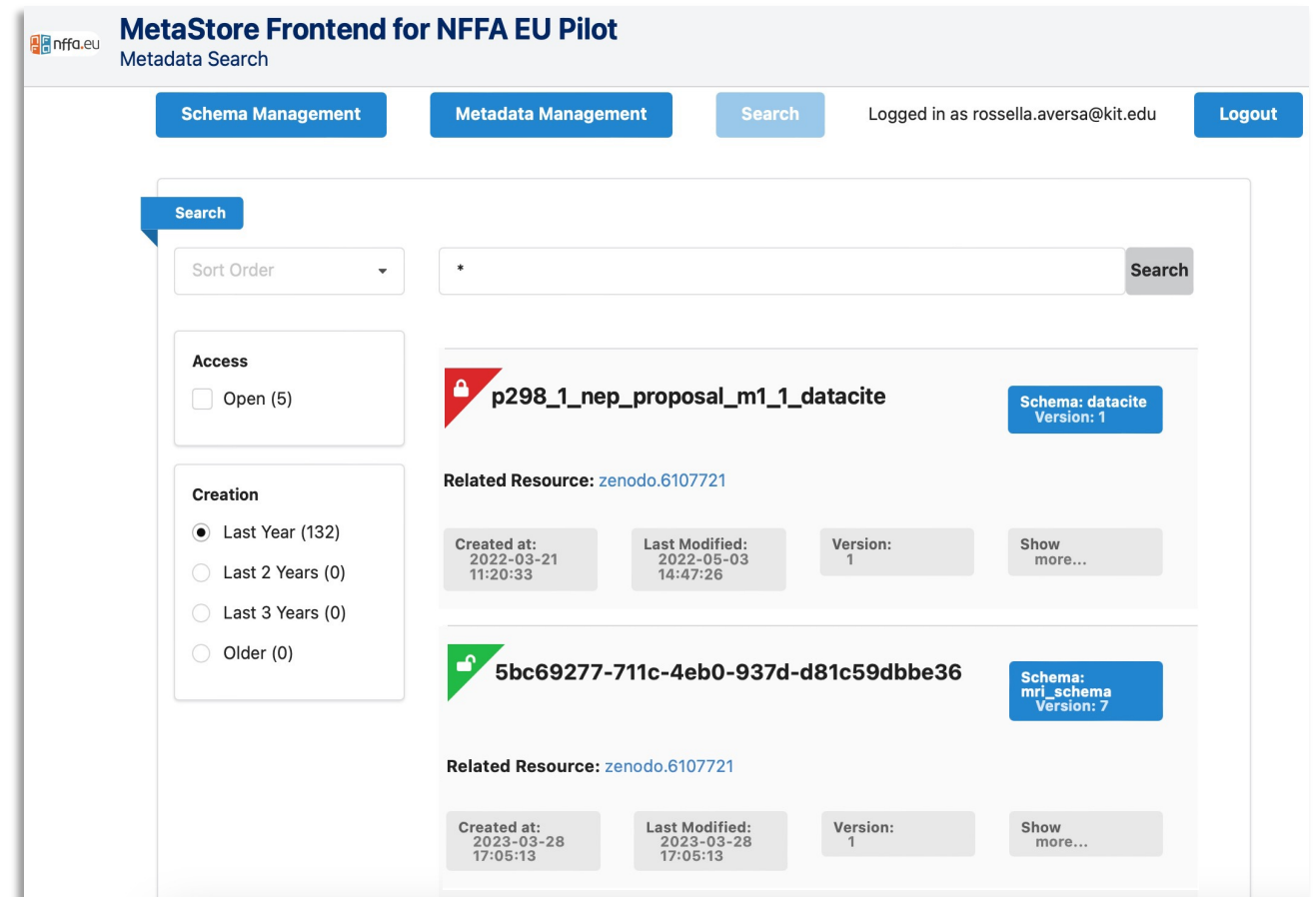
Nicolas Blumenröhr¹

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.

Name	Size	Download
series0.dcm	1.8 MB	Download

10. Find data from metadata

The screenshot shows the MetaStore Frontend for NFFA EU Pilot Metadata Search interface. The page has a header with the logo and title, and navigation buttons for Schema Management, Metadata Management, Search, and Logout. The user is logged in as rossella.aversa@kit.edu. The main content area includes a search bar with a search button, a Sort Order dropdown, and filters for Access (Open (5)) and Creation (Last Year (132), Last 2 Years (0), Last 3 Years (0), Older (0)). Two search results are displayed, each with a lock icon, a title, a schema name, version, and related resource information.

MetaStore Frontend for NFFA EU Pilot
Metadata Search

Schema Management Metadata Management Search Logged in as rossella.aversa@kit.edu Logout

Search

Sort Order *

Search

Access
 Open (5)

Creation
 Last Year (132)
 Last 2 Years (0)
 Last 3 Years (0)
 Older (0)

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 Show more...

5bc69277-711c-4eb0-937d-d81c59dbbe36 Schema: mri_schema Version: 7

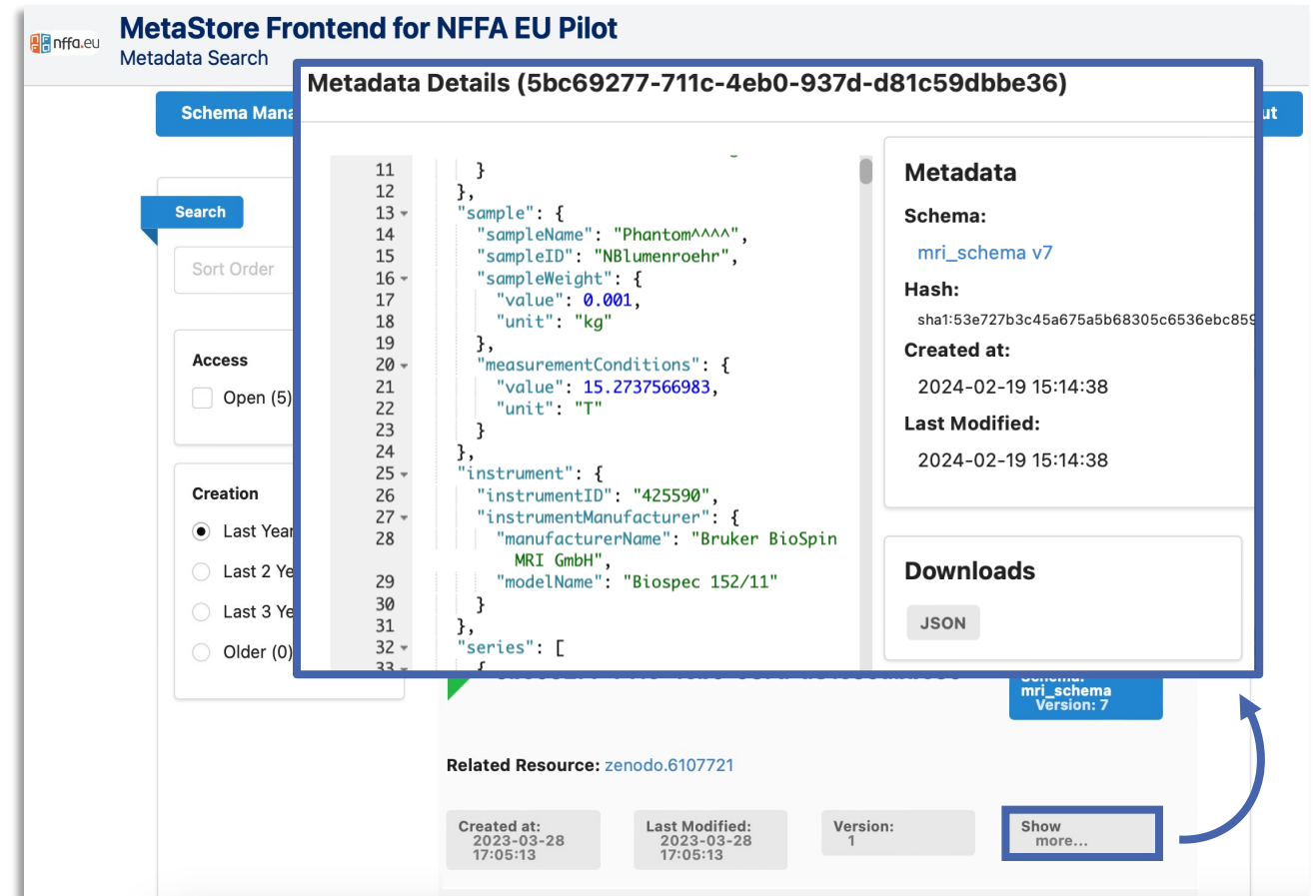
Related Resource: [zenodo.6107721](#)

Created at: 2023-03-28 17:05:13 Last Modified: 2023-03-28 17:05:13 Version: 1 Show more...

10. Find data from metadata

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

Work in Progress



The screenshot displays the 'MetaStore Frontend for NFFA EU Pilot' interface. The main content area shows 'Metadata Details (5bc69277-711c-4eb0-937d-d81c59dbbe36)'. The metadata is presented as a JSON object with the following structure:

```
11  }
12  },
13  "sample": {
14    "sampleName": "Phantom^",
15    "sampleID": "NBlumenroehr",
16    "sampleWeight": {
17      "value": 0.001,
18      "unit": "kg"
19    },
20    "measurementConditions": {
21      "value": 15.2737566983,
22      "unit": "T"
23    }
24  },
25  "instrument": {
26    "instrumentID": "425590",
27    "instrumentManufacturer": {
28      "manufacturerName": "Bruker BioSpin MRI GmbH",
29      "modelName": "Biospec 152/11"
30    }
31  },
32  "series": [
33  ]
```

On the right side, there is a 'Metadata' summary section with the following details:

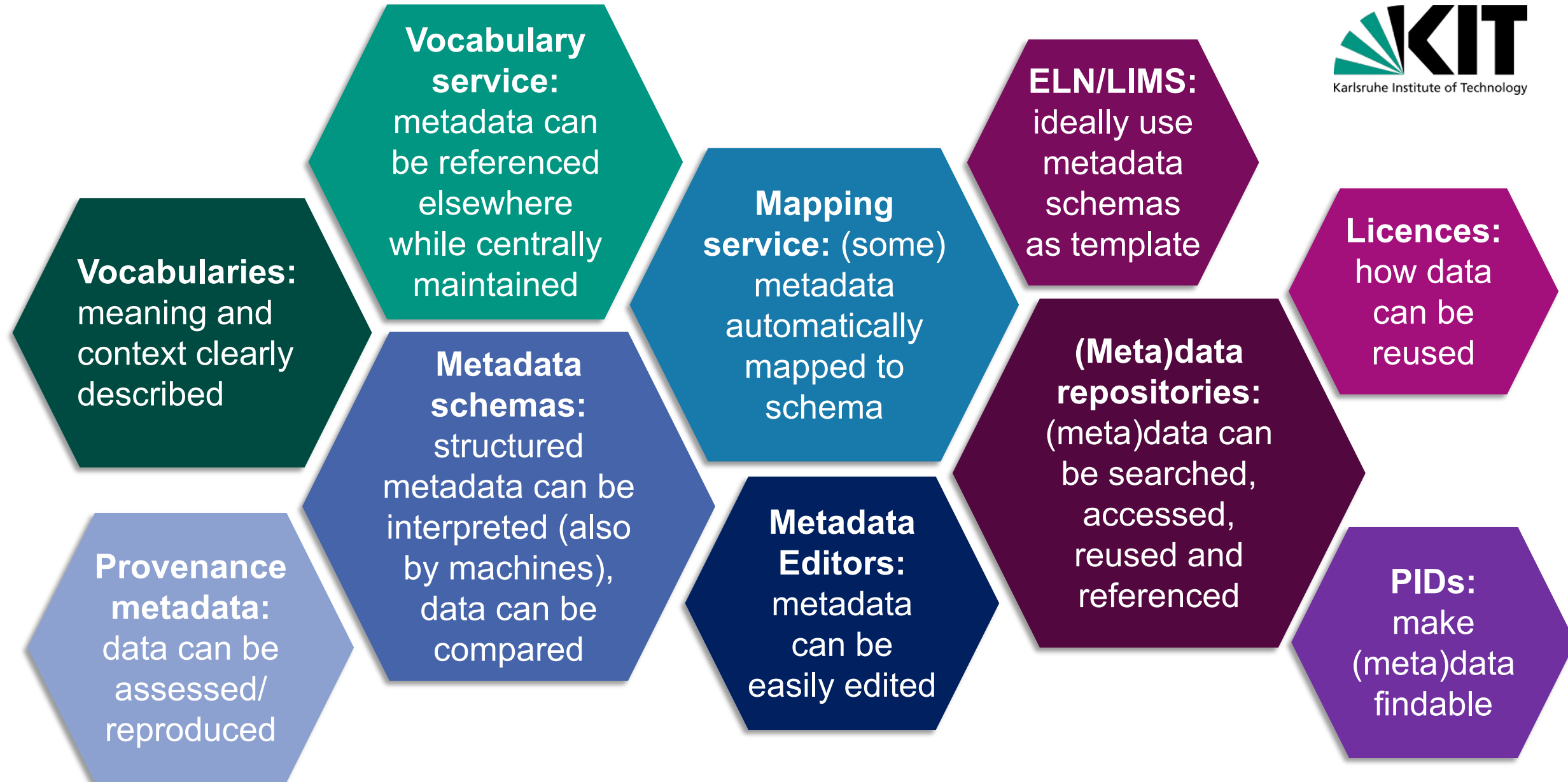
- Schema:** [mri_schema v7](#)
- Hash:** sha1:53e727b3c45a675a5b68305c6536ebc859
- Created at:** 2024-02-19 15:14:38
- Last Modified:** 2024-02-19 15:14:38

Below the summary is a 'Downloads' section with a 'JSON' button. At the bottom, there is a 'Related Resource: zenodo.6107721' and a table of metadata:

Created at:	Last Modified:	Version:	Show more...
2023-03-28 17:05:13	2023-03-28 17:05:13	1	

A blue box highlights the 'mri_schema Version: 7' label and the 'Show more...' button, with a blue arrow pointing from the 'Show more...' button to the 'mri_schema Version: 7' label.

4. Conclusions





Contacts: rossella.aversa@kit.edu

Acknowledgements to:

G. Abdildina, N. Blumenröhr, F. Ernst, V. Haltmann, M. Inkmann, T. Jejkal, A. Kirar, E. Vitali, NEP JA6, JL-MDMC Metadata WG.

Founded by:

The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 38/1 – 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.