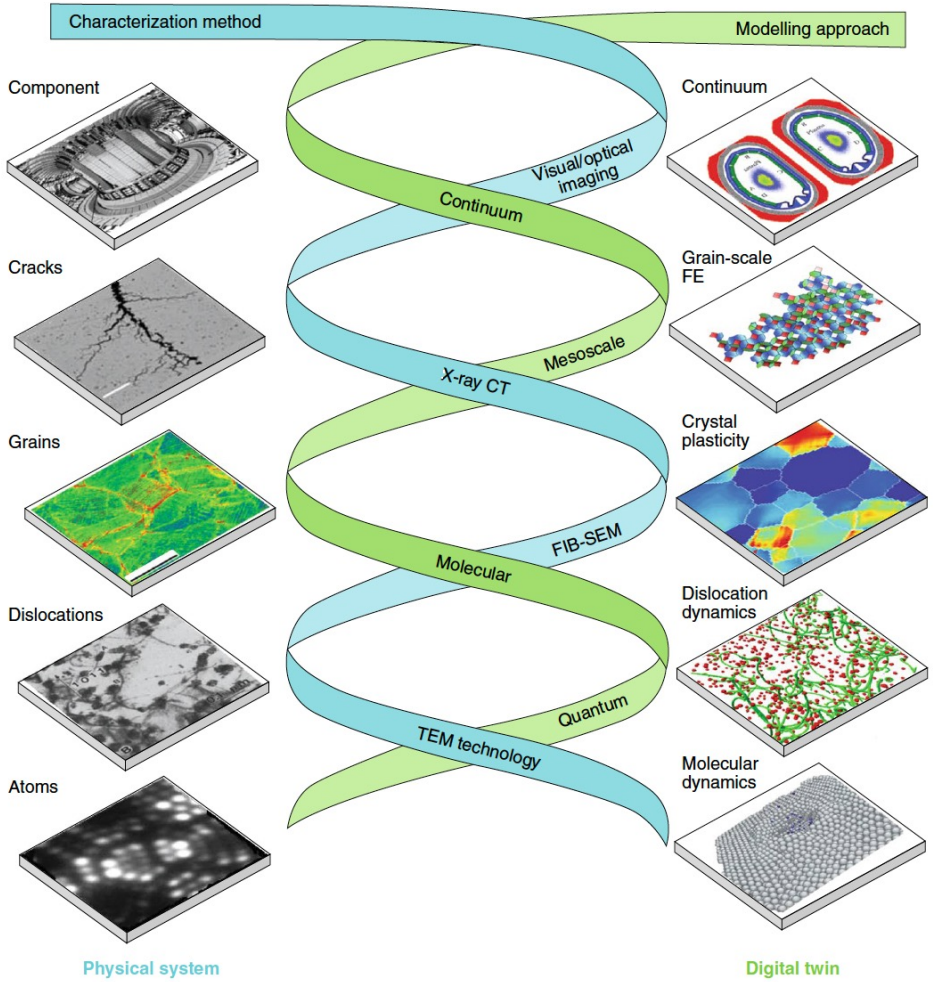
Metadata Management in Correlative Characterization: Tales from the Metadata WG

Rossella Aversa (KIT-SCC)

Italian-German WE-Heraeus-Seminar, Bad Honnef 02-05.04.2024

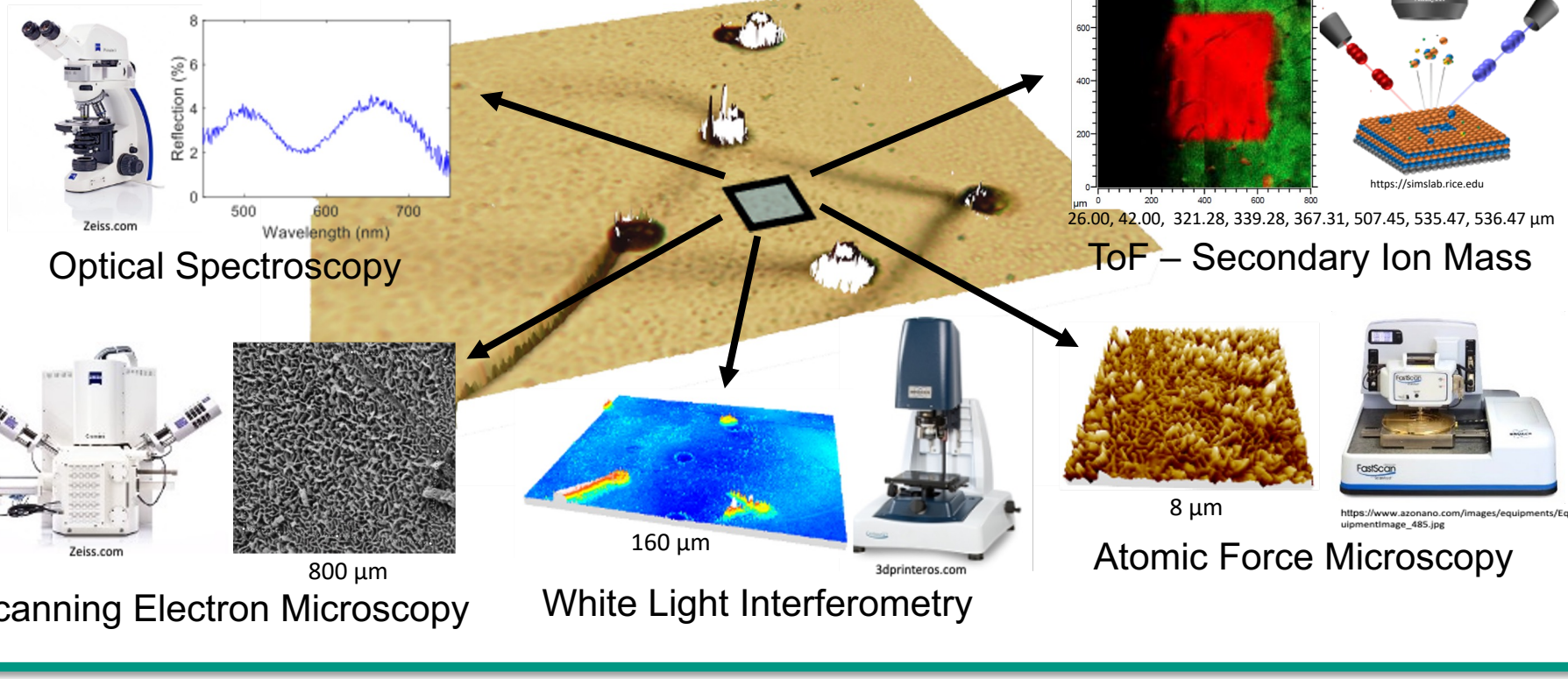
Outline

- Background and motivation
- The Metadata WG
- Achievements
- Next Steps
- Conclusions



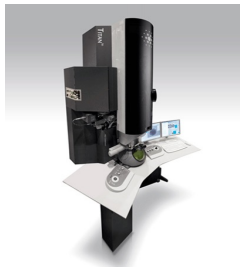
DOI: 10.1038/s41563-019-0402-8

Correlative characterization: butterfly wings

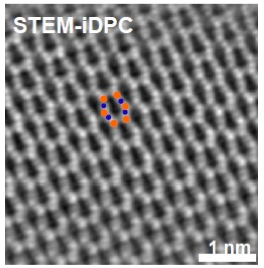


Contact: R. Thelen (KIT)

Correlative characterization: catalyst material

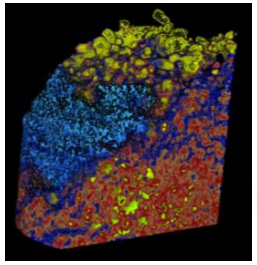


fz-juelich.de



Transmission Electron Microscopy

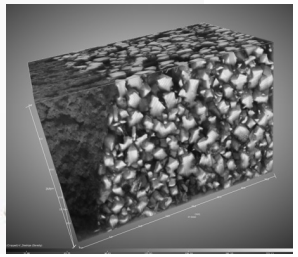
M. Mail, J.-D. Grunwaldt



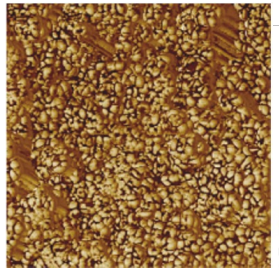
zeiss.de

X-ray microtomography

zeiss.de

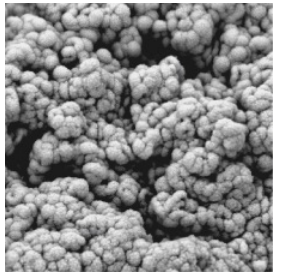


Focused Ion Beam



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

Atomic Force Microscopy



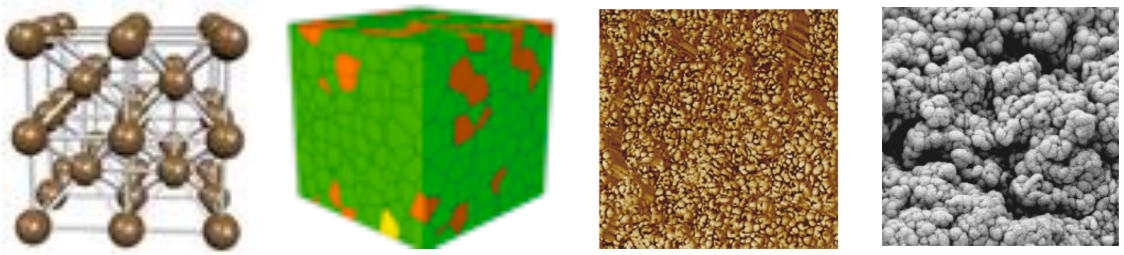
zeiss.de

Scanning Electron Microscopy

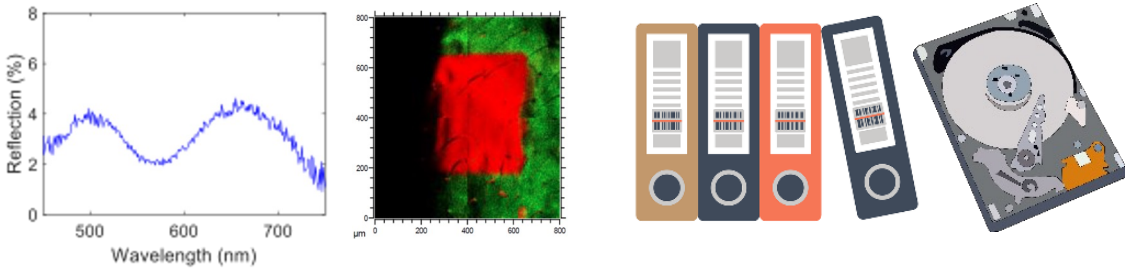
Contact: A. Boubnov (KIT)

Correlative characterization challenges

- Instruments and techniques
- Coordinate systems
- Measurement conditions



- Times
- Length scales
- Regions of the sample



- Data formats
- Software
- Storage systems

Metadata Management

… and use Metadata Management

- **Reproducibility**
  - Describe data with rich metadata
  - Document workflow (provenance)
- **Interoperability**
  - Existing standards and ontologies
  - Common schemas and vocabularies
  - Harmonize as much as possible
- **Infrastructure**
  - Store content and services
  - Human and machine interpretation
  - Easy interaction (GUIs, ELNs)

FAIR Principles

Find existing results

Access available (meta)data

Interoperate data for exchange and comparison

Reuse and reproduce data

The projects

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

- Access to nanoscience research infrastructure
- Integrate synthesis, growth and manipulation of nanostructures with fine analysis, theory and simulation
- 22 international partners, 180 techniques

The projects

Helmholtz Joint-Lab “Model and Data driven Materials Characterization” (JL-MDMC)
https://jl-mdmc-helmholtz.de

- Central methodology platform
- Multiscale and multidimensional characterization, analytics and simulation methods
- 3 Helmholtz Centers (KIT, FZ-Jülich, HZ-Hereon)

The projects

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

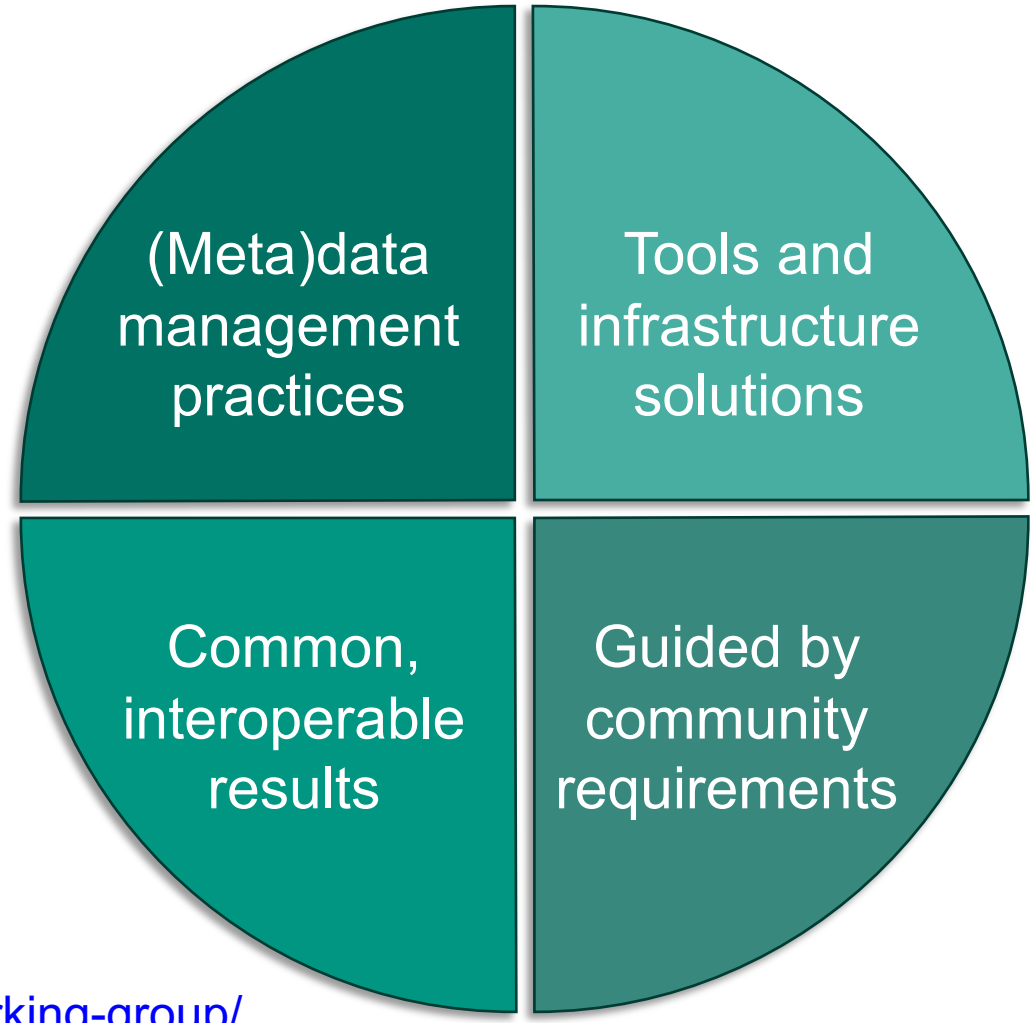
- Infrastructure for the digital representation of materials and their relevant processes
- Involvement of researchers in the development
- 21 Participant Projects, 17 Infrastructure Use Cases

Motivation

- Different communities
- Similar workflows
- Data reproducibility and reuse
- (Meta)data exchange
- Common descriptions and terminologies
- Interoperable solutions

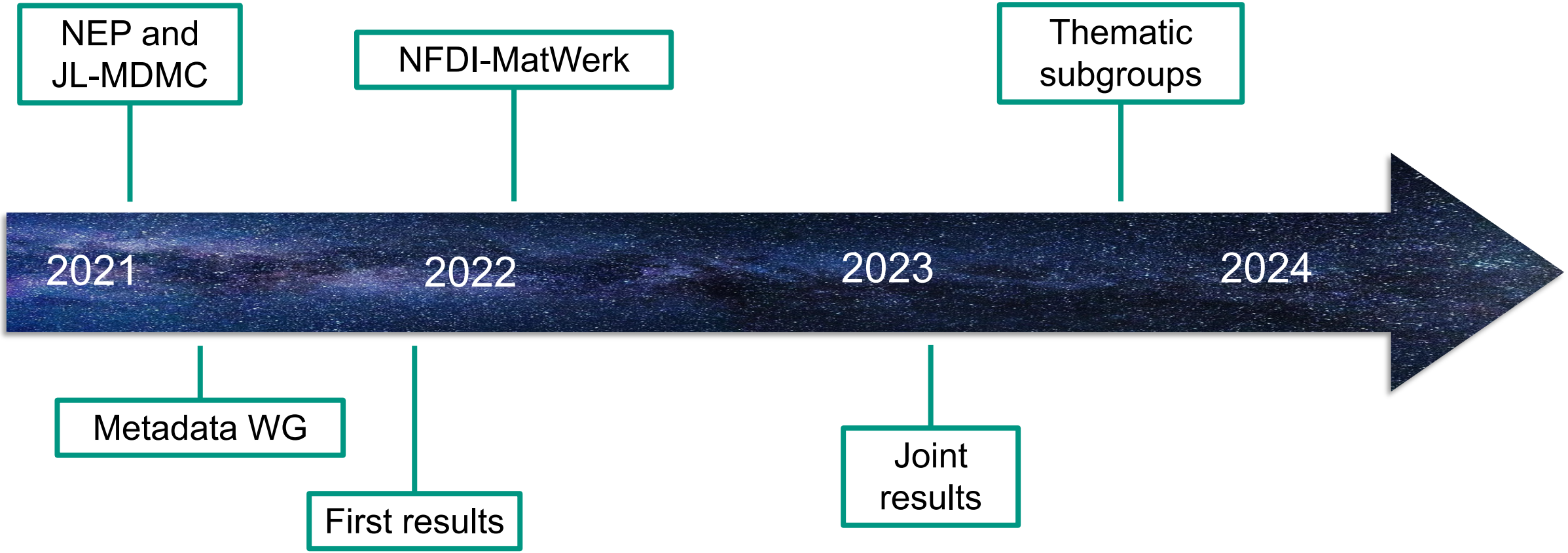
Metadata Working Group

- Collaborative environment
- Consensus-building
- Decision-making
- Voluntary participations
- Grown over time



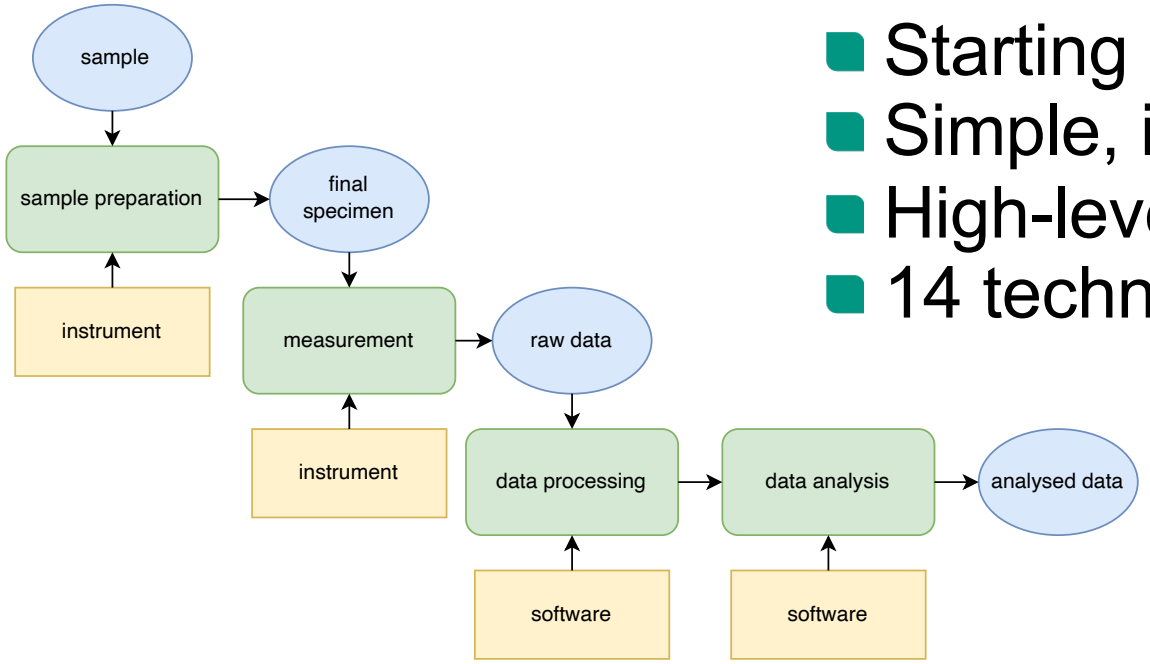
https://jl-mdmc-helmholtz.de/mdmc-activities/metadata-working-group/

Timeline


NEP and JL-MDMC
NFDI-MatWerk
Thematic subgroups
2021
2022
2023
2024
Metadata WG
First results
Joint results

Experimental workflow

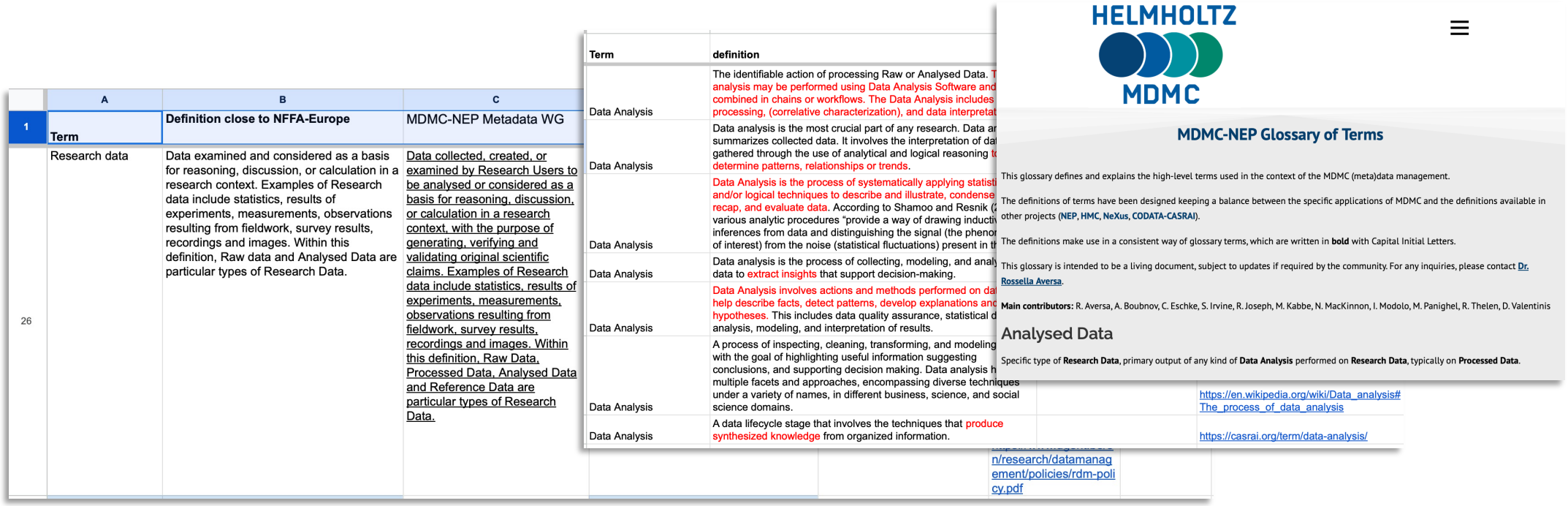
- Define scope and purpose: ensure relevance to specific research goals
- Identify target audience: enhance effectiveness and usability within the scientific community



- Starting point
- Simple, idealized case
- High-level: processes in common
- 14 techniques

Definition of terms

- Shared tables: collaborative work and comments
- Website: promote access and visibility



	A	B	C
1	Term	Definition close to NFFA-Europe	MDMC-NEP Metadata WG
26	Research data	Data examined and considered as a basis for reasoning, discussion, or calculation in a research context. Examples of Research data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, recordings and images. Within this definition, Raw data and Analysed Data are particular types of Research Data.	Data collected, created, or examined by Research Users to be analysed or considered as a basis for reasoning, discussion, or calculation in a research context, with the purpose of generating, verifying and validating original scientific claims. Examples of Research data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, recordings and images. Within this definition, Raw Data, Processed Data, Analysed Data and Reference Data are particular types of Research Data.

Term	definition
Data Analysis	The identifiable action of processing Raw or Analysed Data. T analysis may be performed using Data Analysis Software and combined in chains or workflows. The Data Analysis includes processing, (correlative characterization), and data interpretat
Data Analysis	Data analysis is the most crucial part of any research. Data ar summarizes collected data. It involves the interpretation of da gathered through the use of analytical and logical reasoning t determine patterns, relationships or trends.
Data Analysis	Data Analysis is the process of systematically applying statisti and/or logical techniques to describe and illustrate, condense recap, and evaluate data. According to Shamoo and Resnik (various analytic procedures "provide a way of drawing inducti inferences from data and distinguishing the signal (the phenor of interest) from the noise (statistical fluctuations) present in th
Data Analysis	Data analysis is the process of collecting, modeling, and anal data to extract insights that support decision-making.
Data Analysis	Data Analysis involves actions and methods performed on da help describe facts, detect patterns, develop explanations and hypotheses. This includes data quality assurance, statistical d analysis, modeling, and interpretation of results.
Data Analysis	A process of inspecting, cleaning, transforming, and modeling with the goal of highlighting useful information suggesting conclusions, and supporting decision making. Data analysis h multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.
Data Analysis	A data lifecycle stage that involves the techniques that produce synthesized knowledge from organized information.

n/research/datamanagement/policies/rdm-policy.pdf

https://en.wikipedia.org/wiki/Data_analysis#The_process_of_data_analysis

https://casrai.org/term/data-analysis/

HELMHOLTZ MDMC

MDMC-NEP Glossary of Terms

This glossary defines and explains the high-level terms used in the context of the MDMC (meta)data management.

The definitions of terms have been designed keeping a balance between the specific applications of MDMC and the definitions available in other projects (**NEP, HMC, NeXus, CODATA-CASRAI**).

The definitions make use in a consistent way of glossary terms, which are written in **bold** with Capital Initial Letters.

This glossary is intended to be a living document, subject to updates if required by the community. For any inquiries, please contact Dr. Rossella Aversa.

Main contributors: R. Aversa, A. Boubnov, C. Eschke, S. Irvine, R. Joseph, M. Kabbe, N. MacKinnon, I. Modolo, M. Panighel, R. Thelen, D. Valentinis

Analysed Data

Specific type of **Research Data**, primary output of any kind of **Data Analysis** performed on **Research Data**, typically on **Processed Data**.

Literature review and adoption

- Identify existing terminology: relevant to the research field
- Adoption: ensure alignment with the relevant terms

Crossref

DublinCore

Schema.org

DataCite

MatPortal.org

NeXus

Materials Design Ontology

Metadata for Experiments in Nanoscience Foundries

Vasily Bunakov, Tom Griffin, Brian Matthews & Stefano Cozzini

Using a Core Scientific Metadata Model in Large-Scale Facilities

Brian Matthews, Shoaib Sufi, Damian Flannery, Laurent Lerusse, Tom Griffin, Michael Gleaves, and Kerstin Kleese

Metadata4Ing: An ontology for describing the generation of research data within a scientific activity.

Arndt, Susanne[1]; Farnbacher, Benjamin[2]; Fuhrmans, Marc[3]; Hachinger, Stephan[4]; Hickmann, Johanna[5]; Hoppe, Nils[2]; Horsch, Martin Thomas[6]; Iglezakis, Dorothea[7]; Karmacharya, Ashish[3]; Lanza, Giacomo[8]; Leimer, Sophia[9]; Munke, Johannes[4]; Terzijska, Džulia[10]; Theissen-Lipp, Johannes[11]; Wiljes, Cord[12]; Windeck, Jürgen[3]

Materials Resource Registry

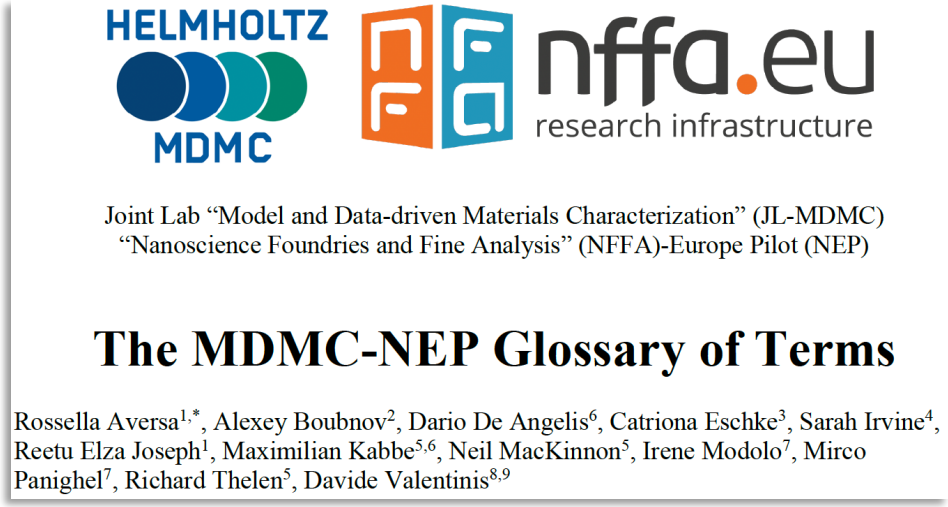
Part of the Materials Genome Initiative

CODATA CASRAI RDM Terminology

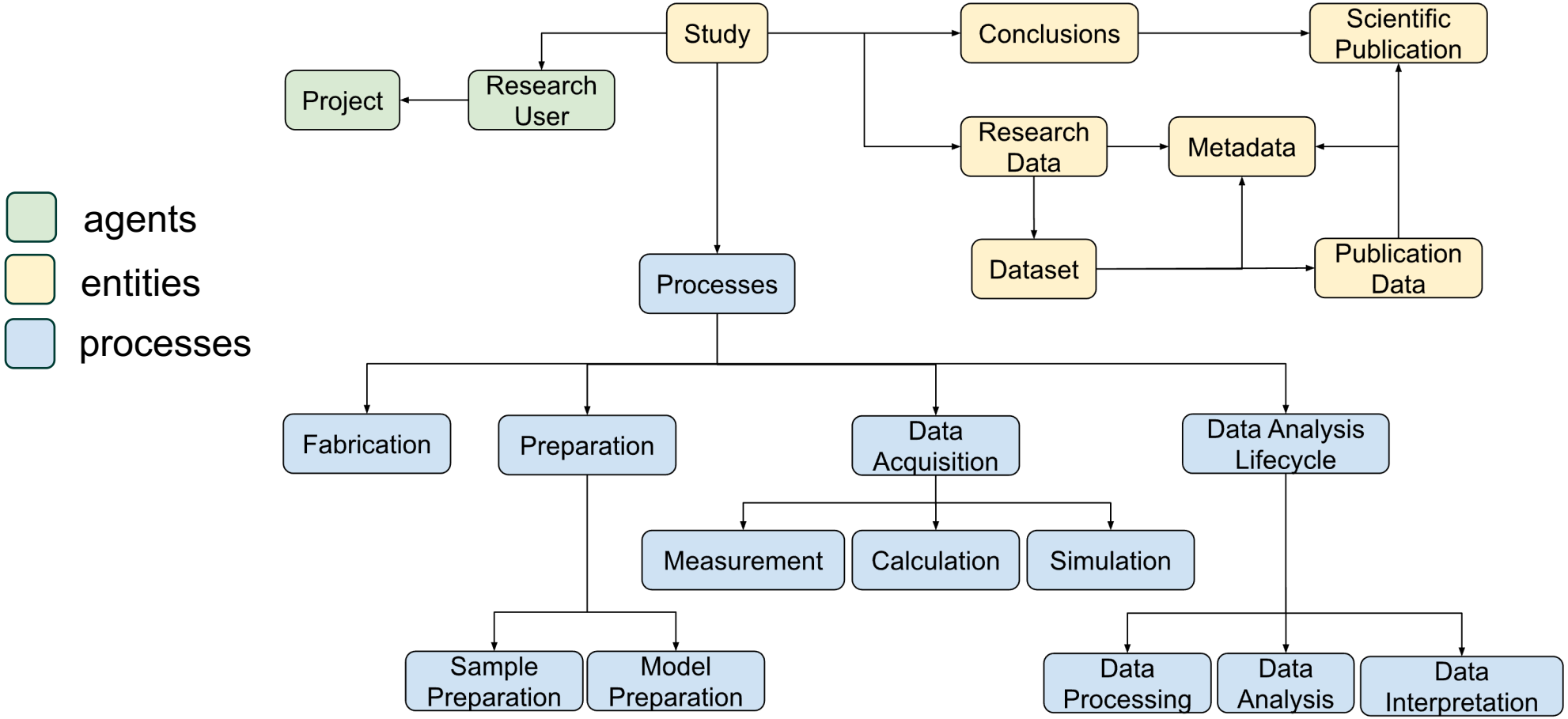
PhySH - Physics Subject Headings

MDMC-NEP Glossary

- 45 high-level terms
- Describes computational/experimental workflows
- Reflects the lifecycle of entities and data
- Framed in the management infrastructure of the involved projects
- Tracks basic provenance information
- Living document
- DOI: 10.5281/zenodo.10663833

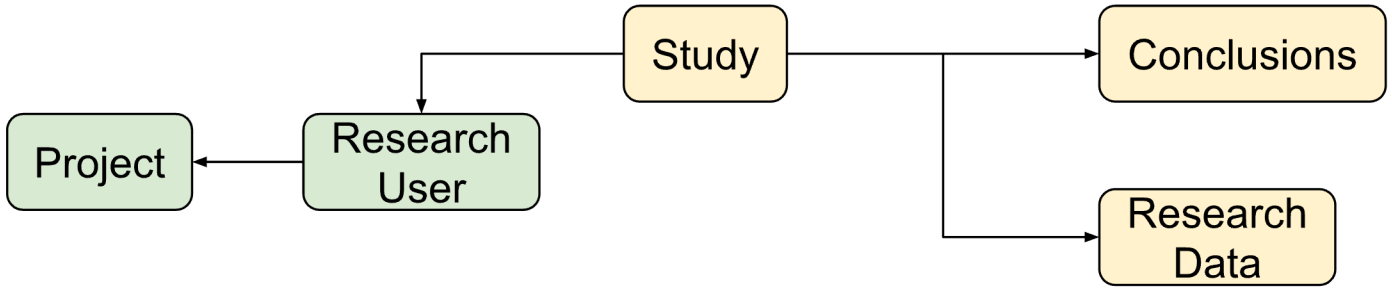


MDMC-NEP Glossary: overview



MDMC-NEP Glossary of Terms. DOI: 10.5281/zenodo.10663833

MDMC-NEP Glossary: overview

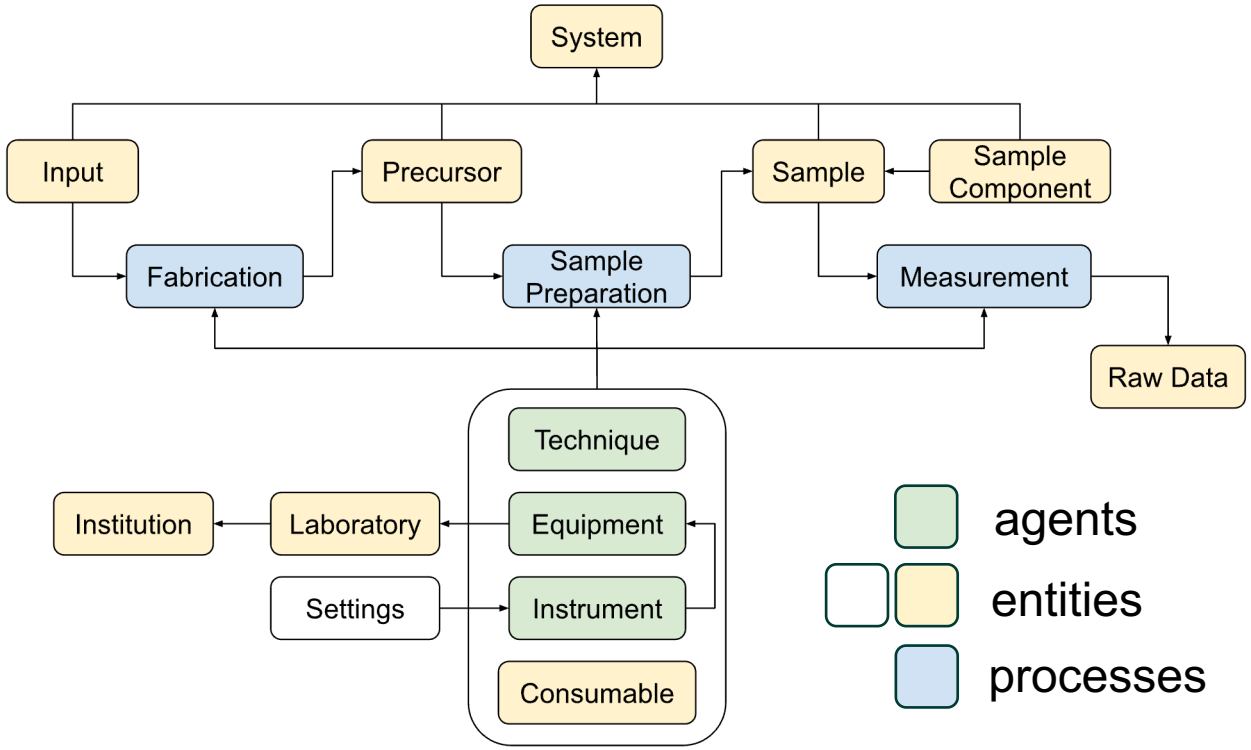


Research User
Person, usually member of a **Project**, who conducts any part of the **Study**, in order to collect and/or analyse **Research Data** or is interested in reusing **Research Data** by a third party (e.g., **Reference Data**) with the final aim to extract insights that support the answer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

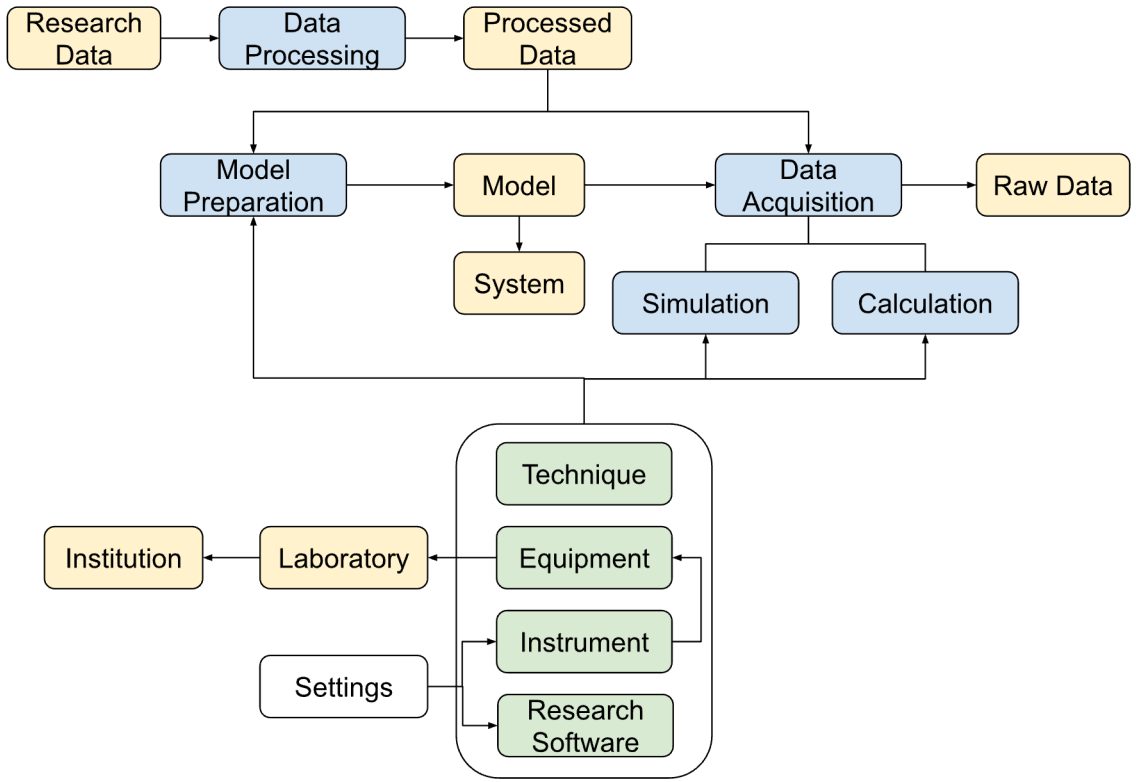
MDMC-NEP Glossary of Terms. DOI: 10.5281/zenodo.10663833

Updated workflows

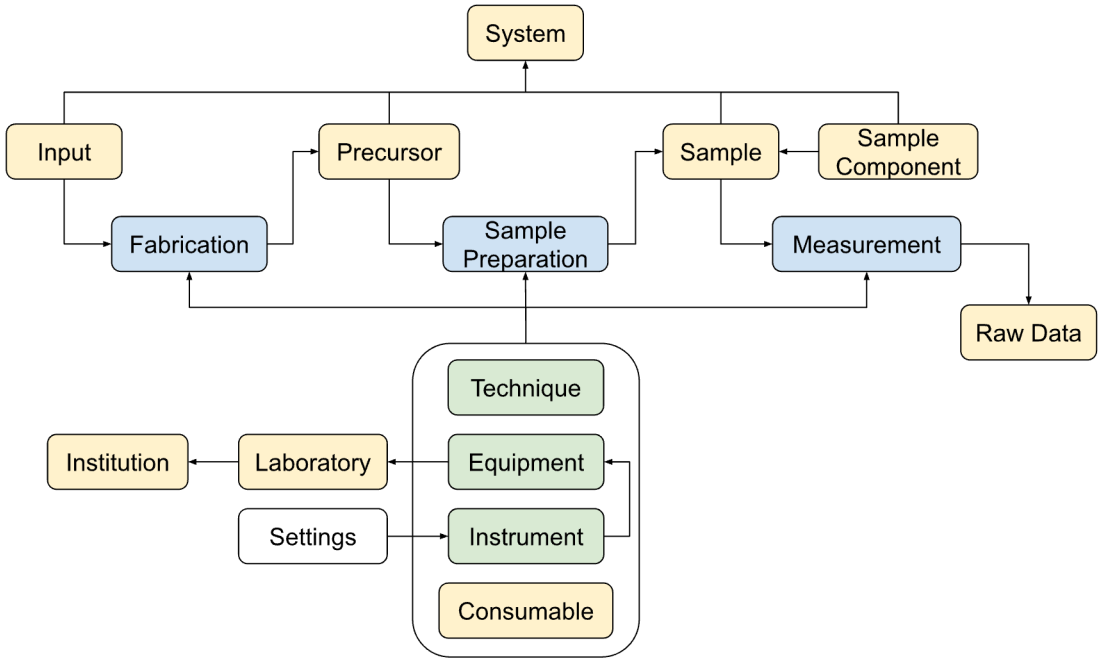
Experimental Workflow



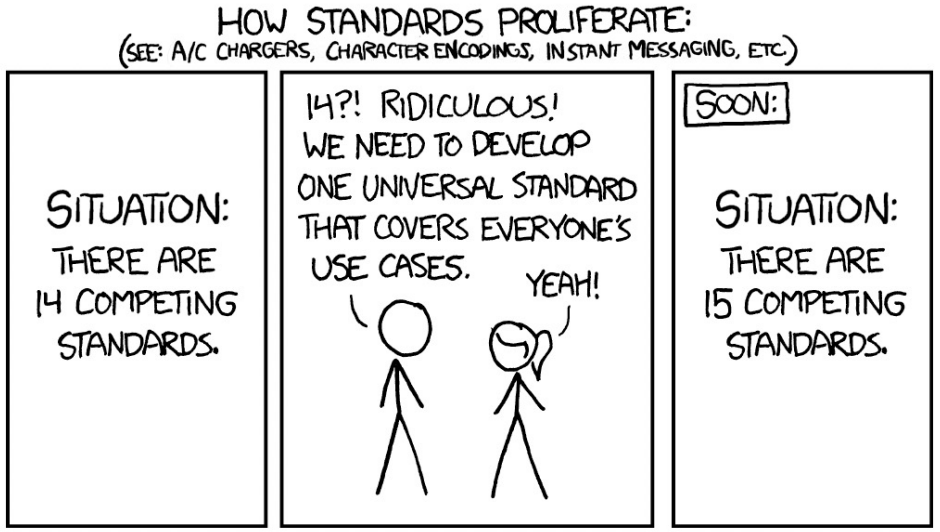
Computational Workflow



Metadata Schemas



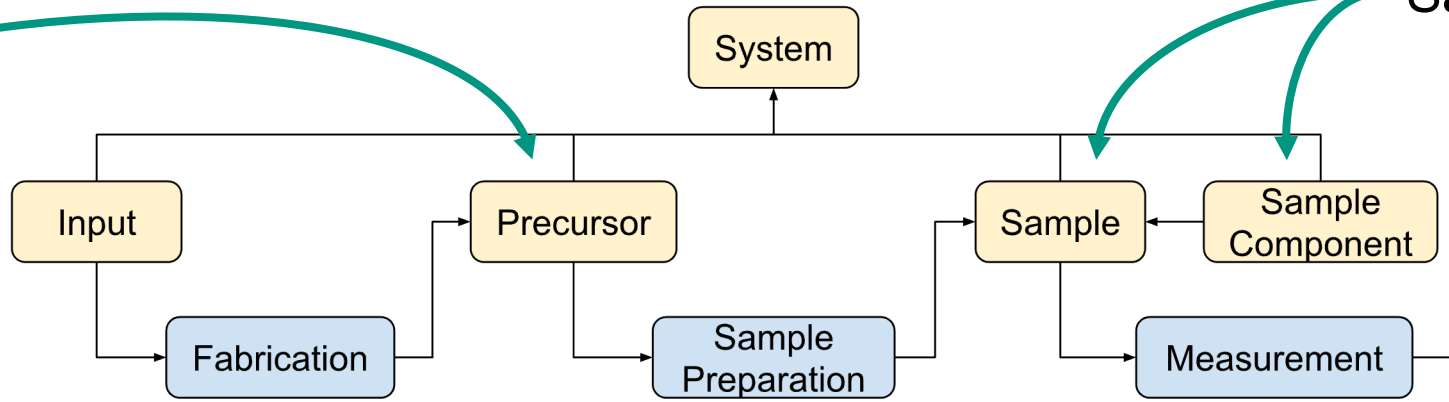
- JSON schema
- Describe inputs/outputs of processes
- Adopt existing solutions
- Avoid proliferation of schemas



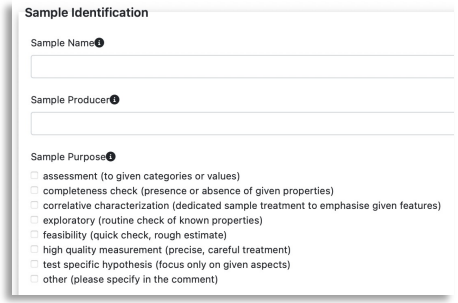
https://xkcd.com/927/

Metadata Schemas

Minimal



Sample description



Ongoing collaboration
with CNR-IOM and FBK

Adopted from the Materials
Data Vocabulary
DOI: 10.5334/dsj-2021-018

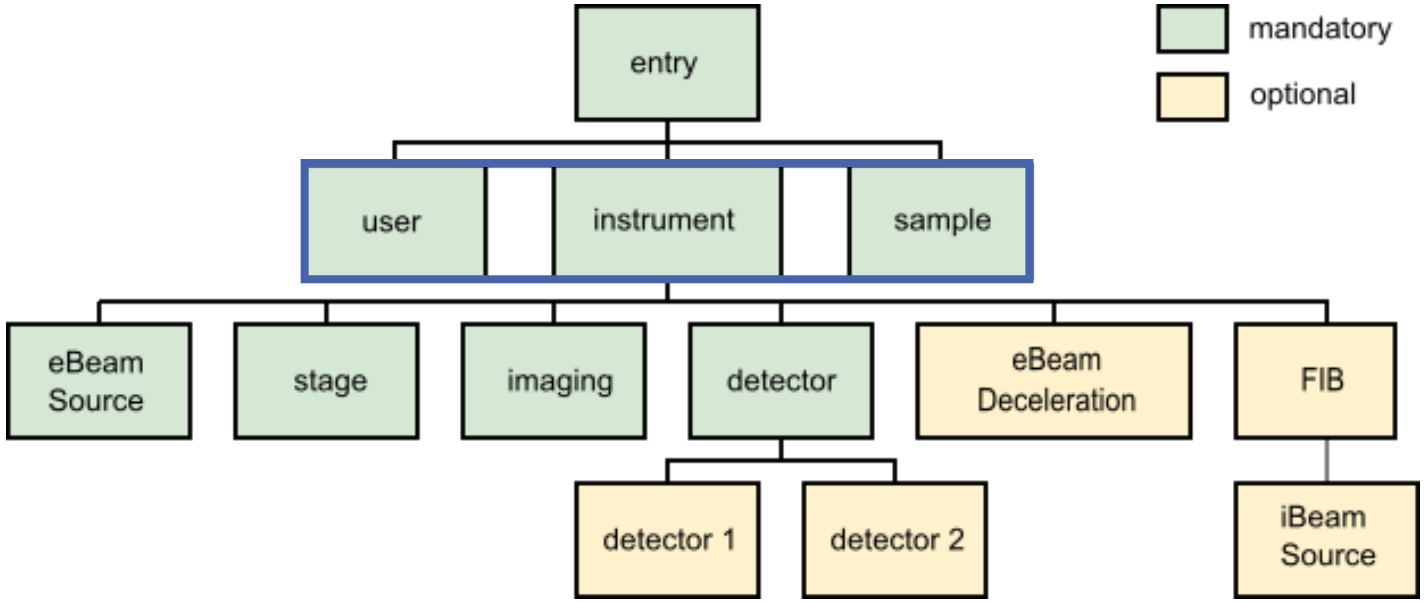
SEM
TEM
MRI
STM
SEM/FIB Tomography
Nano CT/micro CT

Metadata Schema for SEM

Metadata Schema to support FAIR Data in Scanning Electron Microscopy

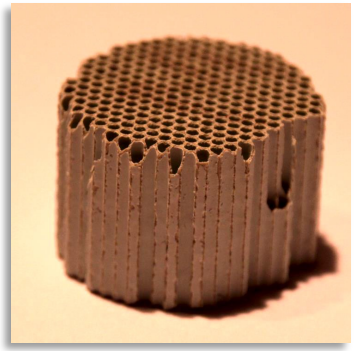
Reetu Joseph[1][0000−0002−1507−9327], Aditya Chauhan[2], Catriona Eschke[3][0000−0002−1033−144X], Ahmad Zainul Ihsan[4][0000−0002−1008−4530], Mehrdad Jalali[5][0000−0003−2465−4933], Ute Jäntsch[6], Nicole Jung[7][0000−0001−9513−2468], C. N. Shyam Kumar[8][0000−0003−4860−5327], Christian Kübel[5,10,11][0000−0001−5701−4006], Christian Lucas[3], Matthias Mail[5,10][0000−0002−9732−8453], Andrey Mazilkin[5], Charlotte Neidiger[5], Mirco Panighel[9][0000−0001−8413−5196], Stefan Sandfeld[4][0000−0001−9560−4728], Rainer Stotzka[1][0000−0003−3642−1264], Richard Thelen[2], and Rossella Aversa[1][0000−0003−2534−0063]

DOI: 10.5445/IR/1000141604



Other Metadata Schemas available at:
https://github.com/kit-data-manager/Metadata-Schemas-for-Materials-Science

Metadata Schema for Sample Description



Size x
value
unit
Size y
value
unit
Size z
value
unit

Sample Holder
Sample Holder Type
✓ Not applicable
stub
dish
cylinder
glass slide
TEM grid
tilting support
custom holder
Other (please add in the comments)
unit
Sample Holder size y
value
unit

Sample Identification
Sample Name
Sample Producer
Sample Purpose
assessment (to given categories or values)
completeness check (presence or absence of given properties)
correlative characterization (dedicated sample treatment to emphasise given features)
exploratory (routine check of known properties)
feasibility (quick check, rough estimate)
high quality measurement (precise, careful treatment)
test specific hypothesis (focus only on given aspects)
other (please specify in the comment)

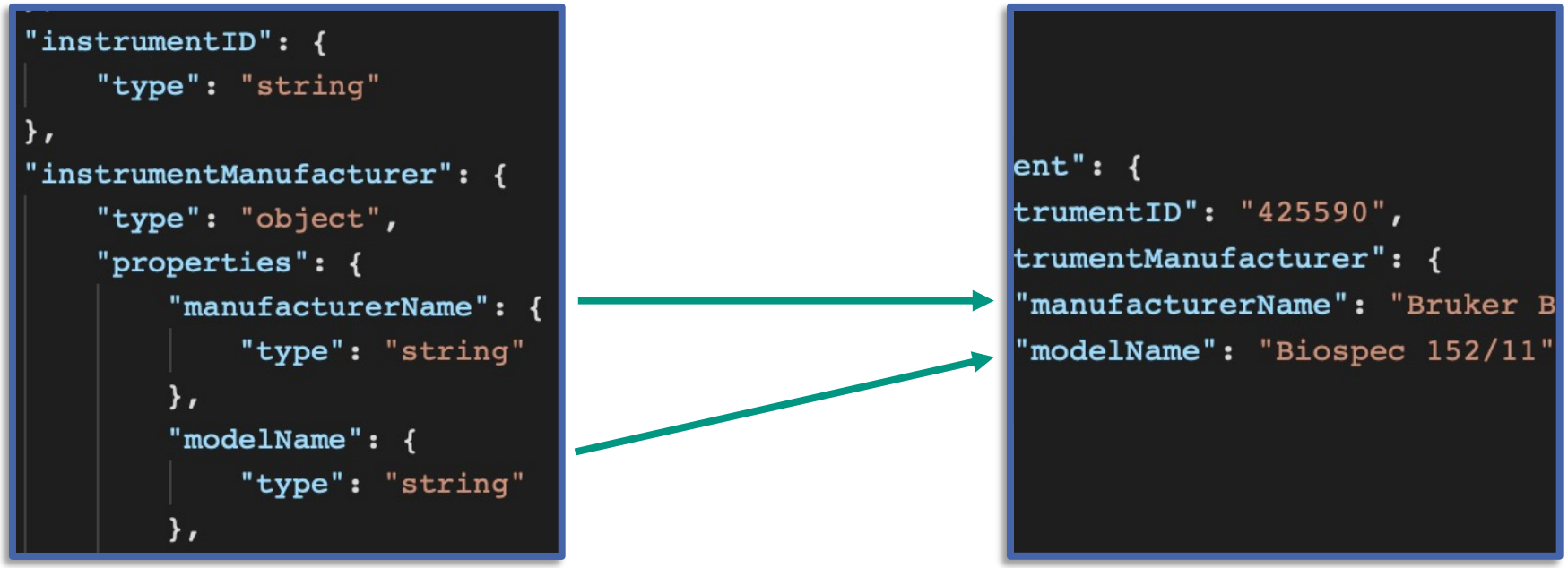
Images: courtesy of M. Mail, R. Thelen

https://kit-data-manager.github.io/Metadata-Schemas-for-Materials-Science/sample_interface/

Metadata Schemas and Documents

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

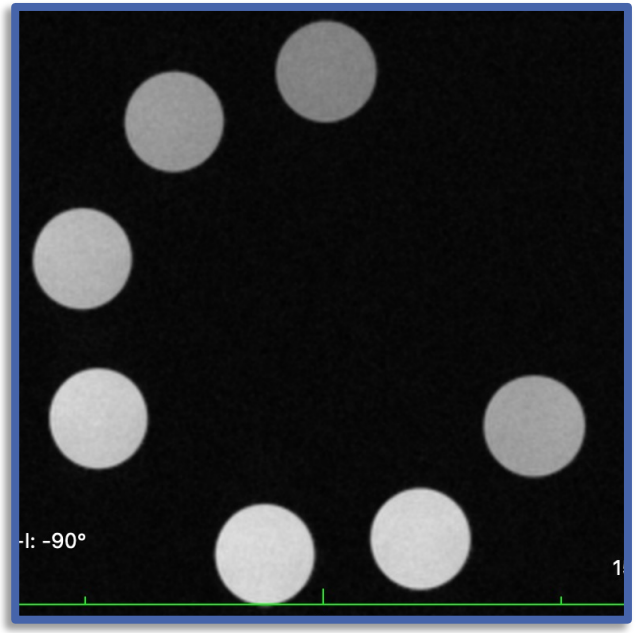
Metadata Document: structured information about a data resource



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

From data to metadata

Raw Data



Metadata Schema

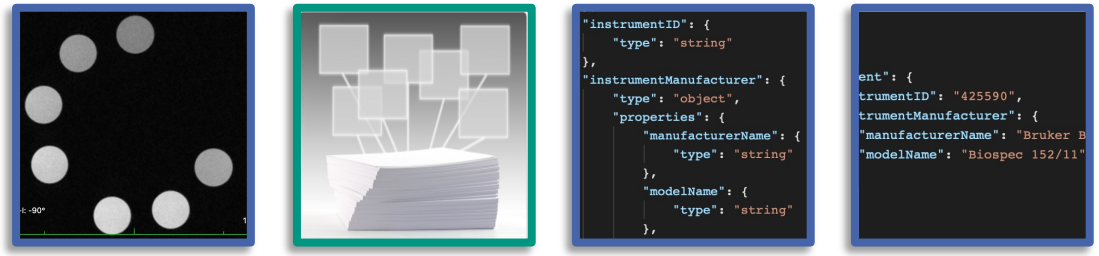
```
"instrumentID": {
    "type": "string"
},
"instrumentManufacturer": {
    "type": "object",
    "properties": {
        "manufacturerName": {
            "type": "string"
        },
        "modelName": {
            "type": "string"
        },
```

Metadata Document

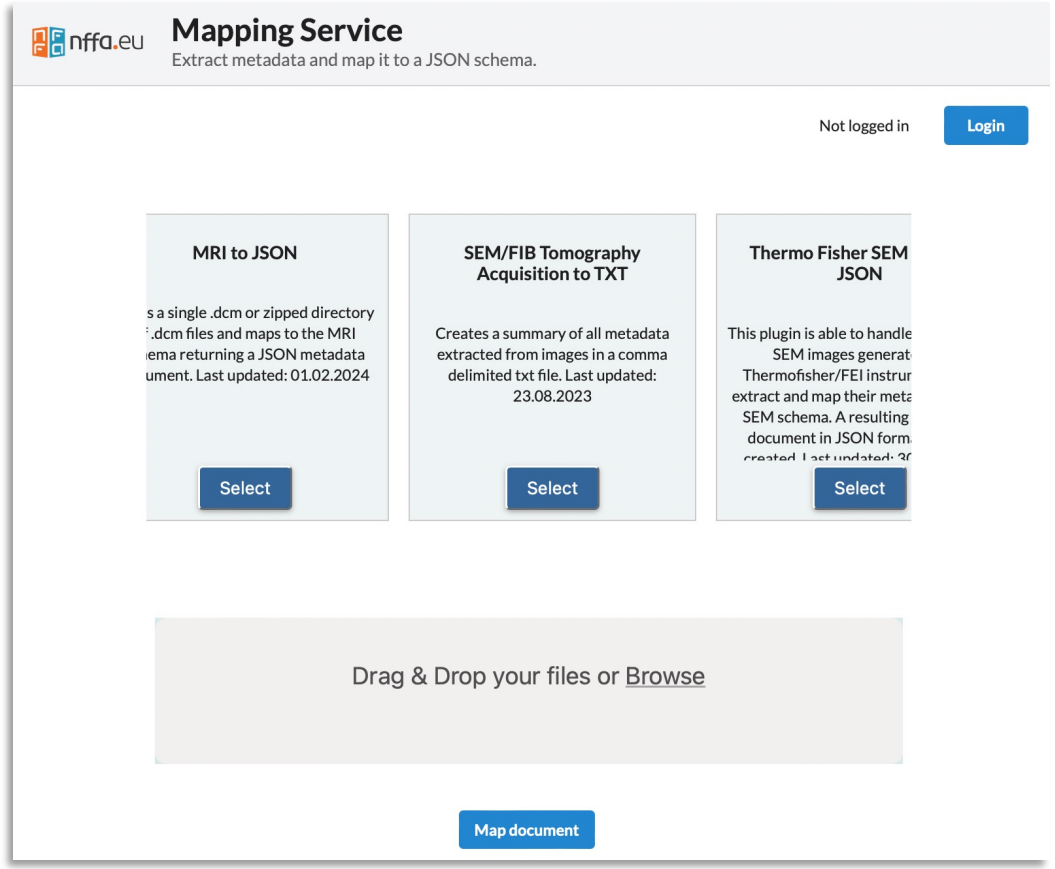
```
ent": {
trumentID": "425590",
trumentManufacturer": {
"manufacturerName": "Bruker B
"modelName": "Biospec 152/11"
```

Image from Magnetic Resonance Imaging Copper Sulfate Dataset. DOI: 10.5281/zenodo.6107720

Mapping service

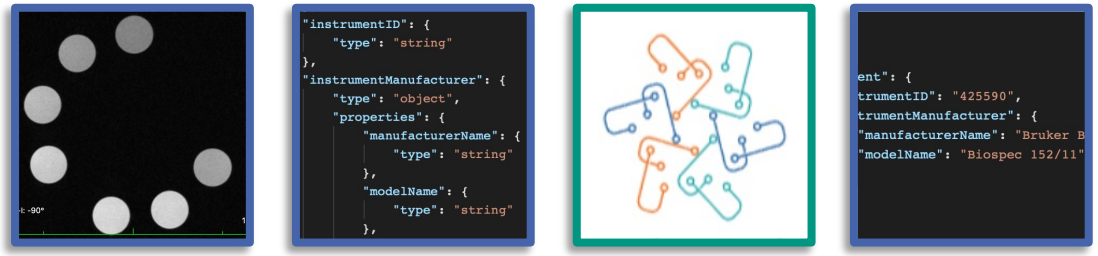


- NEP Virtual Access service, MatWerk
- Input: data file(s)
- Extract unstructured metadata
- Map them to the metadata schema
- Output: structured metadata



https://metarepo.nffa.eu/frontend/mapping-service-ui.html
https://matwerk.datamanager.kit.edu/mapping-service-ui.html

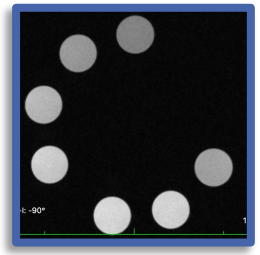
Metadata Editor

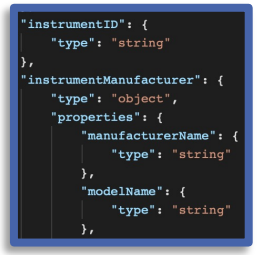


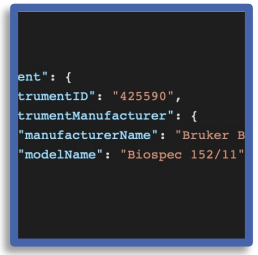
- Local service connected to the metadata repository
- Load schema from registered ones
- Load existing metadata documents
- Manually edit metadata documents
- Download metadata documents
- Register metadata documents
- Create the provenance file

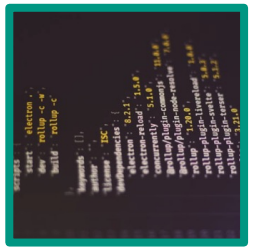
https://metadata-editor.gitlab.io/documentation/

MetaStore

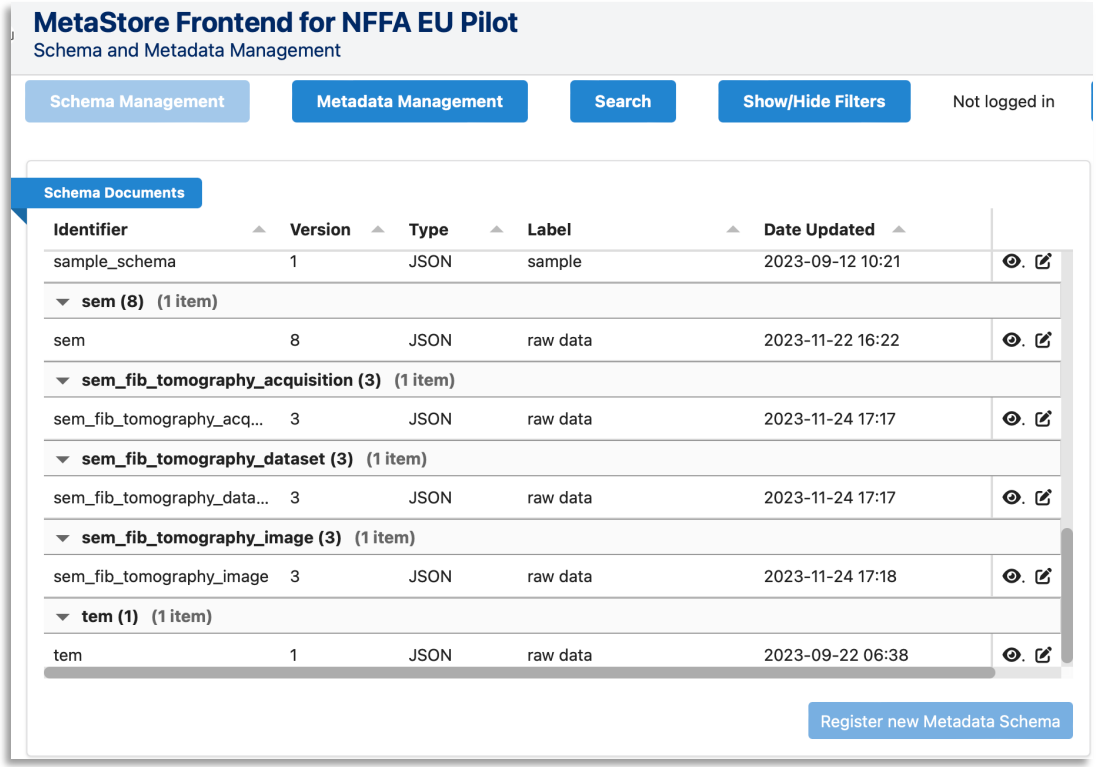






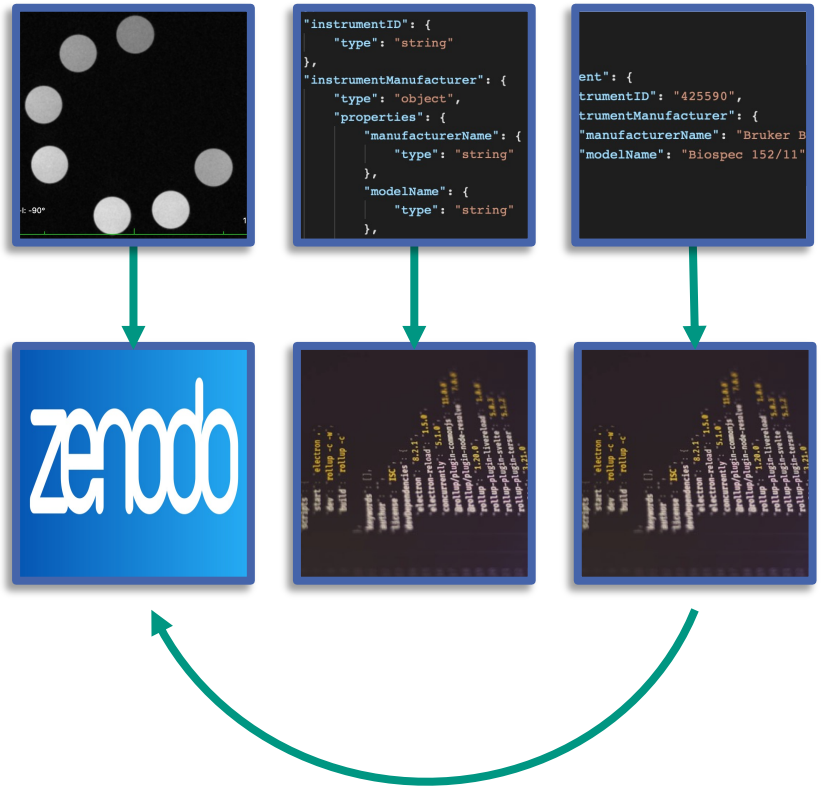


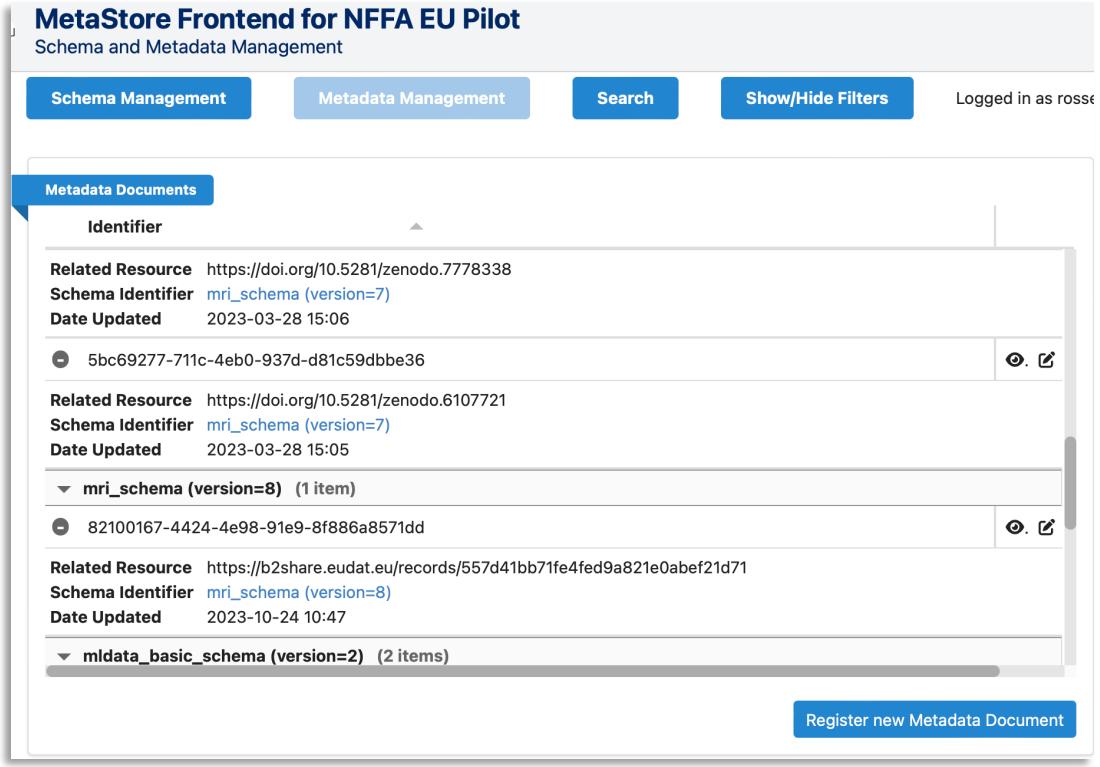
- Metadata repository
- Register/find metadata schemas
- Register/find metadata documents
- Validate metadata documents
- Versioning
- Access control management
- User authentication



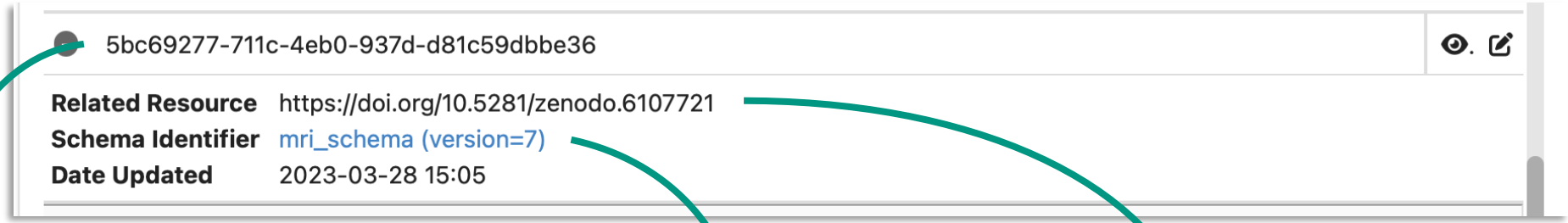
https://metarepo.nffa.eu/frontend/schema-management.html
https://matwerk.datamanager.kit.edu/schema-management.html

Link metadata to data



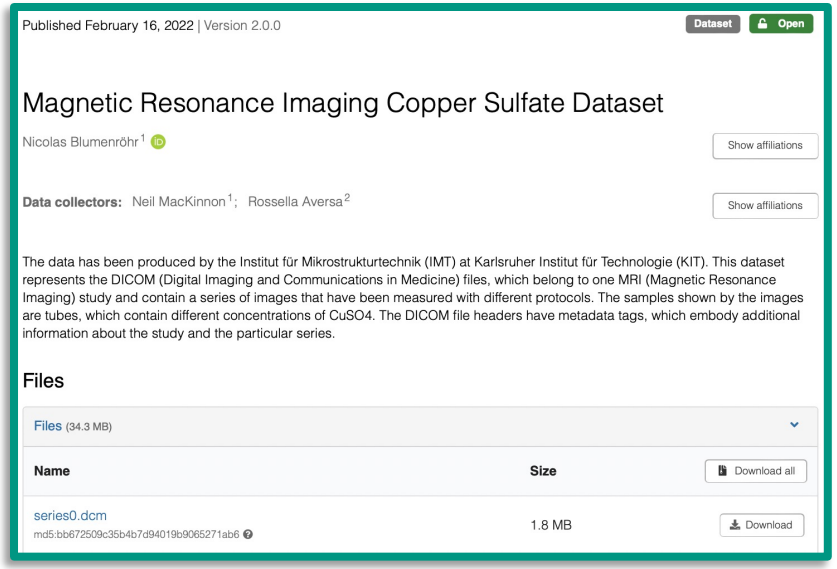


Link metadata to data




```
"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"
        },
```



Application to Correlative Characterization

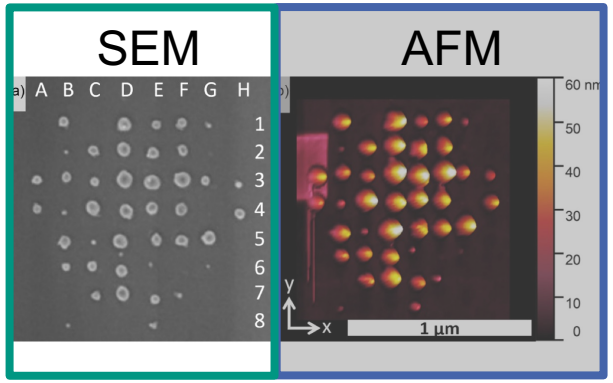
Karlsruhe Institute of Technology

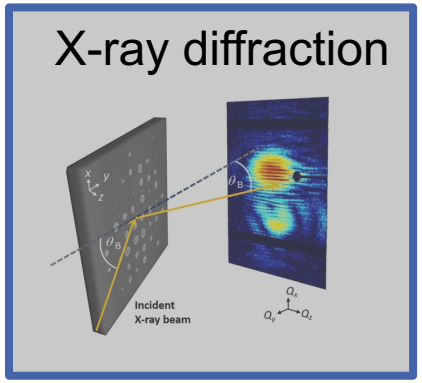
Paper

Coherent x-ray diffraction of a semiregular Pt nanodot array

Thomas F. Keller,[1,2,*,†] Roman Shayduk,[3,†] Chan Kim,[3] Nastasia Mukharamova,[1] Arti Dangwal Pandey,[1] Manuel Abuin,[1] Vedran Vonk,[1] Irene Fernandez-Cuesta,[2] Miriam Barthelmess,[4] Robert Frömter,[2,5] Alexey Zozulya,[3] Artur Erbe,[6] and Andreas Stierle[1,2]

[1]*Centre for X-ray and Nano Science (CXNS), Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany*
[2]*University of Hamburg, Department of Physics, Hamburg, Germany*
[3]*European XFEL GmbH, Schenefeld, Germany*
[4]*Center for Free-Electron Laser Science (CFEL), Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany*
[5]*Institute of Physics, Johannes Gutenberg-Universität Mainz, Mainz, Germany*
[6]*Helmholtz Zentrum Dresden Rossendorf, Dresden, Germany*





Zenodo

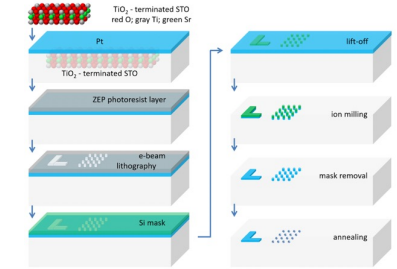
SEM Metadata Document

```
"measurementPurpose": {
    "measurementPurposeOptions": [
        "correlative characterization
        "high quality measurement (pre
    ]
"parents": [
    {
        "parentType": "sample",
        "parentReferenceType": "MetaStore URI",
        "parentReference": "https://metarepo.nffa.eu/api/
    }
],
```

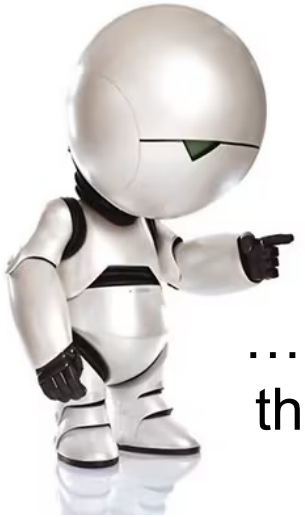
Sample Metadata Document

```
"samplePurpose": {
    "samplePurposeOptions": [
        "correlative characterization
        "high quality measurement (pre
    ]
"preparationAction": "mechanicalAndSurface",
"mechanicalAndSurface": {
    "mechanicalAndSurfaceMethod": "lithography",
    "comments": "creation of pattern and markers",
    "consumables": [
        "ZEP photoresist"
    ]
```

Sample Preparation



Next steps



… and go on with the development

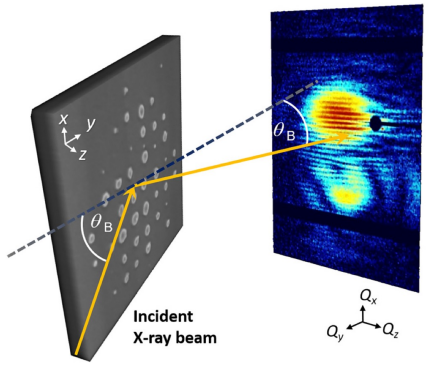
- **Reproducibility**
  - Describe data with rich metadata
  - Document workflow (provenance)
- **Interoperability**
  - Existing standards and ontologies
  - Common schemas and vocabularies
  - Harmonize as much as possible
- **Infrastructure**
  - Store content and services
  - Human and machine interpretation
  - Easy interaction (GUIs, ELNs)

Next steps: include other techniques

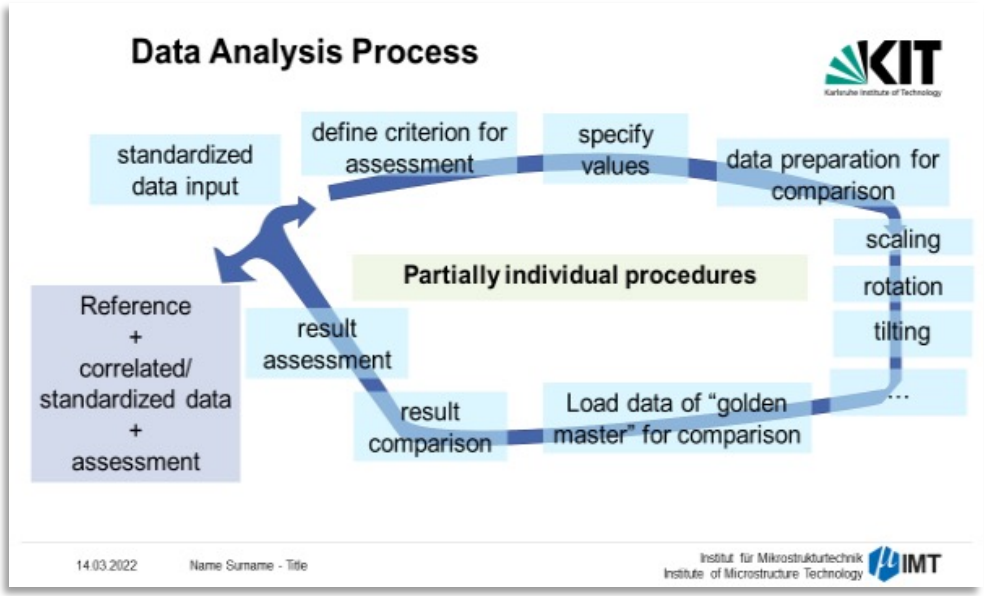
Atomic Force Microscopy

White Light Interferometry

Optical Spectroscopy



X-ray diffraction



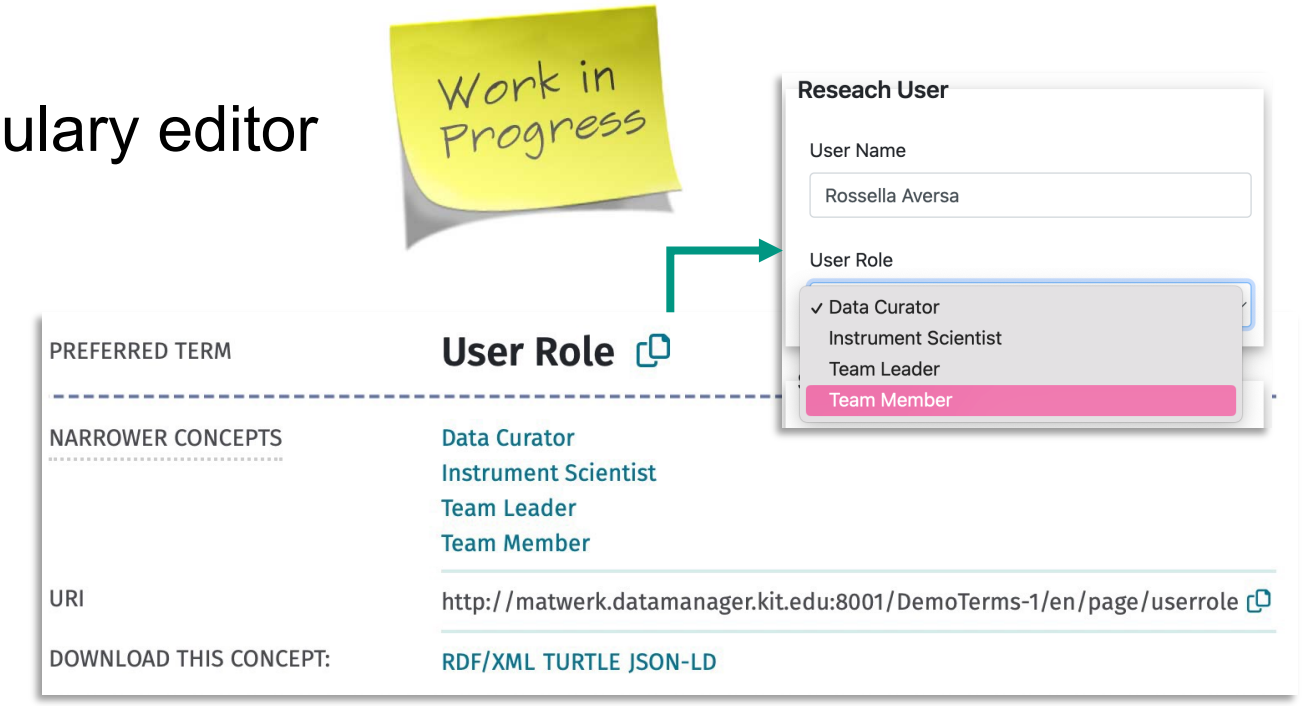
Courtesy of R. Thelen

Next steps: maintenance and updates

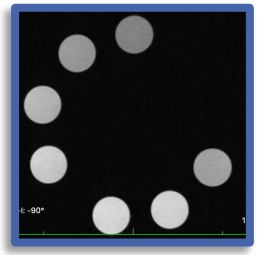
- Incorporate new terms: dynamic resource
- Enhance accessibility and interoperability: vocabulary service

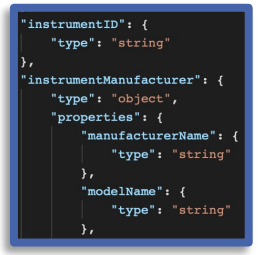
EVOKS: Collaborative online vocabulary editor

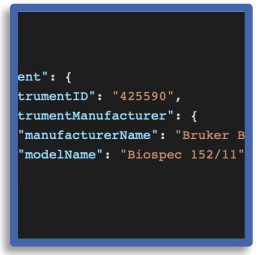
- Developed at KIT
- SKOS model
- Persistent identifier to each term
- Centrally maintained
- Seamless integration of terms



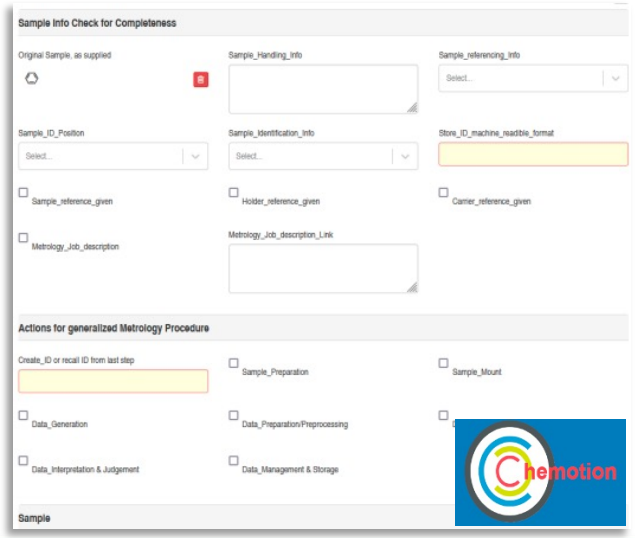
Next steps: ELN and LIMS

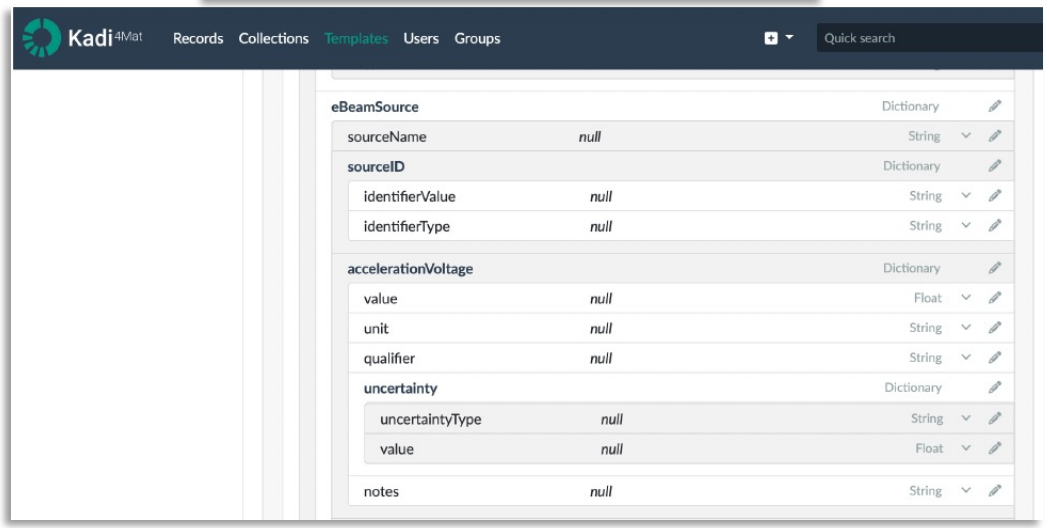






- Electronic Lab Notebooks
- Lab Information Management Systems
- Metadata schemas as templates
- Ongoing collaborations:
  - KIT
  - HZ-Hereon
  - CNR-IOM
  - FBK

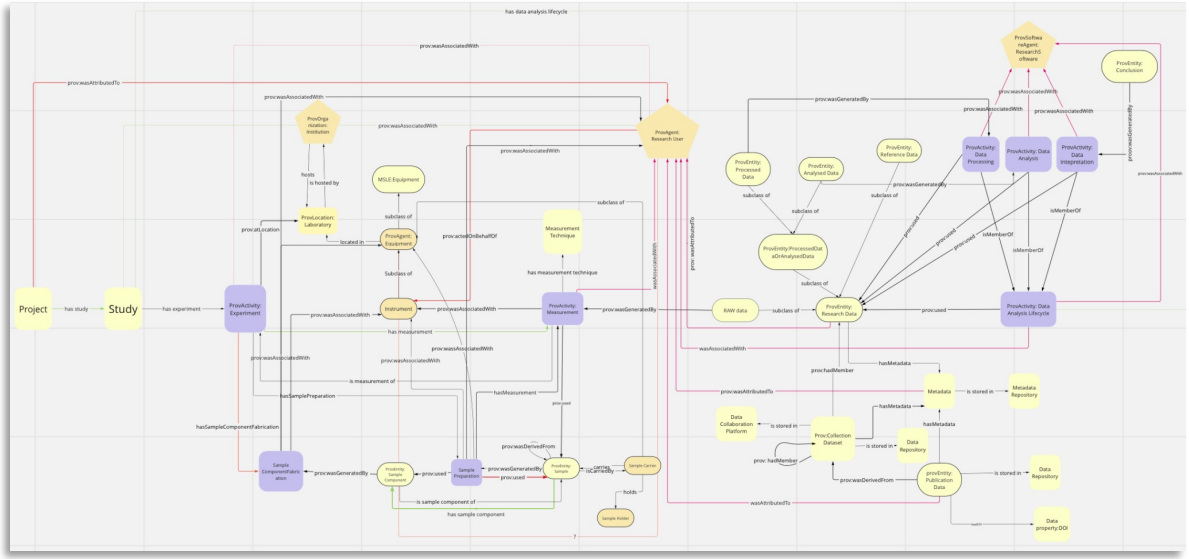


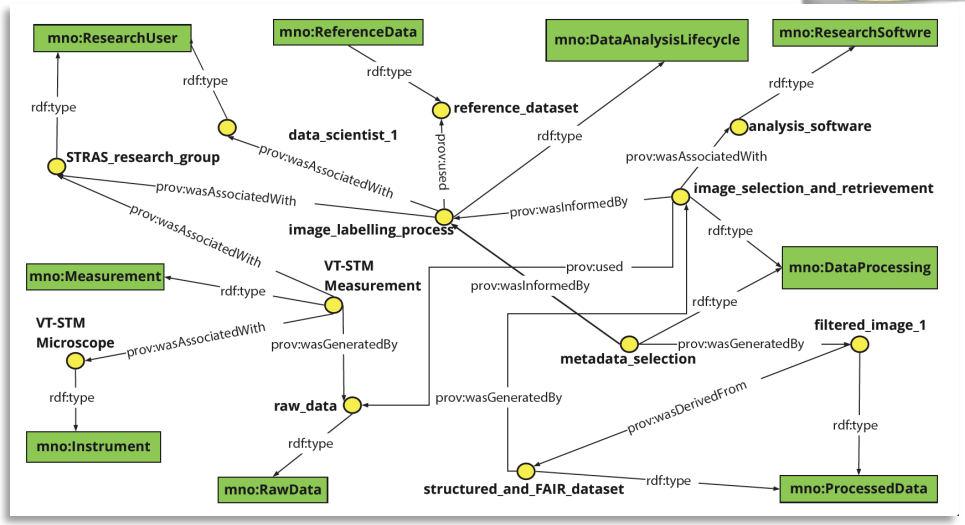


Next steps: from glossary to ontology

- Formalize relationships: add context to the knowledge representation
- Logical constructs (classes, properties, axioms): structured and machine-readable

PRIMA (PRovenance Information for MAterials science) Ontology





Work in Progress

https://jl-mdmc-helmholtz.de/ontology/

Conclusions

- Tools, services and best practices to facilitate (meta)data management
- Guided by community requirements
- Driven by communication with scientists for collecting information
- Applied in daily research life
- Useful for correlative characterization experiments
- Open to exchange with other materials science facilities and projects

Contacts: rossella.aversa@kit.edu

Acknowledgements to:
N. Blumenröhr, F. Ernst, V. Haltmann, M. Inkmann, T. Jejkal, Metadata WG, NEP JA6, NFDI-MatWerk TA-MDI, PP13.

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.