Metadata Management in Scientific Research: an overview

Rossella Aversa (KIT-SCC)
iEntrance Advanced School, 22.02.2024
1. Introduction
Motivation
Metadata

Data describing other data

1. What is metadata?

2. Why is metadata needed?

3. How to manage metadata?
Metadata

To add context and meaning to data

1. What is metadata?
2. Why is metadata needed?
3. How to manage metadata?
We will see it throughout the next slides…
The FAIR Guiding Principles

Findable

Accessible

Interoperable

Reusable

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

Findable  Accessible  Interoperable  Reusable

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

Findable  Accessible  Interoperable  Reusable

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.

Images: courtesy of R. Thelen and M. Mail. Powered by Bing Image Creator
How to describe data?

I2: Metadata use vocabularies that follow the FAIR principles.

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

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)
- [NeXus](http://www.nexusformat.org)
- [Crystallography](https://www.iucr.org/resources/cif)
How to represent structured metadata?

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](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

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

Details

DOI
DOI 10.5281/zenodo.7776338

Resource type
Dataset

Publisher
Zenodo

Languages
English

ORCID

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.
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?

R1.1: (Meta)data are released with a clear and accessible data usage licence.

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
- 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

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).

MDMC-NEP Glossary of Terms. DOI: 10.5281/zenodo.10663833
2. Terms in a Vocabulary Service

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
3. Metadata schemas

- Minimal
- System
- Input
- Precursor
- Sample
- Sample Component
- Fabrication
- Sample Preparation
- Measurement
- Raw Data

Ongoing collaboration with CNR-IOM and FBK (see talk by L. Ferrario)

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

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

Work in Progress
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

MRI schema, DOI: [10.5445/IR/1000159552](http://10.5445/IR/1000159552)
From data to metadata

Image from Magnetic Resonance Imaging Copper Sulfate Dataset. DOI: 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

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

https://metadata-editor.gitlab.io/documentation/
7. ELN and LIMS

- 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

https://metarepo.nffa.eu/
https://github.com/kit-data-manager/metastore2
9. Link metadata to data

MetaStore Frontend for NFFA EU Pilot
Schema and Metadata Management

- Identifier
  - Related Resource: https://doi.org/10.5281/zenodo.7778338
  - Schema Identifier: mri_schema (version=7)
  - Date Updated: 2023-03-28 15:06
  - Related Resource: https://doi.org/10.5281/zenodo.6107721
  - Schema Identifier: mri_schema (version=7)
  - Date Updated: 2023-03-28 15:05
  - Related Resource: https://bitshare.eudat.eu/records/557d041bb71fe4fe4f3a02e0b0/1d71
  - Schema Identifier: mri_schema (version=8)
  - Date Updated: 2023-10-24 10:47

Register new Metadata Document
9. Link metadata to data
10. Find data from metadata
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
4. Conclusions
Vocabularies: meaning and context clearly described

Metadata schemas: structured metadata can be interpreted (also by machines), data can be compared

Provenance metadata: data can be assessed/reproduced

Vocabulary service: metadata can be referenced elsewhere while centrally maintained

Mapping service: (some) metadata automatically mapped to schema

Metadata Editors: metadata can be easily edited

ELN/LIMS: ideally use metadata schemas as template

(Meta)data repositories: (meta)data can be searched, accessed, reused and referenced

Licences: how data can be reused

PIDs: make (meta)data findable
Contacts: rossella.aversa@kit.edu

Acknowledgements to:

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.