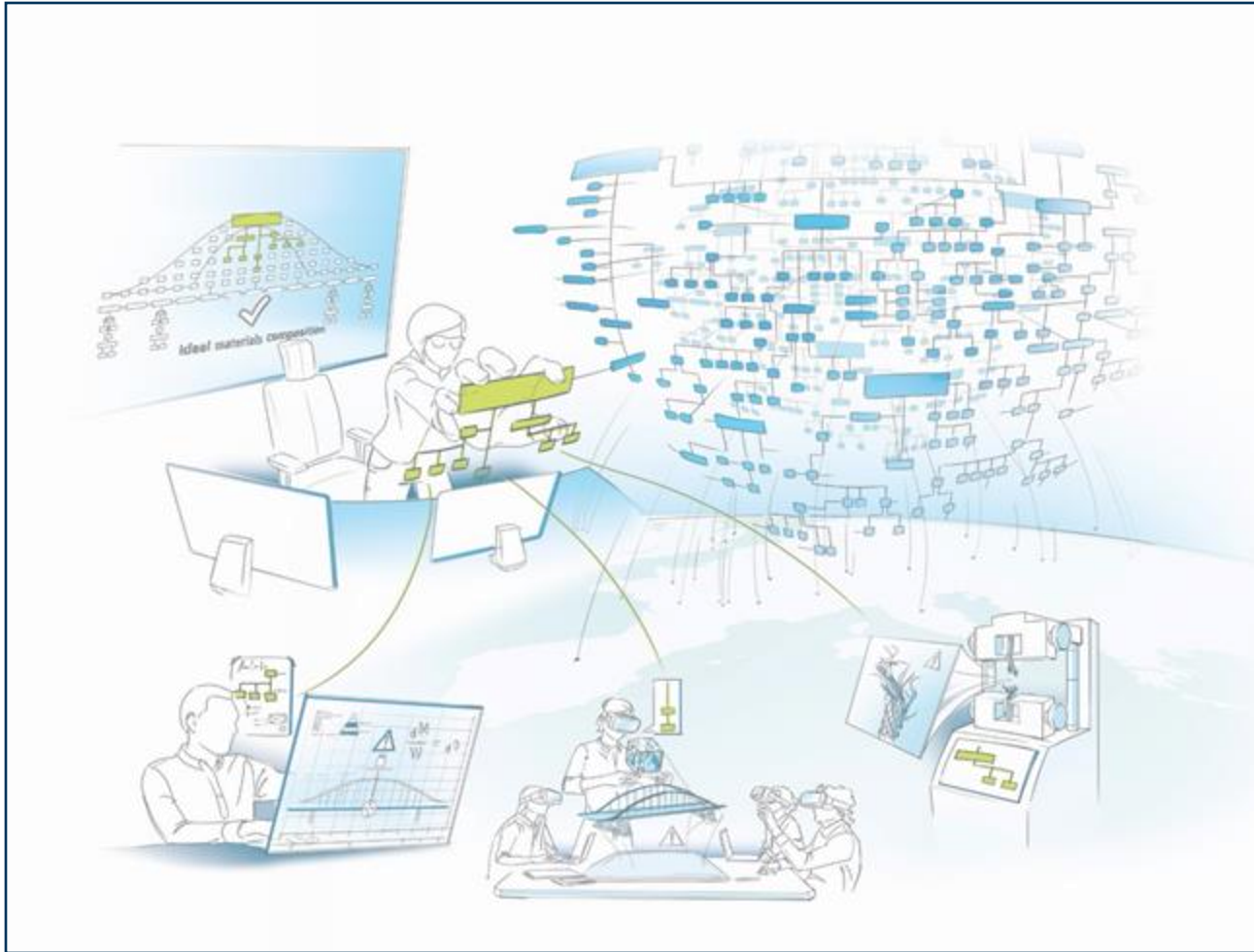
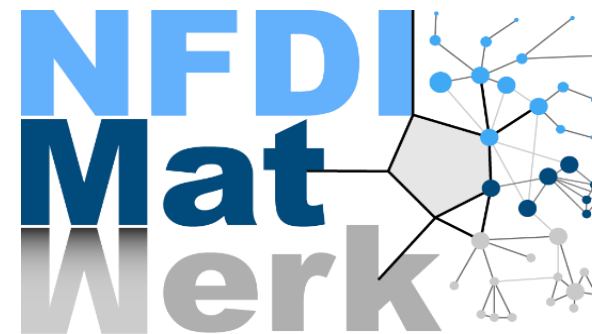


The journey towards Metadata Management



Rossella Aversa
Karlsruhe Institute of Technology (KIT)
Scientific Computing Center (SCC)

NATIONAL RESEARCH DATA
INFRASTRUCTURE FOR
MATERIALS SCIENCE &
ENGINEERING





- Understand the importance and added value of metadata
- Get an overview of JSON Schema
- Learn how to create, edit, and manage metadata
- Search for data from existing metadata

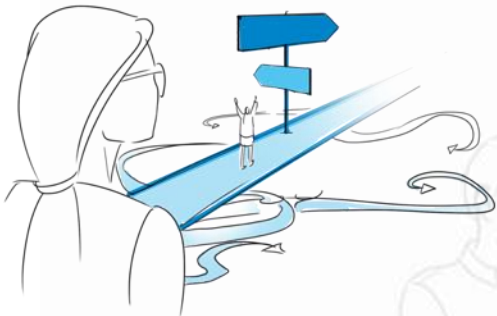
First part (Rossella):

- Motivation
- Recap of FAIR principles
- Basics of metadata
- Basics of JSON Schema and metadata documents

Second part (Sabrine):

- Metadata management in practice
 - Metadata Editor
 - NFDI-MatWerk Metadata Repository
- Survey

Feel free to ask questions!





The course is mainly targeting scientists or researchers, with a focus on the materials science community



Nevertheless, the tutorials are based on generic examples which are not related to any scientific topic, to be digested by a broader audience

You might want to:

- Compare your results with similar ones in the literature
- Reproduce/reuse results available in the literature
- Take delivery of the project handed over by a student/colleague who left
- Exchange data with your colleagues to collaborate on a research project
- Allow others to reproduce/reuse your results to be cited

Data-related research questions

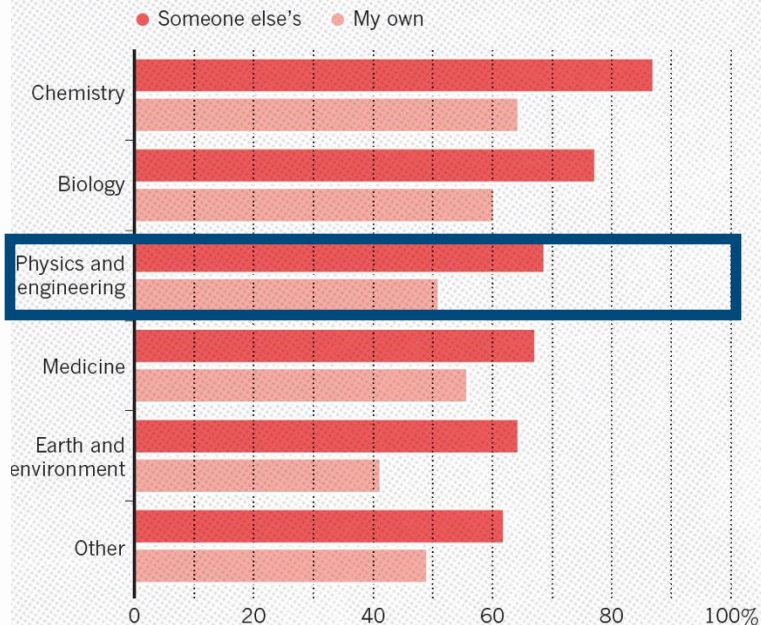
- Which data do support the results in this paper?
- How can I reproduce the data?
- How were these measures achieved/performed?
- How can I search for data with specific features?
- How can I publish my data in such a way that others can reuse it and cite it?



Reproducibility crisis

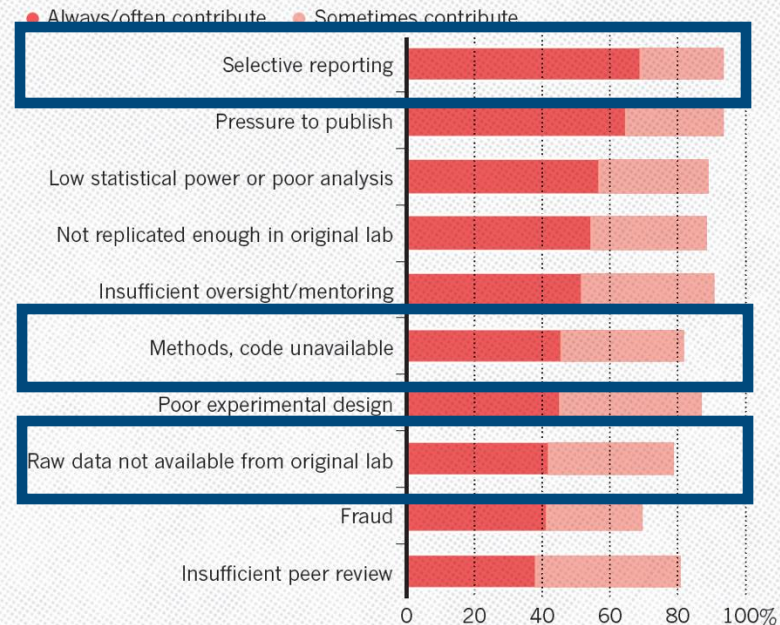
HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



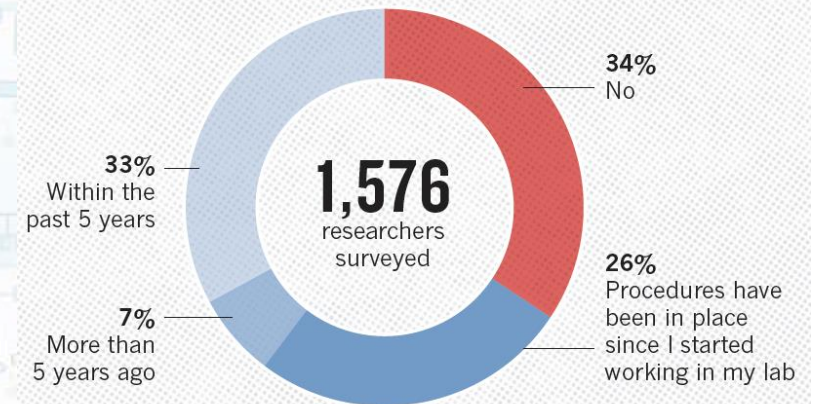
WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?

Among the most popular strategies was having different lab members redo experiments.



Baker (2016) <https://doi.org/10.1038/533452a>

The FAIR Guiding Principles



Findable



Accessible



Interoperable



Reusable

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



Findable

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



PIDs

- F1: (Meta)data are assigned globally unique and persistent identifiers (**PIDs**)
- F3: Metadata clearly and explicitly include the identifier of the data they describe



Accessible

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



Metadata
repositories

- A2: **Metadata should be accessible** even when the data is no longer available
- A1.2: The protocol allows for an **authentication and authorization (AAI)** procedure where necessary.



Interoperable

Data should be exchanged and interpreted by humans and computers



Structured
metadata

- I1: (Meta)data use a formal, accessible, shared, and broadly applicable **language** for knowledge representation
- I2: Metadata use **vocabularies** that follow the FAIR principles



Reusable

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



Standards,
Licence

- R1: Metadata should richly describe the data with a plurality of **accurate and relevant attributes**
- R1.1: (Meta)data are released with a clear and accessible data **usage licence**
- R1.3: Metadata meet domain-relevant **community standards or best practices**



Questions?

Basics of metadata: concepts

Data vs Metadata

Structured
Metadata

Metadata
Standards

Metadata
Repositories

PIDs

Licences

Metadata: data describing data.

I told you,
metadata is data!

It's so much more...



<https://imgflip.com/i/92po3j>

*“**Data** is stuff. It is raw, unprocessed, possibly even untouched by human hands, unviewed by human eyes, un-thought-about by human minds”.*

J. Pomerantz (2015). *Metadata*. The MIT Press.

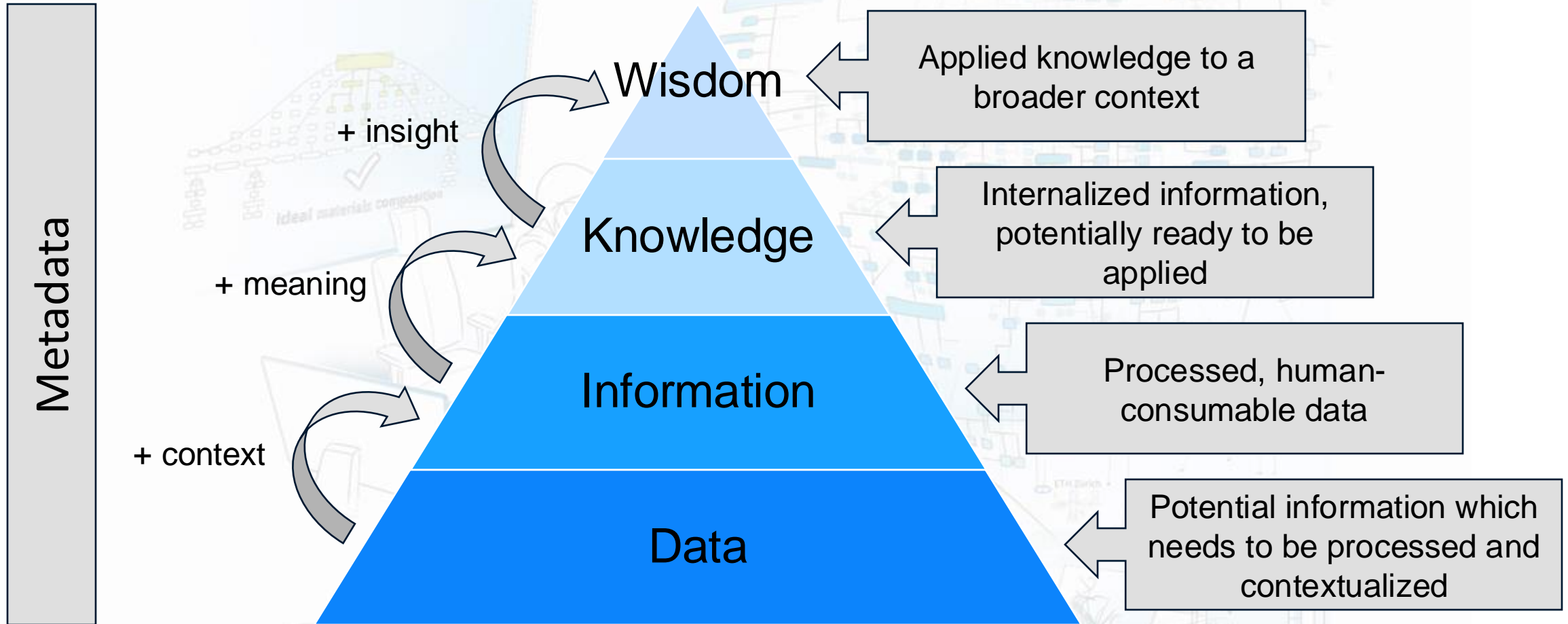
*“**Metadata** is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource”.*

National Information Standards Organization (2004) from “Big Data, Little Data, No Data”, C. L. Borgman (2015)

Data vs Metadata

	Data	Metadata
Nature and Content	Raw facts , measurements, observations	Information providing context and attributes
Usage	Analysis, decision making, research	Data management, discovery, interpretation
Representation	Mostly unstructured	Structured
Relationship	Independent and stand-alone	Linked to data/other metadata
Purpose and Function	Primary source of information	Supporting framework for organization, management, interpretation

Information pyramid

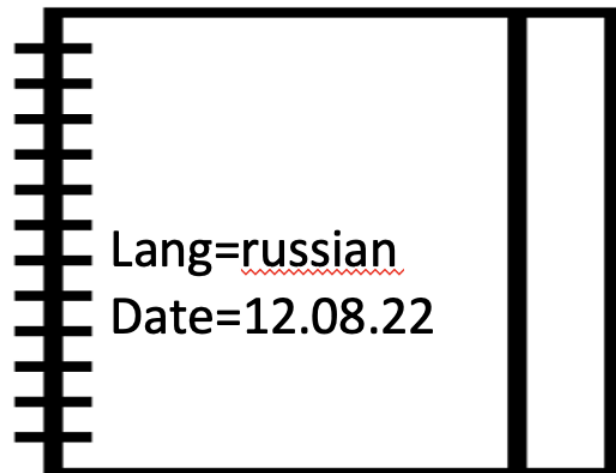


- **Administrative metadata:** facilitate the management of a resource (e.g. technical information regarding the file's creation and format, version, information about copyright, licence and intellectual property rights)
- **Descriptive metadata:** provides information about the intellectual content of a resource (e.g. title, author, date of publication, subject, description, unique identifier)
- **Scientific metadata:** enables discovery, identification, and reproducibility of resources. It usually includes the domain-specific attributes.

Electronic Lab Notebooks (ELNs)

Metadata should be understood by others, too!

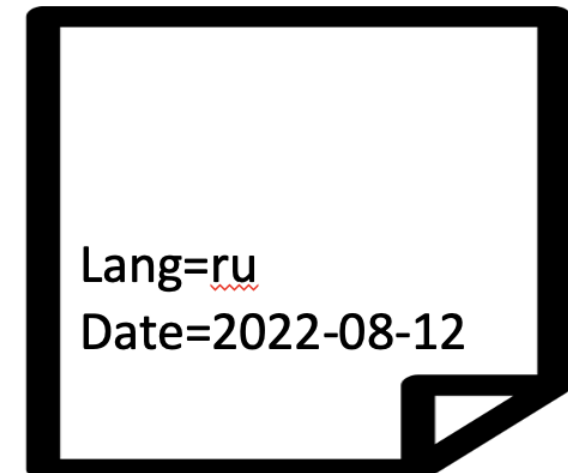
Researcher 1's notebook



Researcher 2's notebook



Researcher 3 using ISO standard



Basics of metadata: concepts

Data vs Metadata

Structured
Metadata

Metadata
Standards

Metadata
Repositories

PIDs

Licences



Format: the language for knowledge representation and exchange



Semantics: the (agreed) terminology to describe the attributes



Schema: the template which specifies the expected attributes and how they should be structured

- eXtensible Markup Language
- Main purpose: transfer and storage of arbitrary data on the World Wide Web
- Software- and hardware-independent
- Human- and machine-readable
- Hierarchical (tree-like) structure
- Elements are wrapped in start `<...>` and end `</...>` “tags”
- Doesn't have arrays
- Extensible: XML tags can be customized

```
<example>  
  <title>This is the example title</title>  
  <description>A simple XML example</description>  
  <wordCount>1</wordCount>  
</example>
```

```
<person>  
  <firstName>John</firstName>  
  <lastName>Doe</lastName>  
</person>
```

- JavaScript Object Notation
- Main purpose: transfer and storage of data
- Software- and hardware-independent
- Human- and machine-readable
- Hierarchical structure
- Elements are defined in key:value pairs
- Elements are wrapped as {objects} or [arrays]

```
{  
  "key":"value",  
  "aString":"string",  
  "anInteger":5,  
  "aFloat":0.5,  
  "aBoolean":true,  
  "anArray": ["item1", "item2", "item3"],  
  "anObject": {  
    "key1":"value1",  
    "key2":"value2",  
    "key3":"value3"  
  }  
}
```


- **Vocabulary:** set of terms pertaining to a particular domain + definitions. Useful to ensure that the data is described consistently
- **Taxonomy:** hierarchical structure of the terms. Useful to organize data into categories which are meaningful in a particular domain
- **Ontology:** formal description of the terms, their properties and their relationships within a particular domain. Useful to consistently represent the knowledge about a domain

- Template which specifies the expected elements and how they are structured:
 - Keys (vocabulary)
 - Value types
 - Rules
 - Mandatory/optional
- XML Schema Definition (XSD): less frequently used for modern standards
- JSON Schema: uses the JSON Schema Vocabulary <https://json-schema.org> to specify & syntactically validate the structure

Metadata Schema



Photo by M. Coghlan on Flickr (licence CC-BY-SA 2.0)

Metadata Document



Photo by M. Carrati on Unsplash

Metadata Schema

Template which specifies the expected elements and how they are structured

```
"givenName": {  
  "type": "string",  
  "description": "(Optional) - Given name of the user"  
},  
"familyName": {  
  "type": "string",  
  "description": "(Optional) - Family name of the user"  
},  
"age": {  
  "type": "number",  
  "description": "(Optional) - Age of the user"  
},
```

Metadata Document

An *instance* of a Metadata Schema which describes a given resource and conforms to the specified definitions

```
"givenName": "Rossella",  
"familyName": "Aversa",  
"age": 26
```

More on this later

- Same parameters for all data, more harmonized description
- Structured metadata can be more easily interpreted (also by machines)
- Results can be more easily reproduced/reused
- Data can be more easily compared/exchanged
- Data can be found based on their attributes (search, filter)
- Metadata can be validated
 - Note: schema validation only checks for **syntactical validity** (required property, corresponding value, and whether the value conforms with the expected data type)

A faint, light-colored illustration serves as the background. It depicts a person wearing a headscarf, looking towards a computer monitor. The monitor displays a bar chart with several vertical bars of varying heights and colors (blue, green, yellow). The overall style is minimalist and sketchy.

Questions?

Basics of metadata: concepts

Data vs Metadata

Structured
Metadata

Metadata
Standards

Metadata
Repositories

PIDs

Licences

- Metadata schemas which are well-established, endorsed, and widely accepted by the user community

General purpose

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

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

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

```
<xs:sequence>
  <xs:choice minOccurs="0" maxOccurs="unbounded">
    <xs:element ref="any"/>
  </xs:choice>
</xs:sequence>
```

```
<xs:element name="givenName" minOccurs="0"/>
<xs:element name="familyName" minOccurs="0"/>
```

```
"colleague": [
  "http://www.xyz.edu/students/alicejones.html",
  "http://www.xyz.edu/students/bobsmith.html"
],
```

Neutron, x-ray, muon

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

```
entry:NXentry
raw:NXsubentry
  definition="NXsas"
reduced:NXsubentry
  definition="NXcanSAS"
fluo:NXsubentry
  definition="NXfluo"
```


- Main purpose: unified description and enhanced identification of resources on the web
- Inspired to libraries



Powered by Bing Image Creator

Dublin Core Elements

Creator	Language	Relation
Contributor	Format	Source
Publisher	Subject	Type
Title	Description	Coverage
Date	Identifier	Rights

Formally standardized for cross-domain resource description

Metadata Standards Registries

FAIRsharing.org

Standards

A registry of terminology artefacts, models/formats, reporting guidelines, and identifier schemas.

Search through current res.

Registry: Standard Query string: dublin

Clear All

Displaying 1 to 15 of 15.

Dublin Core Metadata Element Set

DCES

The Dublin Metadata Element Set, which is often called Dublin Core (DC), is a standardized metadata scheme for description of any kind of resource such as documents in electronic and non-electronic form, digital...

Life Science Subject Ag... Resource ... Annotation Not applic... one more tag

- ▶ Linked Collections 7
- ▶ Linked Databases 72
- ▶ Linked Policies 2
- ▶ Linked Standards 66

RDA Metadata Standards Catalog

Metadata Standards Catalog

Search

Sign in

Dublin Core

A basic, domain-agnostic standard which can be easily understood and implemented, and as such is one of the best known and most widely used metadata standards.

Sponsored by the Dublin Core Metadata Initiative, Dublin Core was published as ISO Standard 15836 in February 2009.

Used in

Multidisciplinary

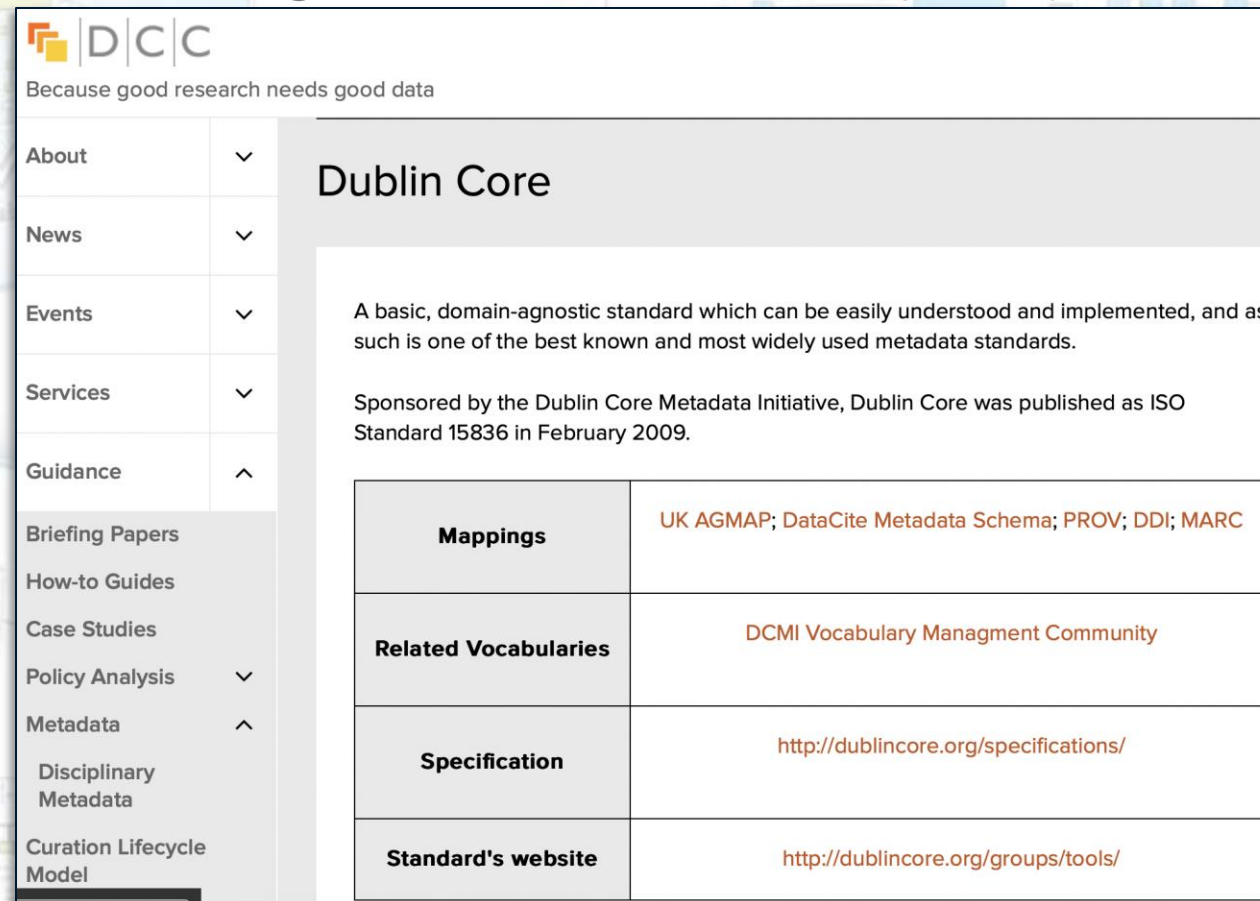
Documentation

[View specification](#)

[Visit website](#)

<https://rdamsc.bath.ac.uk>

Digital Curation Center (DCC)



DCC
Because good research needs good data

- About
- News
- Events
- Services
- Guidance
- Briefing Papers
- How-to Guides
- Case Studies
- Policy Analysis
- Metadata
- Disciplinary Metadata
- Curation Lifecycle Model

Dublin Core

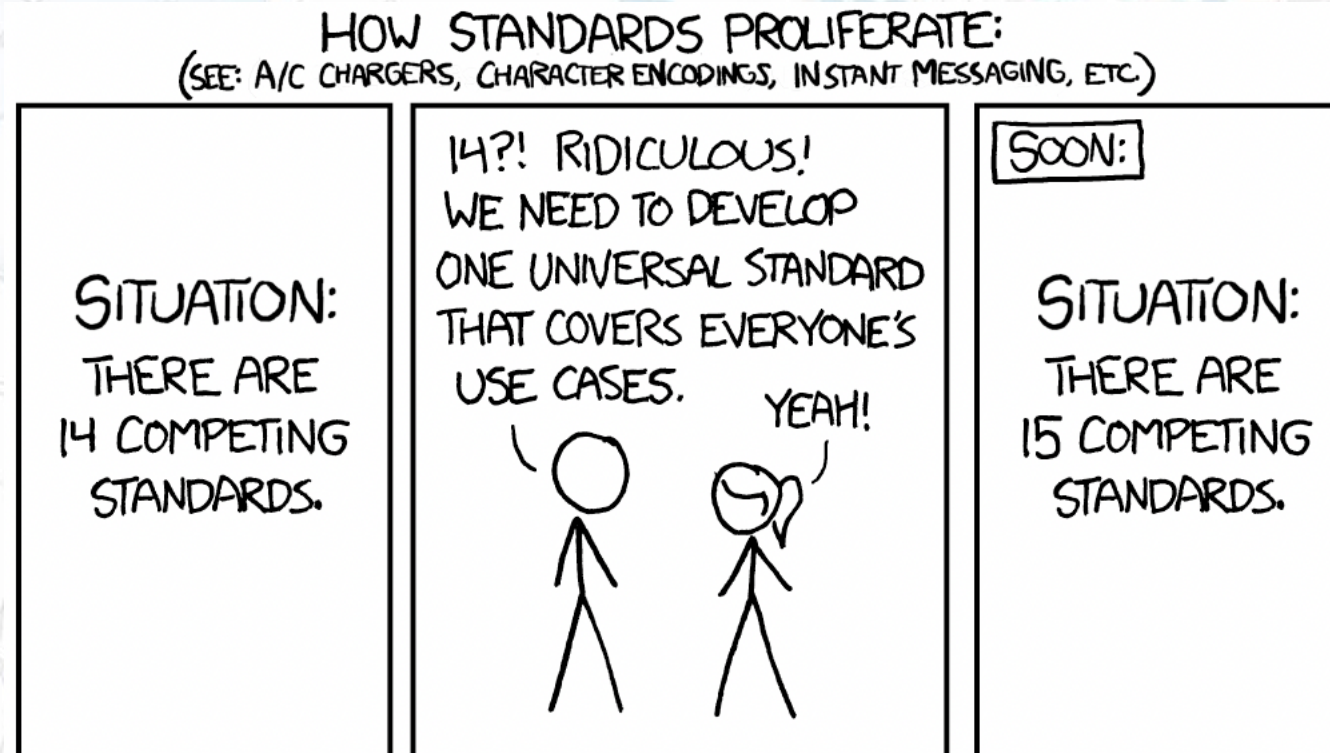
A basic, domain-agnostic standard which can be easily understood and implemented, and as such is one of the best known and most widely used metadata standards.

Sponsored by the Dublin Core Metadata Initiative, Dublin Core was published as ISO Standard 15836 in February 2009.

Mappings	UK AGMAP ; DataCite Metadata Schema ; PROV ; DDI ; MARC
Related Vocabularies	DCMI Vocabulary Management Community
Specification	http://dublincore.org/specifications/
Standard's website	http://dublincore.org/groups/tools/

<https://www.dcc.ac.uk/guidance/standards/metadata/list>

Before describing your data on your own, you should look for existing metadata schemas or standards



What if a standard does not exist?

Where do I publish my metadata?

Questions?

How do I link my data to metadata?

How can my (meta)data be reused and cited?



Data vs Metadata

Structured
Metadata

Metadata
Standards

Metadata
Repositories

PIDs

Licences

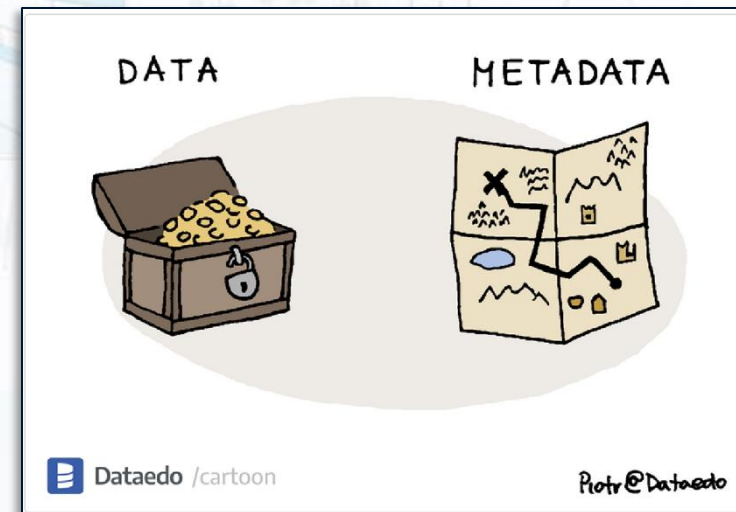
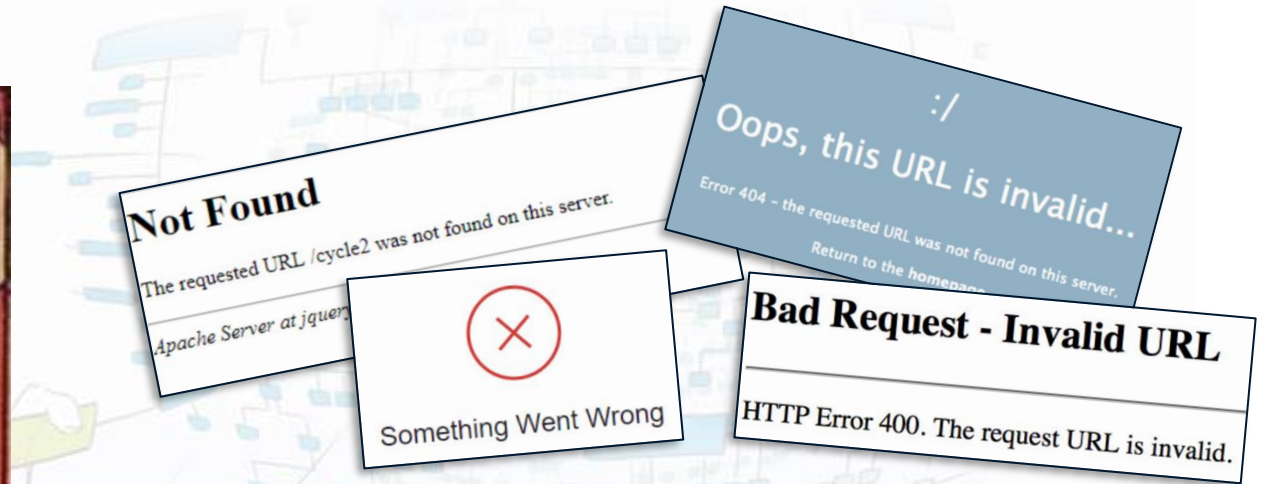
- *“Information system used to store, manage and provide access to Metadata, following a policy or a set of rules that define storage and access norms.”* Aversa R., et al. (2024) DOI: 10.5281/zenodo.10663833
- Register/find metadata schemas
- Register/find metadata documents
- Validate metadata documents against the schema
- Versioning
- Access control management
- User authentication

More on this with Sabrina

Why Metadata Repositories



<https://imgflip.com/i/92pnqc>



<https://dataedo.com/cartoon/data-vs-metadata-2>



Data vs Metadata

Structured
Metadata

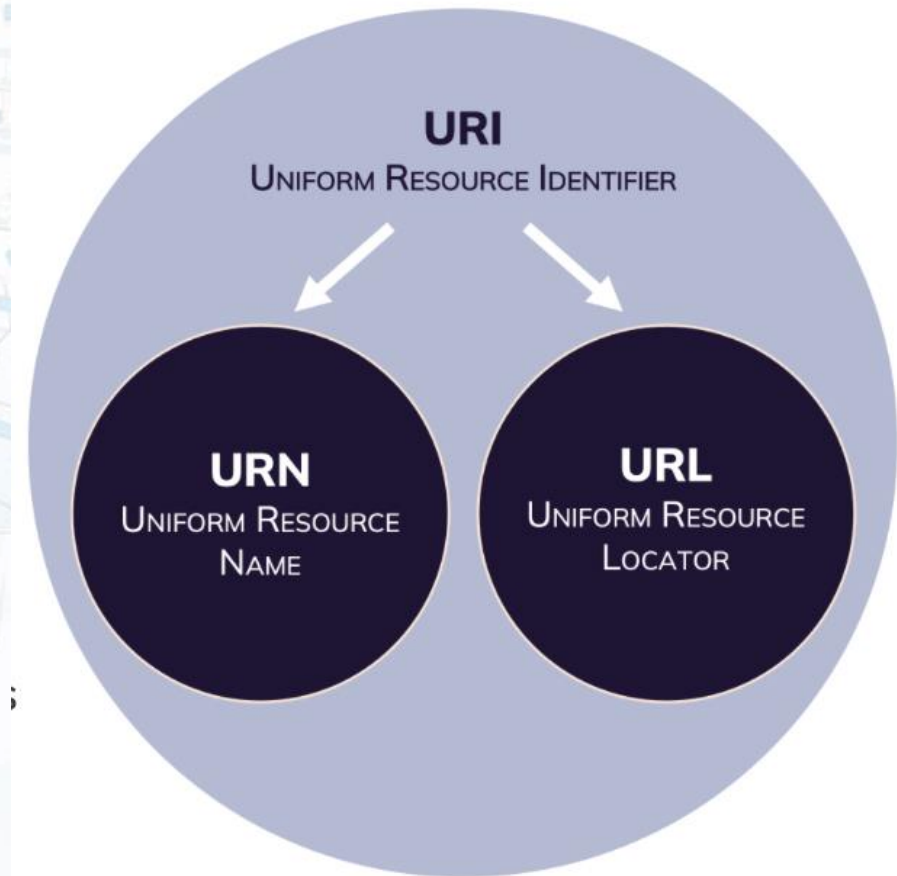
Metadata
Standards

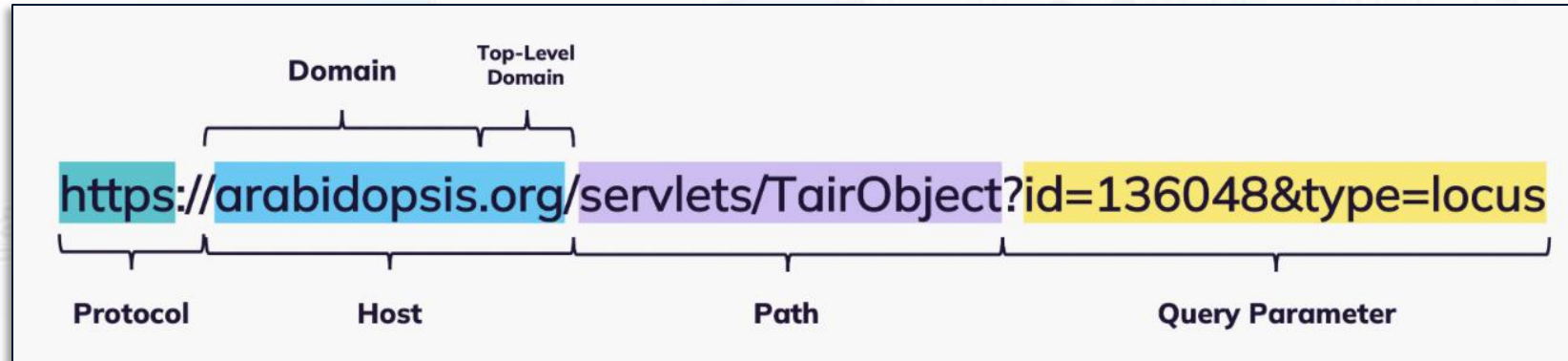
Metadata
Repositories

PIDs

Licences

- **Identifier:** any label used to name a resource – but not necessarily globally (e.g., personal name)
- **Unique identifier:** enables globally unique identification of a resource. The structure is standardized and centrally registered





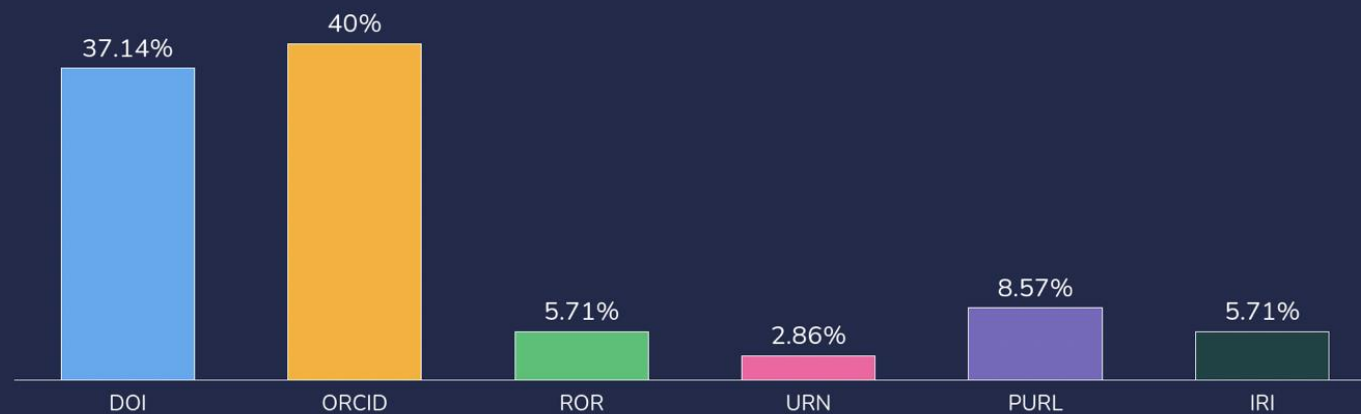
404 response to HTTP GET:

- the location was misspelled
- the file path on the host server changed
- the resource was deleted
- the resource was migrated to a different server

- **Persistent identifier:** long-lasting reference to locate and identify a resource, even if it changes over time
- Globally unique and persistent over time (until the PID provider is maintained!)
- Connected to a set of metadata describing a resource rather than to the resource itself
- **Benefit of PIDs:** allow different platforms to exchange information consistently and unambiguously, e.g. to track citations and reuse

Mint, manage and resolve PIDs (redirecting GET requests for digital resources to their latest URL)

Does it ring a bell? Which of these terms do you know?



- DOI: Digital Object Identifier
- ORCID: Open Researcher and Contributor ID
- ROR: Research Organization Registry
- URN: Uniform Resource Name
- PURL: Persistent Uniform Resource Locator
- IRI: Internationalized Resource Identifier

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>

Link a dataset
and an article

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

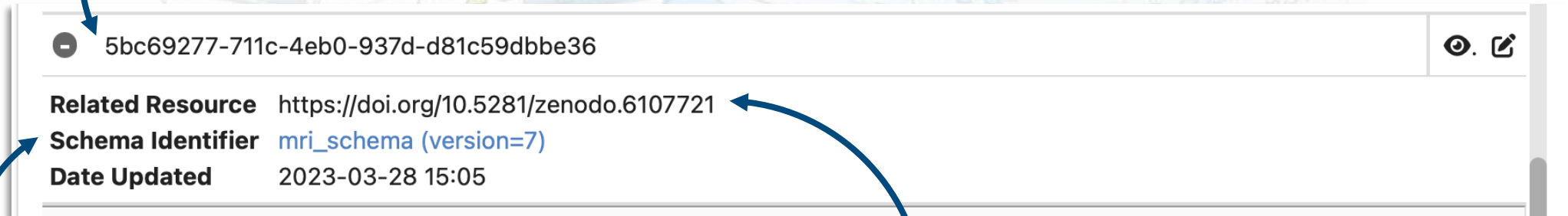
[Preview public record](#)

Link an author
and an article

PIDs can be used in metadata records as identifiers for associated resources

More on this with Sabine

Identifier of the metadata document



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

Identifier of the metadata schema

Identifier of the data

Take-home message

- PIDs are largely used to identify researchers, institutions, research articles, data resources, metadata records, code, software, ...
- Ready to publish?



Basics of metadata: concepts

Data vs Metadata

Structured
Metadata

Metadata
Standards

Metadata
Repositories

PIDs

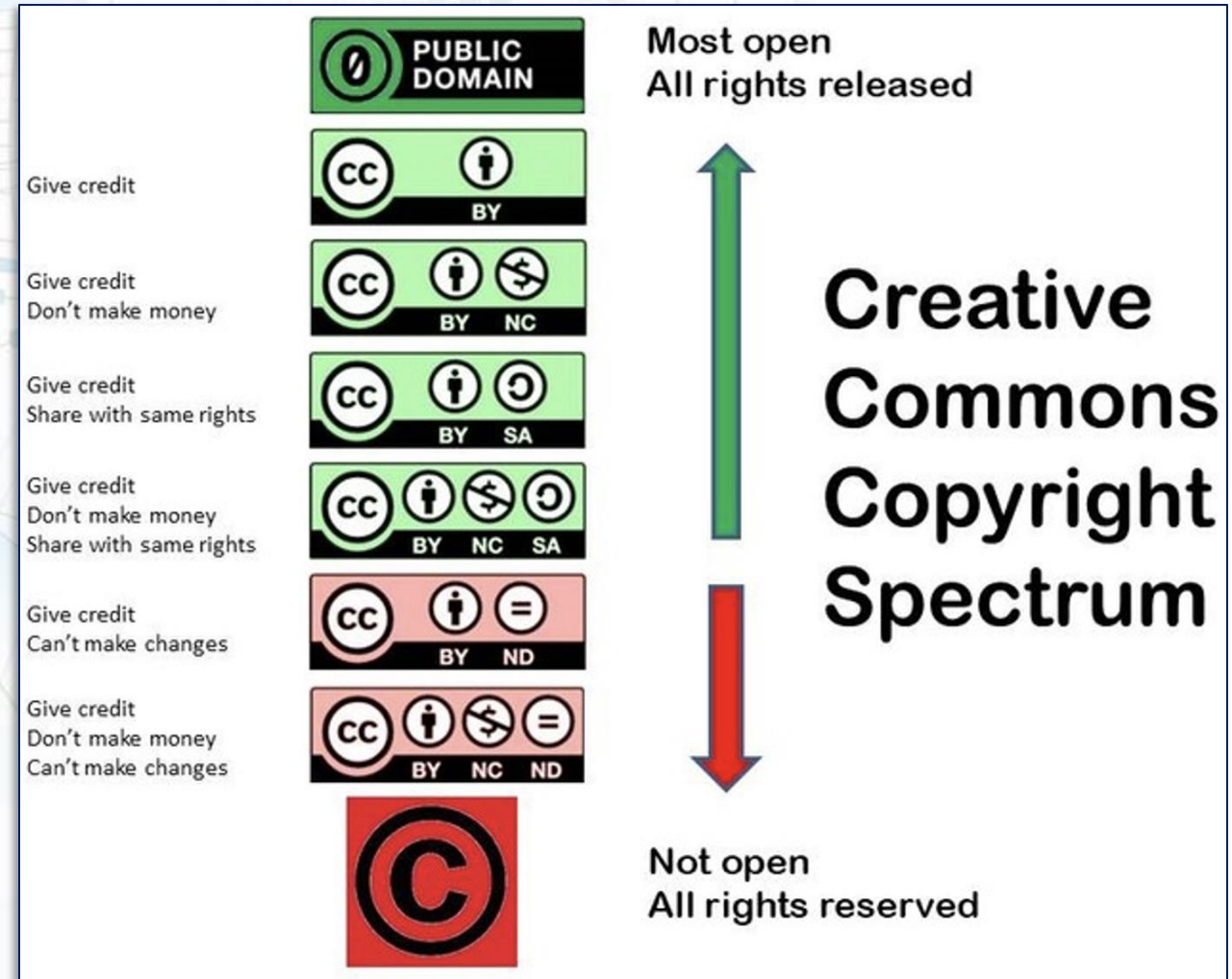
Licences

Legal arrangement between the creator of the data and the end-user, or the place where the data will be deposited, specifying what users can do with the data

Rights



Creative Commons Attribution 4.0 International



- FAIR Principles are guidelines for (meta)data management
- Structured metadata are helpful for better data interpretation, exchange, reuse
- Adoption of existing standards or community best practices avoids proliferation of descriptions and eases data sharing
- Persistent and globally unique identifiers allows you to link resources and cite them
- Licences are keys when you publish or reuse results

Let's put it in practice!

An illustration in a light blue and grey line-art style. On the left, two people are shown from the side, looking at a laptop. The person in the foreground is wearing glasses and has their hand near their chin. In the background, there is a large, complex network diagram with many nodes and connecting lines. To the right of the network diagram is a large rectangular screen or board. In the foreground, there are two laptops, one of which is being held by the person in the foreground. The overall scene suggests a collaborative work environment focused on technology or data.

Demo on JSON Schema

User description

A JSON Schema is an object!

key:value pairs,
keys are always "strings"

```
{  
  "$schema": "https://json-schema.org/draft/2019-09/schema",  
  "title": "user_description",  
  "description": "Basic schema for user description. It can be extended.",  
  "type": "object",  
  "required": [...],  
  "properties": {...}  
}
```

Entities must be separated by commas!

Not sensitive to indentations and line breaks
(but they help readability...)

User description

Schema keyword

Schema of the schema (meta-schema): JSON Schema standard. Version identifier, points to the location of the schema specification. The value must be an URI

Schema annotations

```
{
  "$schema": "https://json-schema.org/draft/2019-09/schema",
  "title": "user_description",
  "description": "Basic schema for user description. It can be extended.",
  "type": "object",
  "required": [...],
  "properties": {...}
}
```

Validation keywords

Case sensitive!

Properties: strings and numbers

Object where each property represents a key that is validated

The name of each property is arbitrary (existing vocabularies!)
The type must be specified

```
"properties": {
  "userName": {
    "type": "string",
    "description": "(Required) - Full name of the user in the format (Family Name, Given Name)"
  },
  "givenName": {
    "type": "string",
    "description": "(Optional) - Given name of the user"
  },
  "familyName": {
    "type": "string",
    "description": "(Optional) - Family name of the user"
  },
  "age": {
    "type": "number",
    "description": "(Optional) - Age of the user"
  },
}
```


Properties: enumerations



```
"role": {  
  "type": "string",  
  "description": "(Required) - Role of the user",  
  "enum": [  
    "Data Curator",  
    "Instrument Scientist",  
    "Team Leader",  
    "Team Member"  
  ]  
},
```

“enum” can be used even without a type, to accept values of different types

Validation keyword used to restrict a value to a fixed set of values. It must be an array with at least one element, where each element is unique

Properties: arrays

Used for ordered elements.
List validation: a sequence of
arbitrary length where each item
matches the same schema

```
"example": {  
  "type": "array",  
  "items": {  
    "type": "string"  
  }  
},
```

```
"example": ["array", "of", "strings"]
```

```
"contact": {  
  "type": "array",  
  "items": {  
    "oneOf": [  
      {  
        "title": "Email",  
        "type": "object",  
        "properties": {  
          "email": {  
            "type": "string",  
            "description": "(Optional) - Email Address"  
          }  
        }  
      },  
      {  
        "title": "ORCID",  
        "type": "object",  
        "properties": {  
          "orcid": {  
            "type": "string",  
            "description": "(Optional) - ORCID URL",  
            "pattern": "^https://orcid\\.org/[0-9]{4}-[0-9]{4}-[0-9]{4}-[0-9]{3}[X0-9]{1}$"  
          }  
        }  
      }  
    ]  
  }  
}
```

Keyword for combining subschemas together. It corresponds to the boolean algebra concept XOR (must be valid against *exactly one* of the subschemas)

Each item of the array is a subschema

keyword used to restrict a string to a regular expression

Required properties

By default, the properties defined by the “properties” keyword are not required.

However, one can provide a list of required properties using the “required” keyword, which takes an array of (unique) strings:

```
"required": [  
  "userName",  
  "role"  
],
```


An illustration in a light blue and grey line-art style. On the left, two people are shown from the side, looking at a laptop. The person in the foreground is wearing glasses and has their hand on their chin in a thoughtful pose. The person behind them is also looking at the laptop. To the right, another laptop is open on a desk. In the background, a large, complex network diagram with many nodes and connecting lines is visible. The overall scene suggests a collaborative technical or data-related activity.

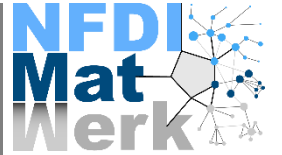
**Hands-on:
write a valid metadata document**

1. Download the metadata schema from: t1p.de/user_schema
2. Open it with a text editor (TextEdit, Visual Studio Code, ...)
3. Open a new empty file
4. Element by element, write the metadata document following the metadata schema
5. Name it as you like, with extension **.json** and save it locally

- This was a very simple metadata schema...
- Writing a valid JSON metadata document for every dataset, measurement, or analysis is tedious and time consuming
- It would be helpful to have the schema as a **form**
 - To fill it out
 - To validate it

More on this with Sabine

Acknowledgements



Contributions:

Thomas Jejkal, Andrea Recchia

Used material:

Gerlich, S., Strupp, A., Hofmann, V., Sandfeld, S. (2023). Fundamentals of Scientific Metadata. The Carpentries Incubator. DOI: 10.5281/zenodo.10091708

Founded by:

the Joint Laboratory Model and Data driven Materials Characterization (JL MDMC), a cross-centre platform of the Helmholtz Association; the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project; the research program "Engineering Digital Futures" of the Helmholtz Association of German Research Centers; the Helmholtz Metadata Collaboration Platform.

Funded by



Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 38/1 – project number 460247524

Illustrations by:
© Fraunhofer IWM, Illustrations: Gebhard|Uhl Freiburg