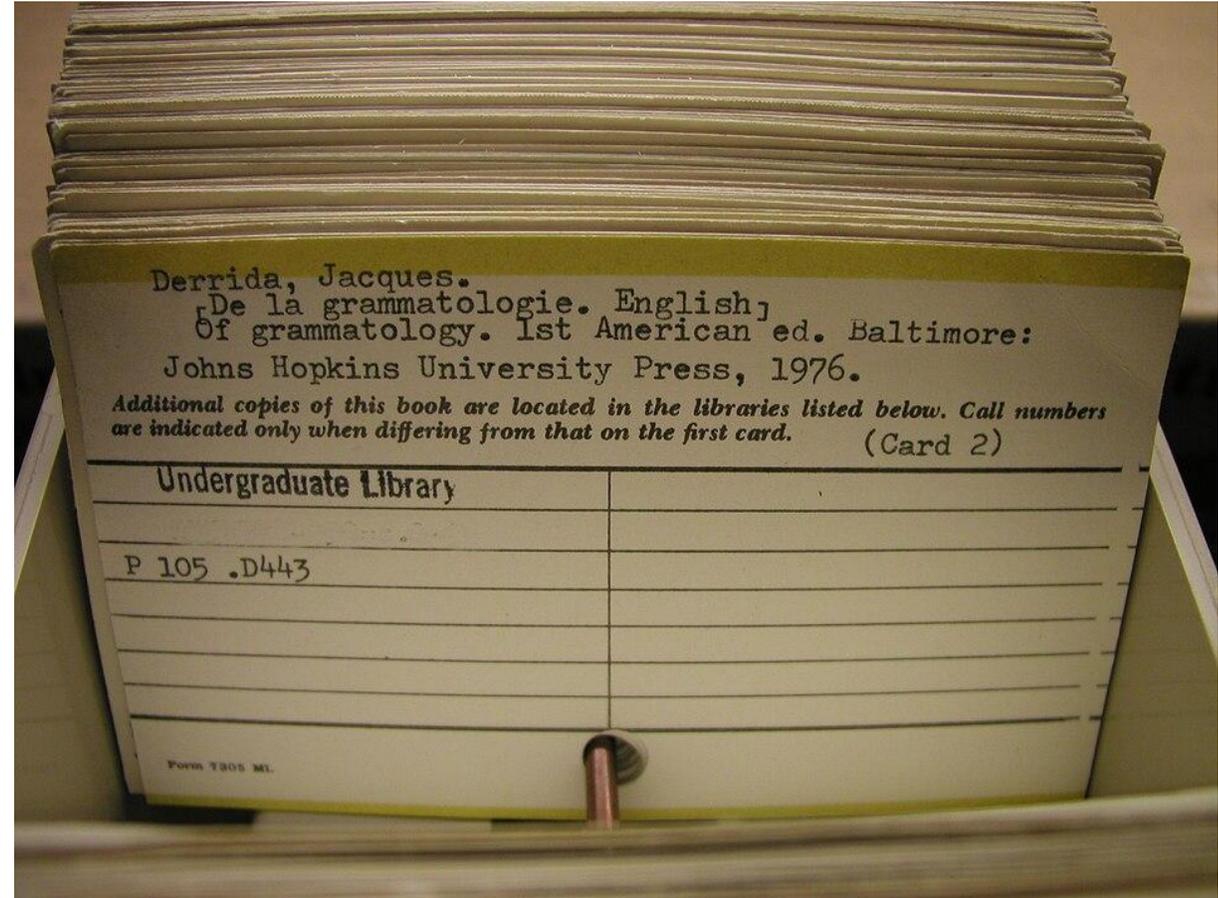




What is metadata?

- Definition(s)
 - Data about Data
 - Structured information that facilitate retrieval, use or management of some information resource
- Everyday examples

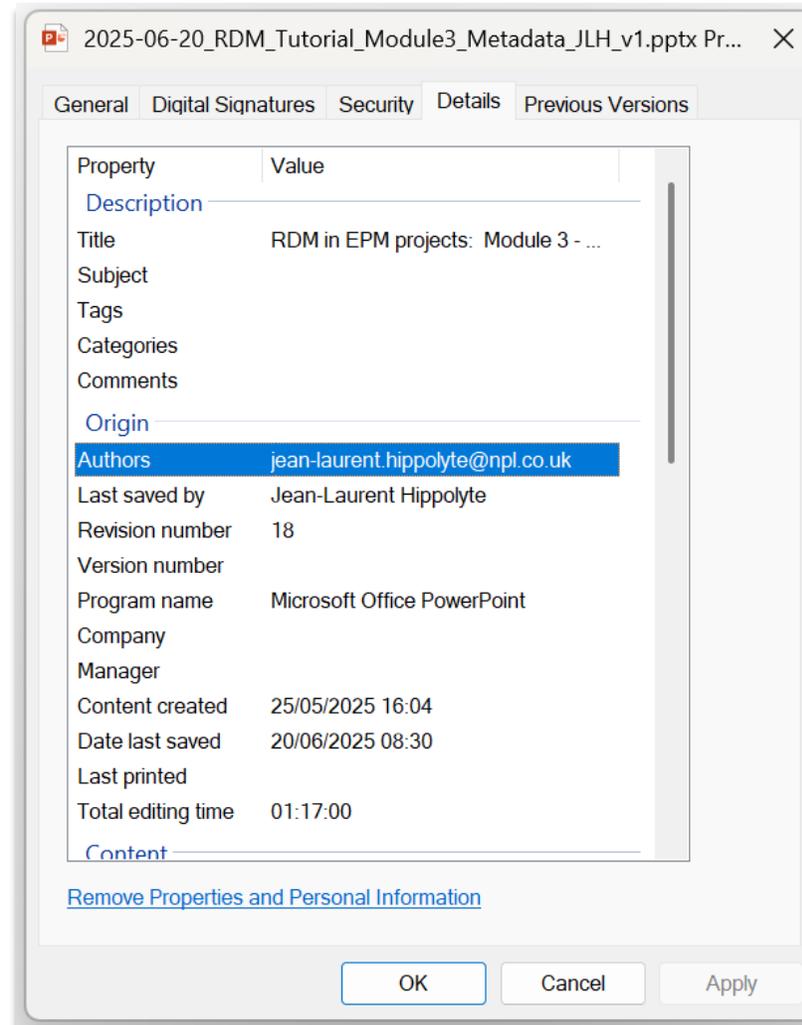


[Photo](#) by David Fulmer / [CC BY](#)



What is metadata?

- Definition(s)
 - Data about Data
 - Structured information that facilitate retrieval, use or management of some information resource
- Everyday examples
 - File properties in Operating Systems





What is metadata?

- Definition(s)
 - Data about Data
 - Structured information that facilitate retrieval, use or management of some information resource
- Everyday examples
 - File properties in Operating Systems
 - Google Knowledge Graph

The screenshot shows a Google search for 'alan turing'. At the top, there are navigation tabs for Images, News, Videos, Short videos, Shopping, Forums, and More. Below the search bar, the 'Alan Turing' Knowledge Graph card is displayed. It includes a portrait of Turing, his birth and death dates (23 Jun 1912, Maida Vale, London; 7 Jun 1954, Wilmslow), and a snippet from Wikipedia: 'Alan Mathison Turing was an English mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist.' The 'About' section of the Knowledge Graph is highlighted with a blue border and contains the following information: 'About Alan Mathison Turing was an English mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Wikipedia Born: 23 June 1912, Maida Vale, London Died: 7 June 1954 (age 41 years), Wilmslow Movies: The Man Who Cracked the Nazi Code: The Story of Alan Turing Education: Princeton University (1936–1938) · See more Influenced by: Kurt Gödel, Ludwig Wittgenstein, Alonzo Church Siblings: John Turing Feedback People also search for Charles Joan Clarke Tommy Herman'. A blue arrow points from the text 'Google Knowledge Graph' in the list to the Knowledge Graph card.



What is metadata?

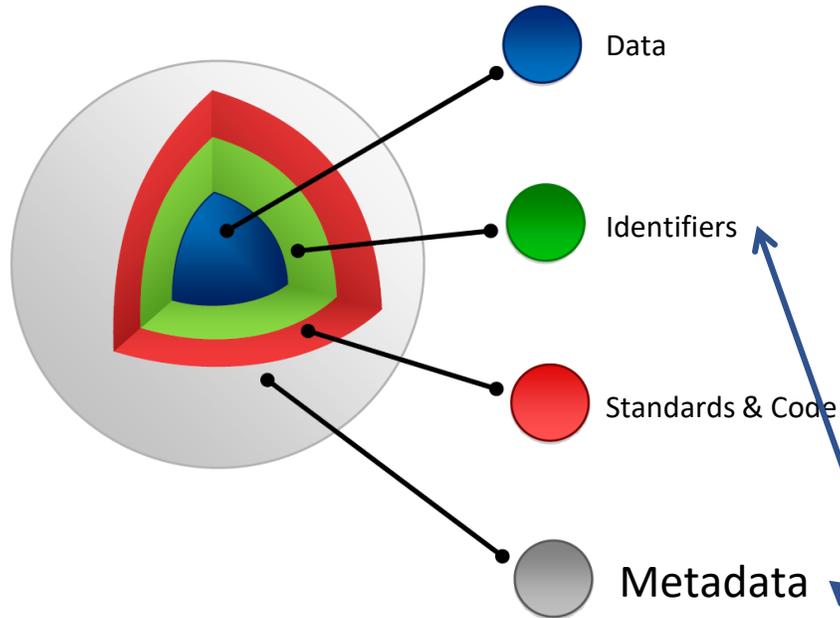


- Gaps in current practices
 - Ad-hoc data organisation
 - file/folder naming conventions
 - Unstandardised description
 - headers in spreadsheets
 - Knowledge embedded in human
 - Understanding of data lost due to employee turnover
 - Data unusable by third parties



How is it useful to RDM?

- Realization of FAIR relies on metadata
 - Findable, Accessible, Interoperable, Reusable



Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. **Machine-readable metadata are essential** for automatic discovery of datasets and services.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

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

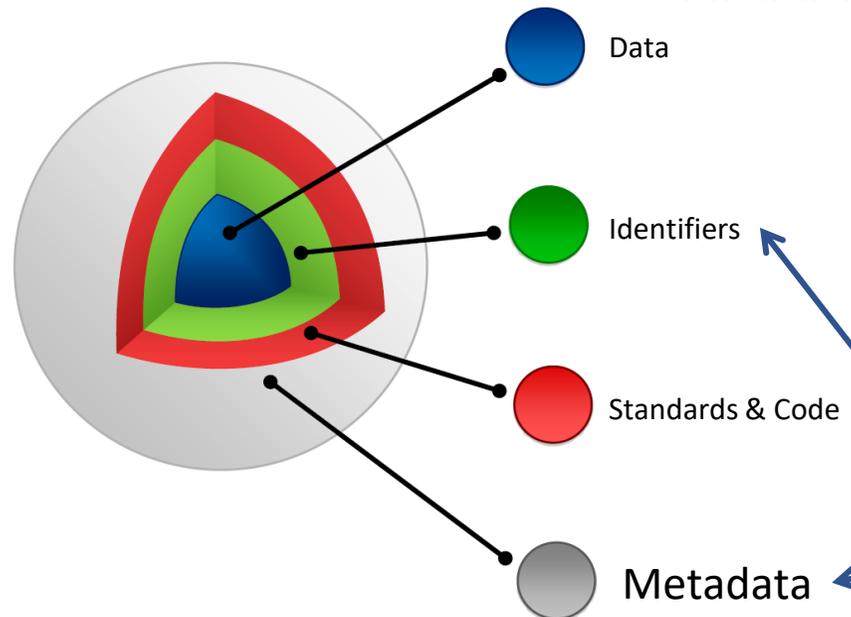
F4. (Meta)data are registered or indexed in a searchable resource

Hodson, Simon et al. 2018. FAIR Data Action Plan: Interim recommendations and actions from the European Commission Expert Group on FAIR data. (Jun. 2018). <http://doi.org/10.5281/zenodo.1285290>



How is it useful to RDM?

- Realization of FAIR relies on metadata
 - Findable, Accessible, Interoperable, Reusable



Accessible

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

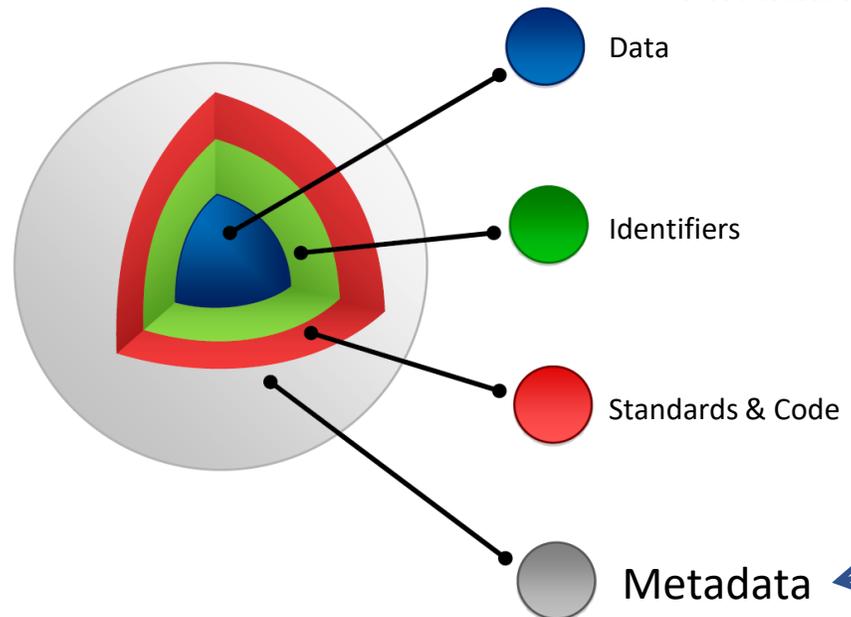
A2. Metadata are accessible, even when the data are no longer available

Hodson, Simon et al. 2018. FAIR Data Action Plan: Interim recommendations and actions from the European Commission Expert Group on FAIR data. (Jun. 2018). <http://doi.org/10.5281/zenodo.1285290>



How is it useful to RDM?

- Realization of FAIR relies on metadata
 - Findable, Accessible, Interoperable, Reusable



Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

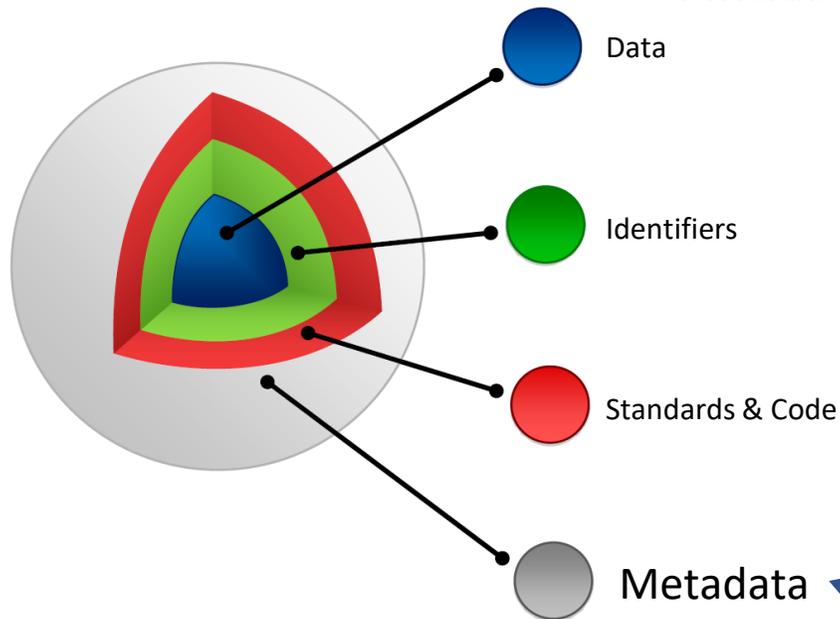
1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
2. (Meta)data use vocabularies that follow FAIR principles
3. (Meta)data include qualified references to other (meta)data

Hodson, Simon et al. 2018. FAIR Data Action Plan: Interim recommendations and actions from the European Commission Expert Group on FAIR data. (Jun. 2018). <http://doi.org/10.5281/zenodo.1285290>



How is it useful to RDM?

- Realization of FAIR relies on metadata
 - Findable, Accessible, Interoperable, Reusable



Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

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

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

Hodson, Simon et al. 2018. FAIR Data Action Plan: Interim recommendations and actions from the European Commission Expert Group on FAIR data. (Jun. 2018). <http://doi.org/10.5281/zenodo.1285290>



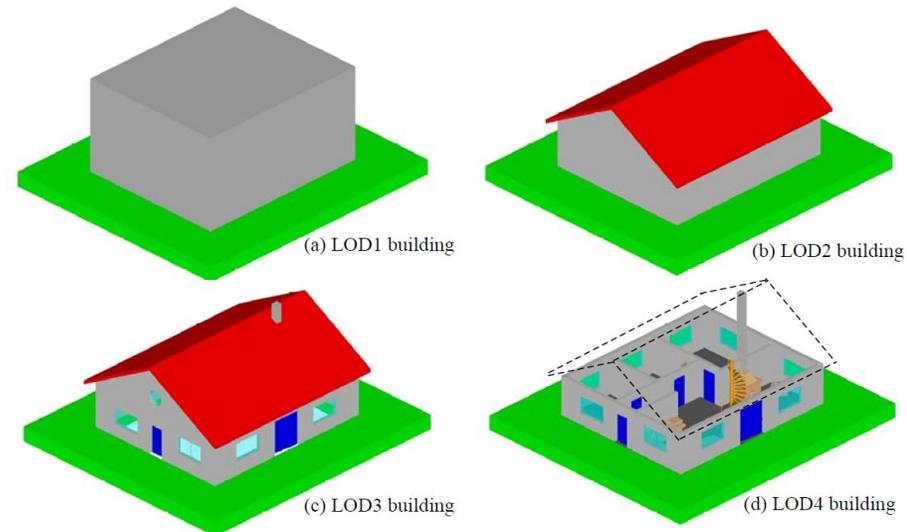
How is it useful to RDM?

- EURAMET reporting guide
 - Mandates that
 - Data used for research validation is accompanied by metadata
 - *“rich information on the datasets (description, date of deposit, author(s), venue and embargo); the EMPIR funding; the project acronym and number; licensing terms; persistent identifiers for the dataset, the authors involved, and, if possible, their organisations and the project.”*
 - Recommends that
 - Datasets are shared via open access repositories, therefore made **searchable through metadata**
 - Metadata to comply with standard vocabularies or schemas, particularly metadata standards **specific** to the targeted discipline (if they exist...)



How is it useful to RDM?

- Leveraging the FAIR principles
 - Data quality
 - Provenance
 - Reproducibility
 - Transparency
 - Trustworthiness
- The more comprehensive the metadata, the more value added to data
- Level of details also depend on intended/potential reuse



<http://www.ogc.org>



How is it useful to RDM?

- Many desirable aspects of data quality can't be achieved without metadata:
 - believability, objectivity, reputation, relevancy, interpretability...
 - Mathmet Data Quality Assurance planning tool

<https://www.euramet.org/european-metrology-networks/mathmet/activities/quality-assurance-tools>

- Available as an online interactive form on the MetRDMO platform

<https://dmp.metrology-rdm.eu/>

The screenshot shows the 'Create new project' form on the MetRDMO platform. The form has a blue header with the PTB logo and navigation links for 'Management', 'Feedback', and 'Language'. The main content area is titled 'Create new project' and includes fields for 'Title', 'Description', and 'Catalog'. The 'Catalog' section has two radio button options: 'Mathmet Software Quality Assurance Plan: Version 0.2.0' (selected) and 'Mathmet Data Quality Assurance Plan'. A sidebar on the right, titled 'Quality assurance', contains three questions: 'Q25 How will data quality be documented and made available for review?', 'Q26 How will the project team review data quality issues? *', and 'Q27 How will the customer or customer proxy review data quality issues? *'. Each question has a corresponding text input field.



Specifying metadata requirements

- Metadata requirements often formally described.
- Example: metadata for scientific papers
 - A BibTeX entry includes mandatory and optional tags which characterize a bibliographic reference (author, title, year, etc.)
 - Multiplicity of tags allows cross-checking of the reference

[1] P. J. Cohen. The independence of the continuum hypothesis. *Proceedings of the National Academy of Sciences*, 50(6):1143–1148, 1963.

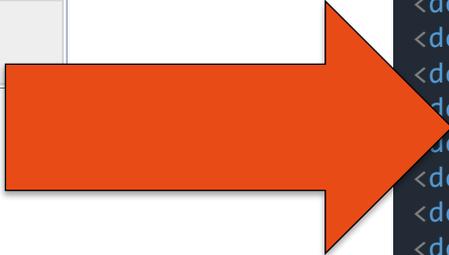
```
@article{CitekeyArticle,  
  author   = "P. J. Cohen",  
  title    = "The independence of the continuum hypothesis",  
  journal  = "Proceedings of the National Academy of Sciences",  
  year     = 1963,  
  volume  = "50",  
  number  = "6",  
  pages   = "1143--1148",  
}
```



Specifying metadata requirements

Web form

Title?	[+][-]
Creator?	[+][-]
Subject?	[+][-]
Description?	[+][-]
Publisher?	[+][-]
Contributor?	[+][-]
Date?	[+][-]
Type?	[+][-]
Format?	[+][-]
Identifier?	[+][-]
Source?	[+][-]
Language?	[+][-]
Relation?	[+][-]
Coverage?	[+][-]
Rights?	[+][-]



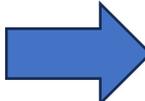
Machine-interpretable metadata

```
<?xml version="1.0" encoding="UTF-8"?>
<dc:title>Metadata for RDM and publications submission for EMPIR projects</dc:title>
<dc:creator>Jean-Laurent Hippolyte</dc:creator>
<dc:creator>Julia Neumann</dc:creator>
<dc:subject>Metadata</dc:subject>
<dc:subject>Research Data</dc:subject>
<dc:description>Brief overview of metadata for scientific datasets</dc:description>
<dc:publisher>EURAMET TC-IM 1449</dc:publisher>
<dc:date>11/03/2021</dc:date>
<dc:type>Presentation</dc:type>
<dc:format>Microsoft PowerPoint</dc:format>
<dc:source>https://www.euramet.org/</dc:source>
<dc:language>en</dc:language>
<dc:rights>https://creativecommons.org/licenses/by/4.0/</dc:rights>
```

Generated using an online generator:

https://nsteffel.github.io/dublin_core_generator

- Metadata capture as part of the experimental process log

 Following talk



Specifying metadata requirements

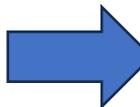
- Example of domain-specific metadata schema(s):
 - Open Biological and Biomedical Ontology (OBO) Foundry
 - Metadata concept search engine (OntoBee)
- SI and CIPM MRA
 - SI Digital Framework

The screenshot shows the OntoBee search engine interface. At the top, there is a navigation bar with links for Home, Intro, Statistics, SPARQL, Ontobee, Annotator, Tutorial, and FAQs. Below the navigation bar, there is a search input field with the text "Please select an ontology (optional)" and a dropdown arrow. The "Keywords:" field contains the text "staining". To the right of the keywords field are buttons for "Search terms" and "Batch Search". Below the search input, there is a section titled "Terms with 'staining' included in their label:" followed by a list of three results:

1. http://purl.obolibrary.org/obo/OBI_0302887 (OBI):
 - staining in Ontobee: OBI, BCGO, CIDO, ECO, ERO, ICO, OBIB
2. http://purl.obolibrary.org/obo/IDOMAL_0000551 (IDOMAL):
 - staining in Ontobee: IDOMAL
3. http://purl.obolibrary.org/obo/NCIT_C50753 (NCIT):

Below the list, there is a detailed view for the class "staining". The "Class: staining" section includes the "Term IRI" (http://purl.obolibrary.org/obo/OBI_0302887) and the "Definition": "Staining is a process which results in the addition of a class-specific (DNA, proteins, lipids, carbohydrates) dye to a substrate to qualify or quantify the presence of a specific compound." The "Annotations" section lists several properties: "definition editor: Philippe Rocca-Serra", "definition source: adapted from Wikipedia: <http://en.wikipedia.org/wiki/Staining>", "example of usage: PMID: 18540298. Role of modified bleach method in staining of acid-fast bacilli in lymph node aspirates. Acta Cytol. 2008 May-Jun;52(3):325-8.", and "has curation status: pending final vetting". The "Class Hierarchy" section shows the following structure:

```
Thing
+ entity
+ occurrent
+ process
+ planned_process
+ material_processing
+ sample_preparation_for_assay
+ transplantation
+ cell_co-culturing
+ enzymatic_cleavage
+ artificially_induced_nucleic_acid_hybridization
- histological_sample_preparation
- ionize_process
- cell_cycle_synchronization
- manufacturing
+ material_combination
+ library_preparation
- vaccine_preparation
- cross_linking
- denaturing
```

 Following talk



Making datasets accessible

- **Zenodo** an open-access repository hosted by CERN
- Attempts to comply with FAIR principles as best as possible
- But some compliance aspects are up to the users!

Zenodo provides online tools to:

- assign and resolve dataset persistent identifiers (DOIs)
- generate basic metadata
- search datasets through cross-domain metadata

<https://about.zenodo.org/principles/>

The screenshot shows the Zenodo website interface. At the top, the Zenodo logo is on the left, a search bar containing 'metrology' is in the center, and 'Upload' and 'Communities' links are on the right. Below the search bar, it says 'All versions' and 'Found 5 results.' with a pagination control showing '1'. The first result is for a dataset titled 'December 1, 2020 (v.1)' with 'Open Access' status. The title is 'Road siting experiment' and the authors are Graziano Coppa, Chiara Musacchio, and Andrea Merlone. The description states: 'This dataset comes from an experiment performed at the Italian National Metrology Institute (INRiI) MeteoMet 2. The purpose of the experiment was to evaluate the effect of the presence of a road in dataset is composed by'. It was uploaded on December 1, 2020. The second result is for a dataset titled 'April 22, 2020 (0.2)' with 'Open Access' status. The title is 'MPU9250 MEMS IMU Sine wave acceleration excitation along the Z axis' and the author is Seeger Benedikt.



Making datasets accessible

- **DataCite** a not-for-profit organization
- Aims to improve data citation for :
 - accessible research data
 - transparent and reproducible research
- Datacite provides online tools to:
 - assign and resolve dataset persistent identifiers (DOIs)
 - generate metadata
 - search datasets through cross-domain metadata



Find what you're looking for by searching millions of records with extensive, reliable metadata.



Share your data and reuse the data of others to create the highest impact in the research community.



Cite your research sources with confidence, and receive proper credit when your work is reused.



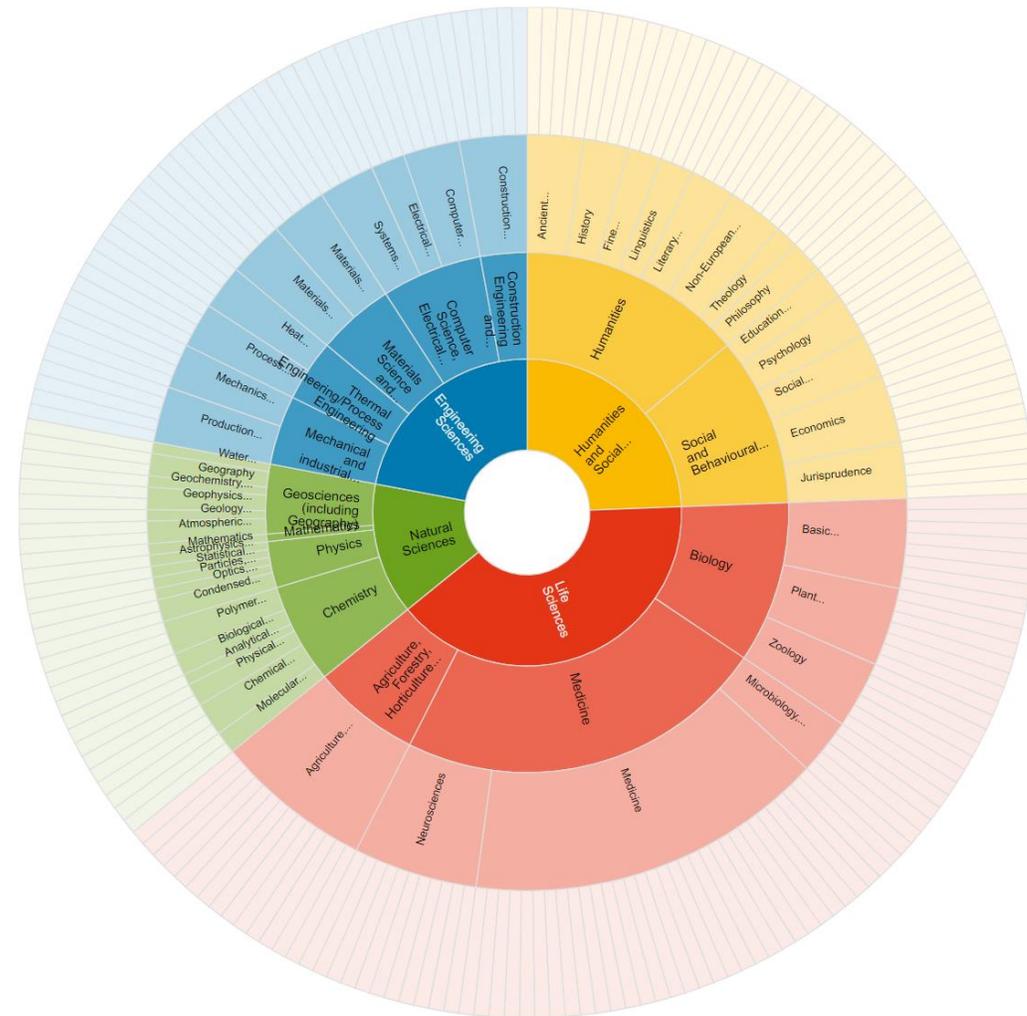
Connect your research – publications, datasets, software, authors, institutions, and funding data all in one place.

<https://datacite.org/>



Making datasets accessible

- Datacite also provides an online tool to identify what **online repository** is right for your dataset according to:
 - Topic
 - Content type (text, database, source code...)
 - Country





Thank you!



<https://www.w3.org/DesignIssues/LinkedData>

Thank you for your attention!

Thanks to Julia Neumann, Federico Toro Grasso and TC-IM 1449 members who helped prepare and review this presentation.

This work is licensed under a Creative Commons Attribution 4.0 International (CC-BY 4.0) license, which allows a free reuse and share for any purpose, as long as appropriate credit to the original source is provided. Please see <https://creativecommons.org/licenses/by/4.0/> for more details.



APPENDIX

- More resources about metadata standards and scientific data sharing:
 - Research Data Alliance
 - <https://www.rd-alliance.org/>
 - FAIRSharing search metadata standards
 - <https://fairsharing.org/>
 - CODATA
 - <https://codata.org/>
 - CASRAI RDM glossary
 - <https://casrai.org/>