

Metadata for Research Data Management and publications submission for EMPIR projects

EURAMET TC IM 1449:
Research Data Management and the European Open Science Cloud

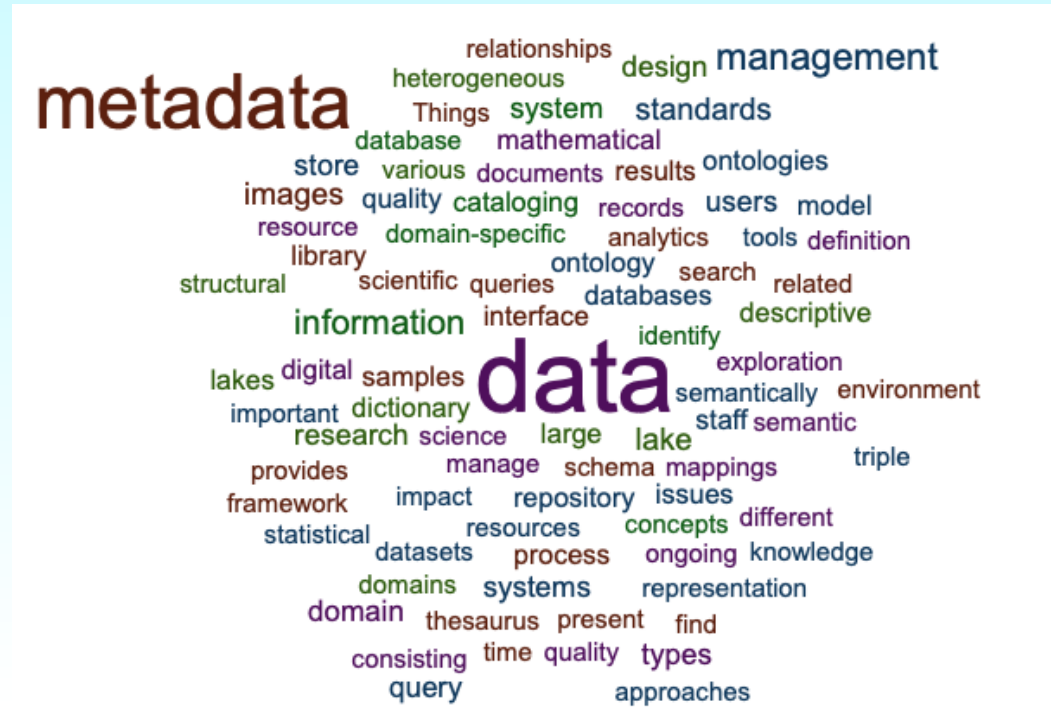
Dr Jean-Laurent Hippolyte (NPL)

Ms Julia Neumann (PTB)



Outline

1. What is metadata and how is it useful to **Research Data Management (RDM)**?
2. Specifying metadata requirements
3. Scientific metadata processing at NPL



METAS



NPL



What is metadata?

Definitions

- “Data about Data”
- Structured information that facilitate retrieval, use or management of some information resource

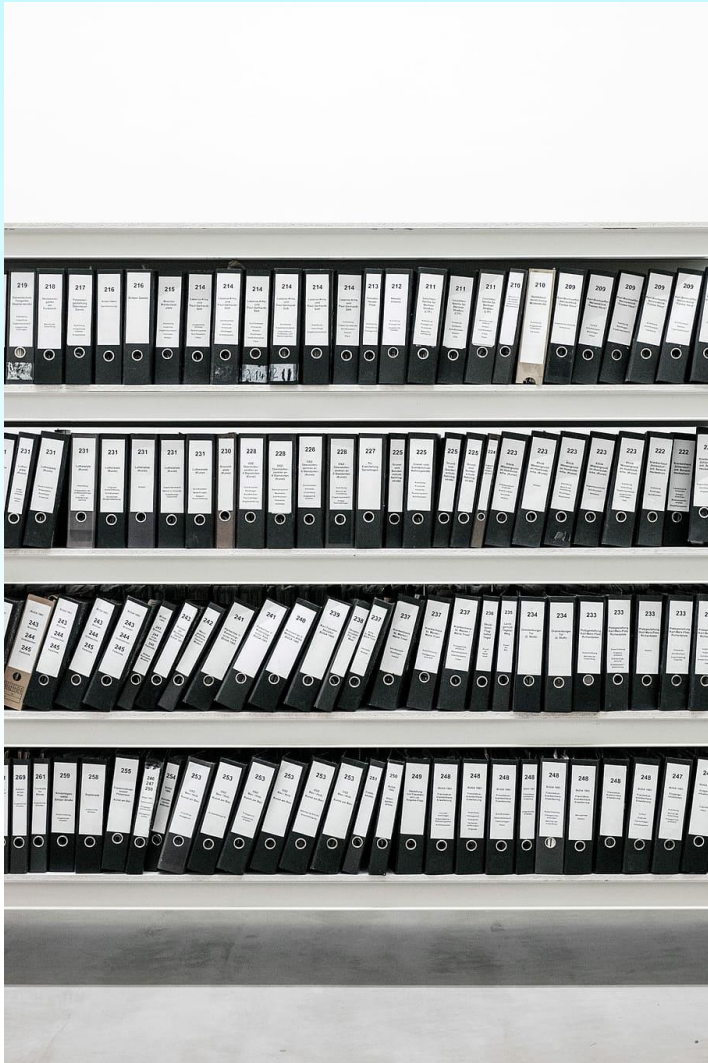
Everyday examples

- File properties in Operating Systems
- Google Knowledge Graph

The image shows a Google search for "alan turing". The search results include a Wikipedia entry, a list of "People also ask" questions, and snippets from the New York Times and Britannica. A yellow arrow points from the search bar area to the Knowledge Graph panel on the right. The Knowledge Graph panel for Alan Turing includes a grid of images, his name and profession, a brief biography, key dates (born: 23 June 1912, died: 7 June 1954), education (Princeton University), and known works. It also features a "Books" section with several book covers and a "People also search for" section with portraits of related figures like Joan Clarke, John von Neumann, and Ada Lovelace.

<http://www.niso.org/publications/understanding-metadata-2017>

What is metadata?

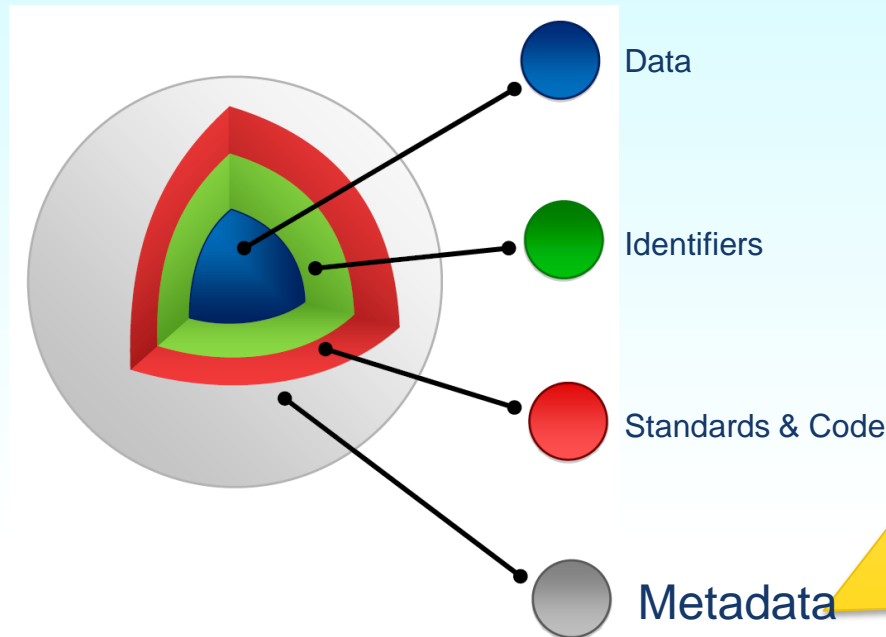


- Gaps in current practices
 - Ad-hoc data organisation
 - file/folder naming conventions
 - Unstandardised description
 - headers in spreadsheets
 - Knowledge embedded in human
 - data loss due to employee turnover

www.data.cam.ac.uk/data-management-guide/organising-your-data

How is it useful to RDM?

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



- Basic metadata
 - Discovering data
- Richer information and provenance
 - Understanding data
- “plurality of relevant attributes” + data usage license
 - Reusing data

How is it useful to RDM?

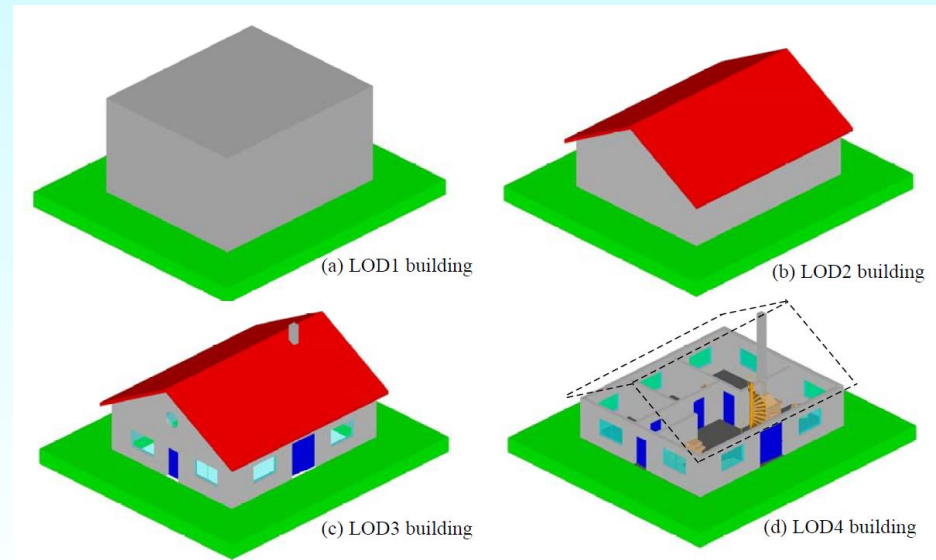
- EURAMET Data Management Plan templates recommend:
 - Sharing datasets via open access repositories, searchable through metadata
 - Metadata to comply with standard vocabularies or schemas
- Many desirable aspects of data quality can't be achieved without metadata:
 - believability, objectivity, reputation, relevancy, interpretability...
 - [MathMet](#) data quality management system

How is it useful to RDM?

Beyond the FAIR principles

- Data quality
- Traceability
- Reproducibility
- Transparency
- Trustworthiness

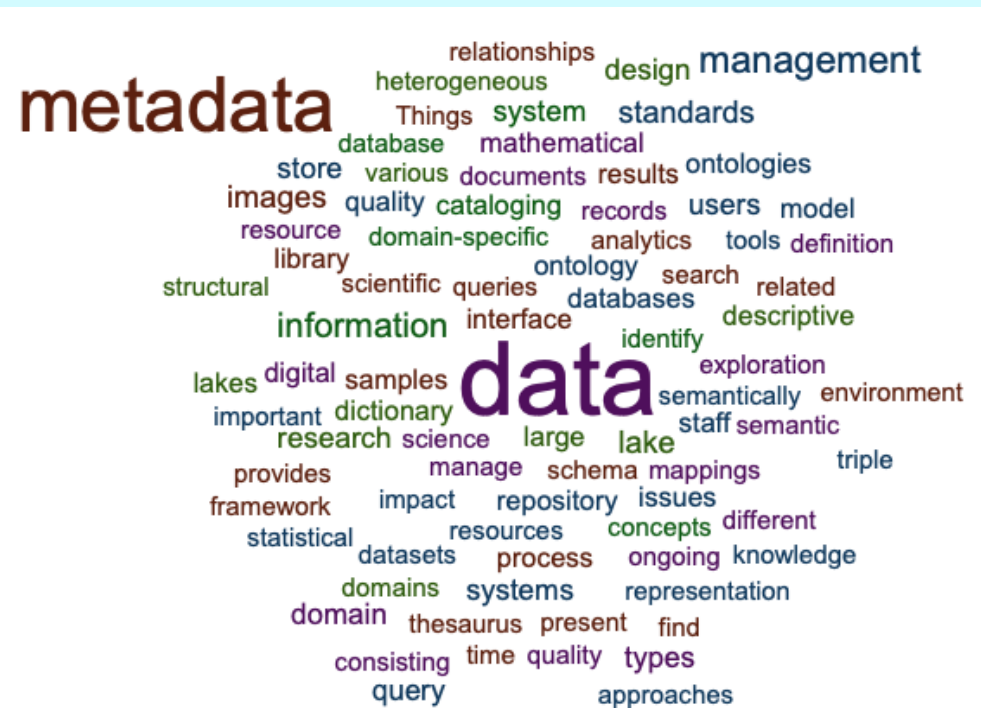
The more comprehensive the metadata, the more value added to data



CityGML Levels Of Detail (source: www.ogc.org)

Outline

1. What is metadata and how is it useful to RDM?
2. Specifying metadata requirements
3. Scientific metadata processing at NPL



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

```
@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",  
}
```

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

Specifying metadata requirements

- In the same way, metadata schemas specify elements to characterize data unambiguously
- Some metadata automatically generated by data acquisition/processing software
- Use general-purpose metadata models to
 - enrich the description of your dataset with non-scientific aspects (organisational, commercial)
 - make your dataset discoverable by non-specialists
 - link your dataset with web resources

[SKOS: captures common concepts of knowledge organisation systems such as taxonomies, glossaries etc..](#)

[DUL: provides upper concepts to leverage interoperability between ontologies](#)

[DCTERMS: standardised metadata elements for resource description](#)

[PROV-O: represent and interchange provenance information generated in different systems and under different contexts](#)

[FOAF: link people and information](#)

[VANN: a vocabulary to annotate vocabularies](#)

[GeoSparql: representing and querying geospatial data](#)

Commonly used generic ontologies

Specifying metadata requirements

- Machine-interpretable metadata languages:
 - [XML/XSD](#),
 - [RDF](#),
 - [OWL](#)
- Open file container formats, metadata+datasets in one file:
 - [NetCDF](#),
 - [HDF5](#),
 - [ADF](#)



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

Specifying metadata requirements

- Metadata for this presentation using dcterms schema:

```
<?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

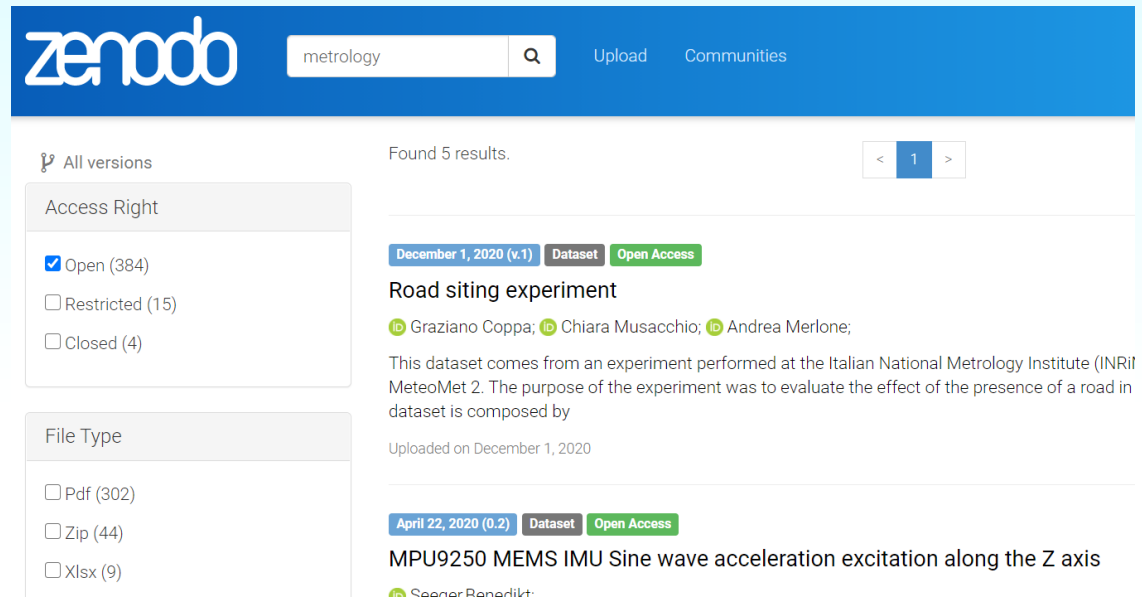
Making datasets accessible

- Generating metadata is not enough to make datasets accessible
 - Datasets and metadata must be uniquely identifiable online
 - Associated metadata must be made searchable
- Restricted VS open repositories
- Cross-domain VS domain-specific

Making datasets accessible

- Zenodo is an open-access repository hosted by CERN
- Zenodo attempts to comply with FAIR principles as best as possible
- Zenodo provides online tools to:

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



The screenshot shows the Zenodo website interface. At the top, there is a search bar with the text 'metrology' and a magnifying glass icon. To the right of the search bar are links for 'Upload' and 'Communities'. Below the search bar, there is a filter section with 'All versions' selected. The filter section has two sub-sections: 'Access Right' and 'File Type'. Under 'Access Right', there are three options: 'Open (384)' (checked), 'Restricted (15)', and 'Closed (4)'. Under 'File Type', there are three options: 'Pdf (302)', 'Zip (44)', and 'Xlsx (9)'. The main content area shows search results. The first result is for a dataset titled 'Road siting experiment' uploaded on December 1, 2020 (v.1). It is labeled as a 'Dataset' and 'Open Access'. The authors listed are Graziano Coppa, Chiara Musacchio, and Andrea Merlone. The description states that the dataset comes from an experiment performed at the Italian National Metrology Institute (INRiM) MeteoMet 2. The second result is for a dataset titled 'MPU9250 MEMS IMU Sine wave acceleration excitation along the Z axis' uploaded on April 22, 2020 (0.2). It is also labeled as a 'Dataset' and 'Open Access'. The author listed is Seener Benedikt.

Making datasets accessible

<https://datacite.org/>

- **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.

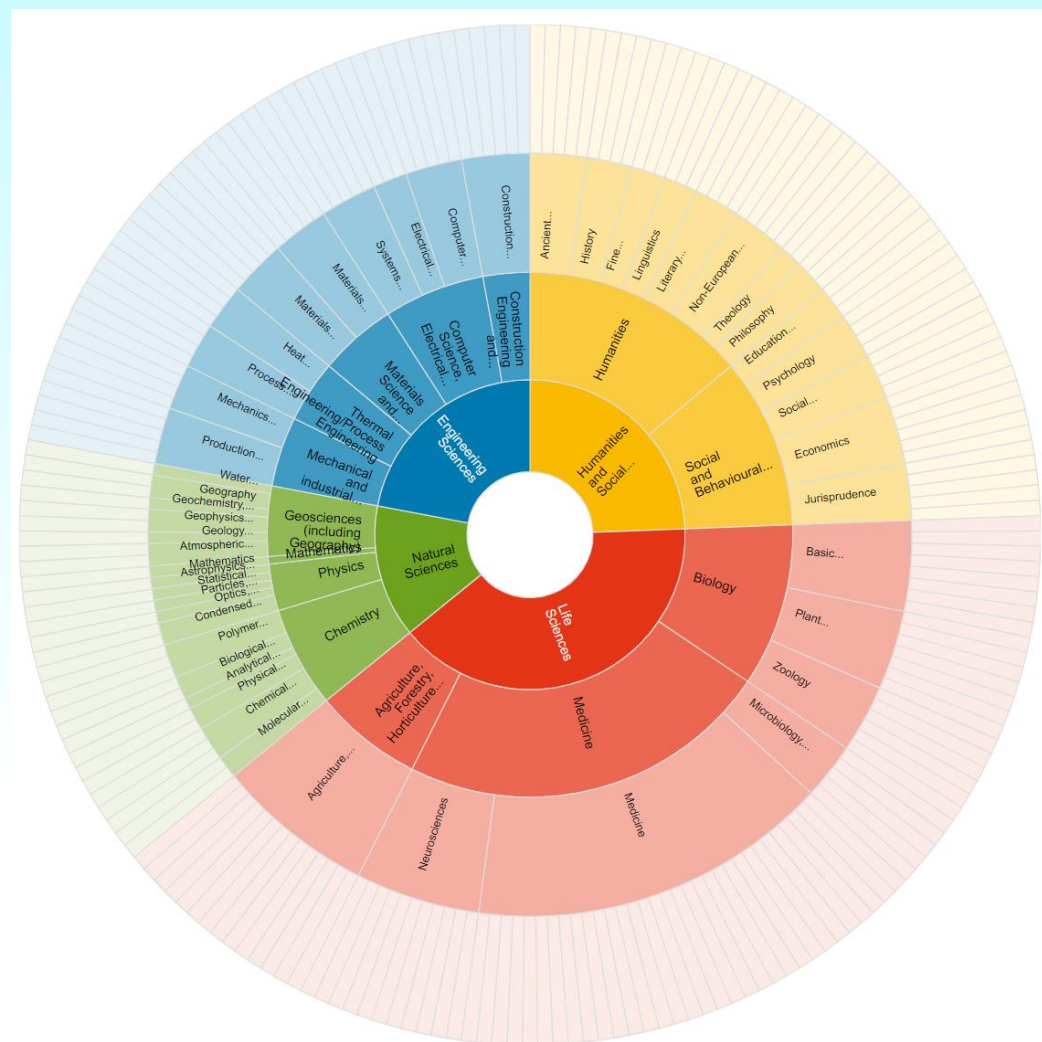
Specifying metadata requirements

- Zenodo and Datacite are not-domain specific
 - [Datacite metadata schema](#)
- Domain-specific metadata standards and repositories exist to enhance discoverability, interoperability and reusability
 - [FAIR R1.3](#) *“If community standards or best practices for data archiving and sharing exist, they should be followed.”*
- For more resources about metadata standards and scientific data sharing:
 - [Research Data Alliance](#)
 - [FAIRSharing](#) search metadata standards
 - [CODATA](#)
 - [CASRAI](#) RDM glossary

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

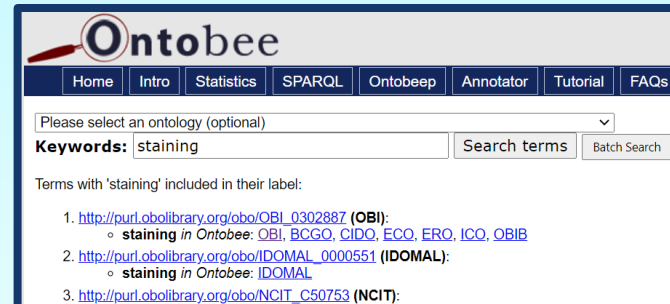


<https://www.re3data.org/browse/>

Specifying metadata requirements

Example of domain-specific metadata schema(s):

- Open Biological and Biomedical Ontology (OBO) Foundry
- Metadata concept search engine (OntoBee)



The screenshot shows the OntoBee search 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. To the right of the input field is a "Search terms" button and a "Batch Search" button. Below the search input, the text "Keywords: staining" is displayed. Underneath, 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, OBI/B
2. http://purl.obolibrary.org/obo/IDOMAL_0000551 (IDOMAL):
 - staining in Ontobee: IDOMAL
3. http://purl.obolibrary.org/obo/NCIT_C50753 (NCIT):



The screenshot shows the detailed search results for the term "staining". The page is structured as follows:

- Class: staining**
- Term IRI:** http://purl.obolibrary.org/obo/OBI_0302887
- 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.
- Annotations:**
 - **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.
 - **has curation status:** pending final vetting
- Class Hierarchy:**
 - 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

<http://www.obofoundry.org/>

Scientific metadata processing at NPL

Knowledge Management System

The screenshot displays the NPL Knowledge Management System interface. It features a search bar with the term 'nuclear fission' and a 'Search' button. Below the search bar, there are sections for 'Wildcard Search' and 'Metadata Search'. The main content area shows search results for Record ID 252, which is an Article titled 'Angular momentum generation in nuclear fission'. The record is in an 'Approved' state. A detailed view of this record is shown below, including a workflow diagram with stages: DRAFT DOCUMENT, GL/SAL REVIEW, IP OFFICE REVIEW, PRE-PUBLICATION, POST PUBLICATION, REPROGRAPHICS REVIEW, and APPROVED. The document details section includes fields for Document Type (Article), Classification (Public), Document Title, Responsible Author (Paddy Regan), Group Leader (Angelo Bella), Science Area Leader (Peter Ivanov), Group (SED/MMN/NUCLEAR), Funding Source (NMS), and Technical Review Team (Andrew Robinson). An abstract is also visible at the bottom.

Search Results
Total 11 result/s found

Record ID	Document Type	Document Title	Process Status
252	Article	Angular momentum generation in nuclear fission	Approved
249	Abstracts	Measurement of fission product gases using a high-resolution beta-	Submit to Pre-Publication
269			
266			
257			
263			
268			

Document Details

Document Type	Classification	Document Title	
Article	Public	Angular momentum generation in nuclear fission	
Responsible Author	Group Leader	Science Area Leader	Group
Paddy Regan	Angelo Bella	Peter Ivanov	SED/MMN/NUCLEAR
Funding Source	Technical Review Team		
NMS	Andrew Robinson		

Abstract

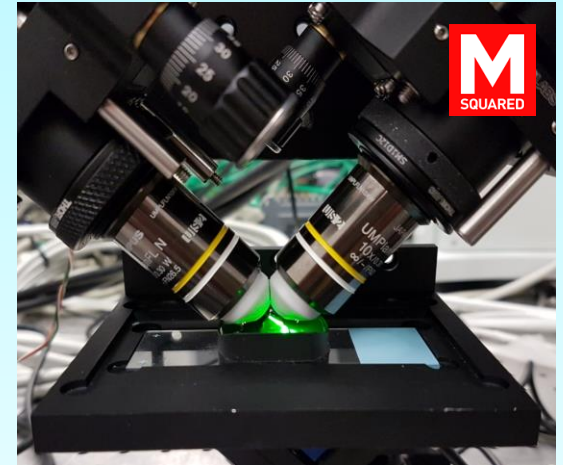
When a heavy atomic nucleus fissions, the resulting fragments are observed to emerge spinning this phenomenon has been an outstanding mystery in nuclear physics for over 40 years . The internal generation of around 6-7 units of angular momentum in each fragment is particularly puzzling for systems which start with zero angular momentum. These systems are considered anomalous which could indicate a transition between the conventional and the superfluid state.

- Centralised repository for NPL publications
- Searchable through metadata
- Basic document metadata, technical review and IP approval workflows

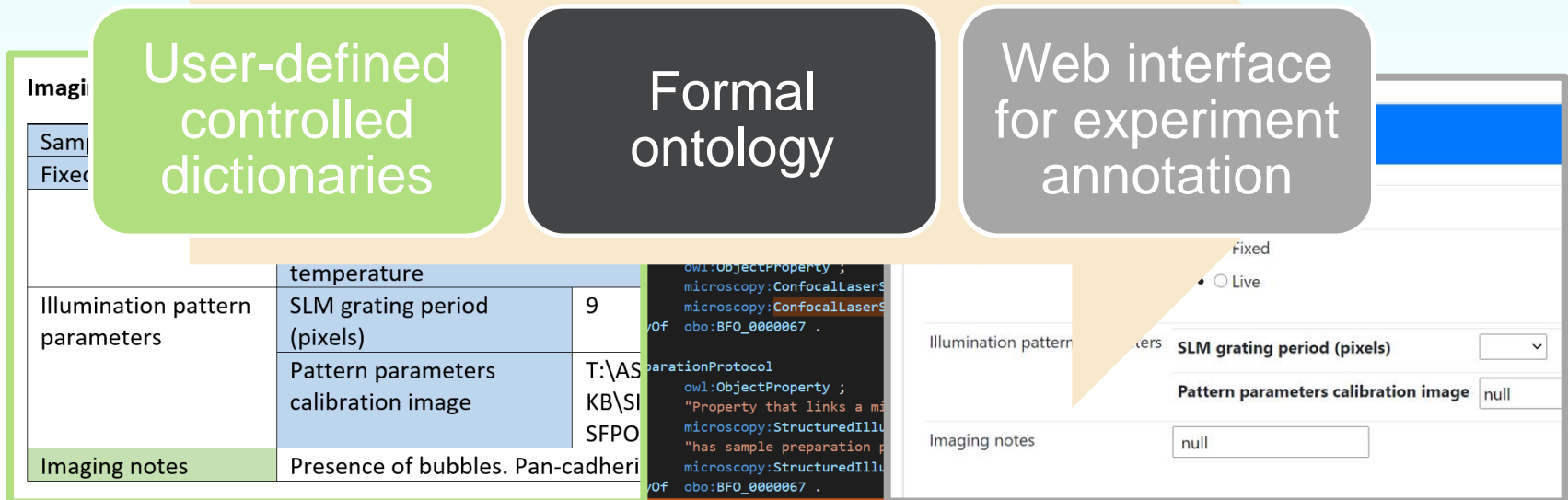
Scientific metadata processing at NPL

Custom microscopy assay metadata generator

- To capture lab-specific experimental setup
- Metadata specification extends community vocabularies from the Open Biological and Biomedical Ontology (OBO) Foundry



Source: Ebeling, C. G., Nat. Biotech., **31** (2013)



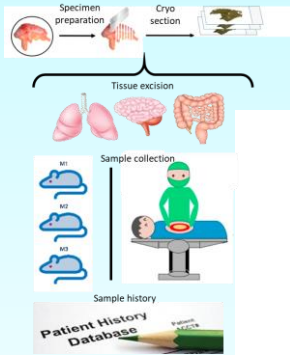
Scientific metadata processing at NPL

Cancer Research UK MSI Data Curation

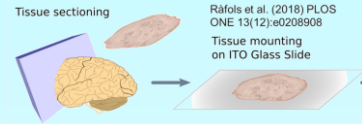
Sample Storage



Sample history

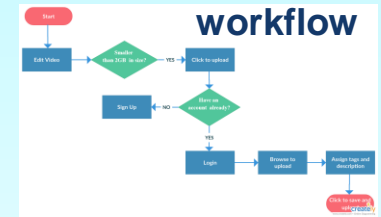


Sample Preparation (@ NPL)

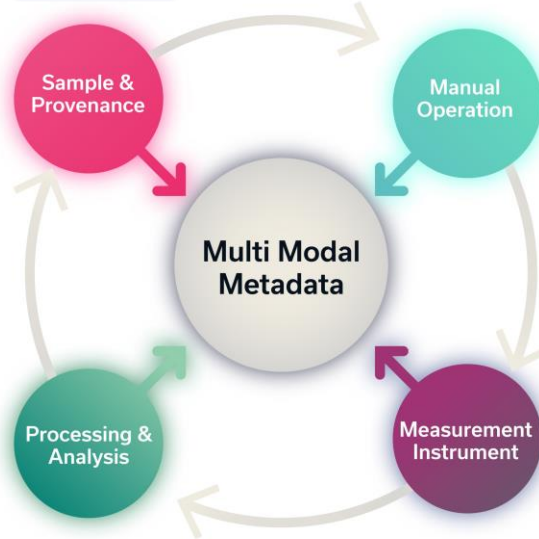


Rafols et al. (2018) PLOS ONE 13(12):e0208908
Tissue mounting on ITO Glass Slide

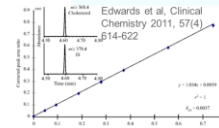
Experimental workflow



Users and Study Information



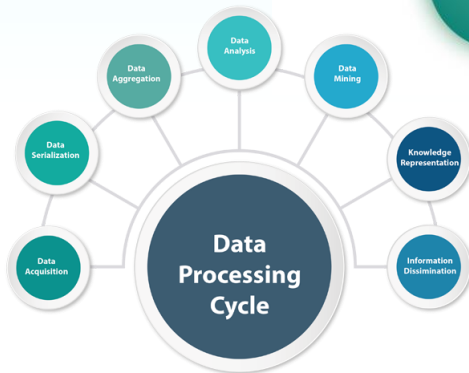
Calibration Data



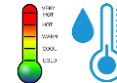
Health and Safety Risk assessment, COSHH

Likelihood	Severity			
	Very Low	Low	Medium	High
Fatality	High	High	High	High
Major trauma	High	High	Medium	Medium
Minor trauma	High	Medium	Medium	Low
Minor injury	Medium	Medium	Low	Low

Data handling



Environmental Conditions Experimental Settings



Study	Beaton SLC75
Sample provider	Beaton
Acquisition start time	10/04/2019 14:13:46
Acquisition completion time	10/17/18 14:28:08
File name	2019_04_30_SLC75_4635_081_081_75um
Sample provider	Beaton
Unique sample ID	00000001
Modality	DESI
Polarity	Negative
Instrument	DESI Kevo
Operator(s) name	Chelsea Nisuke, Tony Steven



Thank you for your attention!

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.



METAS



NPL



Appendix 1

- Some scientific journals focussing on processes for contextualisation, processing of data incl. metadata management:
 - https://www.forschungsdaten.org/index.php/Data_Journals
 - <https://www.nature.com/sdata/>
 - <https://datascience.codata.org/>
 - <https://www.journals.elsevier.com/data-in-brief>