

Web Tutorial 3: Metadata for Research Data Management (RDM) and publications submission for EMPIR projects

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

Dr Jean-Laurent Hippolyte (NPL)

Ms Julia Neumann (PTB)



This work is licensed under a Creative Commons Attribution 4.0 International (CC-BY 4.0) license.

Outline

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



METAS



PTB

NPL



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

alan turing

12,100,000 results (0.67 seconds)

Images News Videos Books More Settings Tools

Alan Turing - Wikipedia

Alan Mathison Turing OBE FRS (/ˈtjʊərɪŋ/; 23 June 1912 – 7 June 1954) was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist.

Use of death: Suicide (disputed) by cyanid... Partner(s): Joan Clarke; (engaged in 1941;...
wards: Smith's Prize (1936) Resting place: Ashes scattered in gardens...

Turing law · Legacy of Alan Turing · Category:Alan Turing · Alan Turing Year

People also ask

Who cracked the Enigma code?

What is Alan Turing most famous for?

Did Alan Turing invent the computer?

What did Alan Turing invent?

What was Alan Turing's IQ?

Is Joan Clarke real?

Who is the real father of computer?

Who is God of computer?

Who made the first computer?

Feedback

www.nytimes.com · 2019/06/05 · obituaries · alan-turin...
Overlooked No More: Alan Turing, Condemned Code Breaker ...
5 Jun 2019 — On June 7, 1954, **Alan Turing**, a British mathematician who has since been acknowledged as one the most innovative and powerful thinkers of the 20th century — sometimes called the progenitor of modern computing — died as a criminal, having been convicted under Victorian laws as a homosexual and forced to endure chemical ...
Place of death: Wilmslow Place of birth: United Kingdom
Date of death: June 7, 1954 Born: June 23, 1912

www.britannica.com · Science · Mathematics ·
Alan Turing | Biography, Facts, & Education | Britannica
Alan Turing , in full **Alan Mathison Turing**, (born June 23, 1912, London, England—died June 7, 1954, Wilmslow, Cheshire), British mathematician and logician, who made major contributions to mathematics, cryptanalysis, logic, philosophy, and mathematical biology and also to the new areas later named computer science, ...
Subjects of study: artificial intelligence, com... Died: June 7, 1954 (aged 41); Wilmslow, E...
Born: June 23, 1912; London, England Role in: World War II

Alan Turing
Mathematician

Alan Mathison Turing OBE FRS was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist. [Wikipedia](#)

Born: 23 June 1912, Maida Vale, London
Died: 7 June 1954, Wilmslow
Education: Princeton University (1936–1938), [MORE](#)
Known for: Cryptanalysis of the Enigma, Turing's proof, [MORE](#)

Quotes [View 4+ more](#)

We can only see a short distance ahead, but we can see plenty there that needs to be done.

I propose to consider the question, 'Can machines think?'

Science is a differential equation. Religion is a boundary condition.

Books [View 5+ more](#)

[The Essential Turing: S...](#) 2004
[Mathem... logic](#)
[Mechanical Intelligence](#) 1992
[Morphog...](#)
[Pure mathem...](#) 1992

People also search for [View 15+ more](#)

[Joan Clarke](#) [John von Neumann](#) [Ada Lovelace](#) [Charles Babbage](#) [Benedict Cumber...](#)

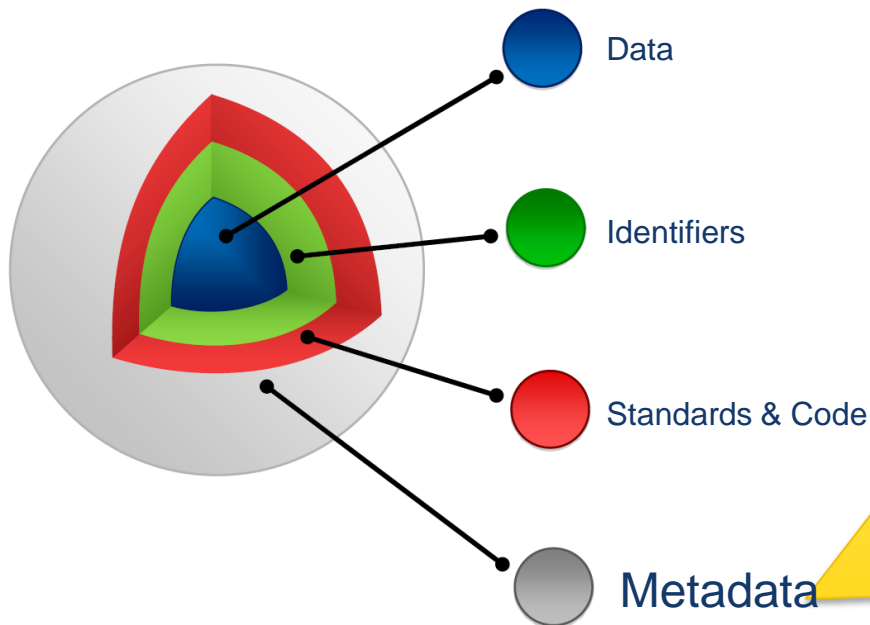
What is metadata?



- 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

How is it useful to RDM?

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



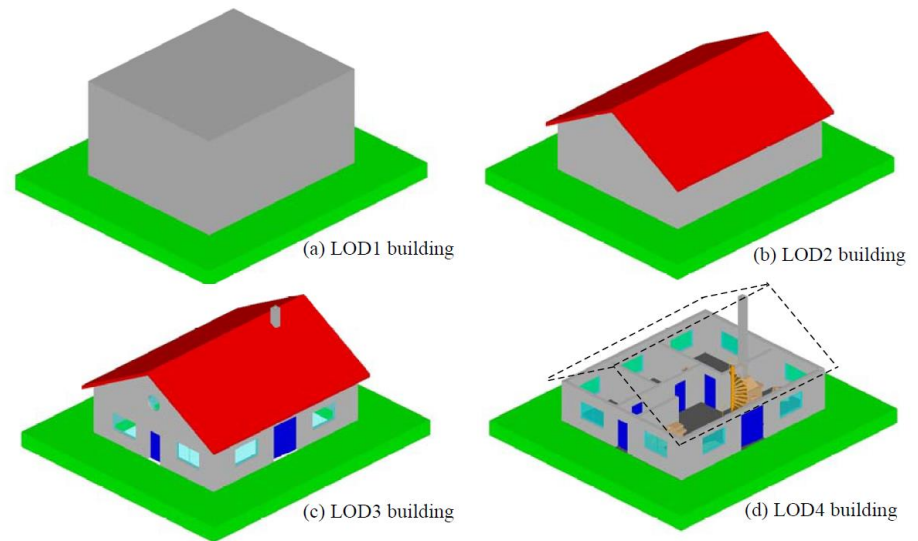
- Basic metadata
 - Discovering data
- Richer information and provenance
 - Understanding data
- “plurality of relevant attributes” + data usage license
 - Reusing data

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



METAS



PTB

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

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

- 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

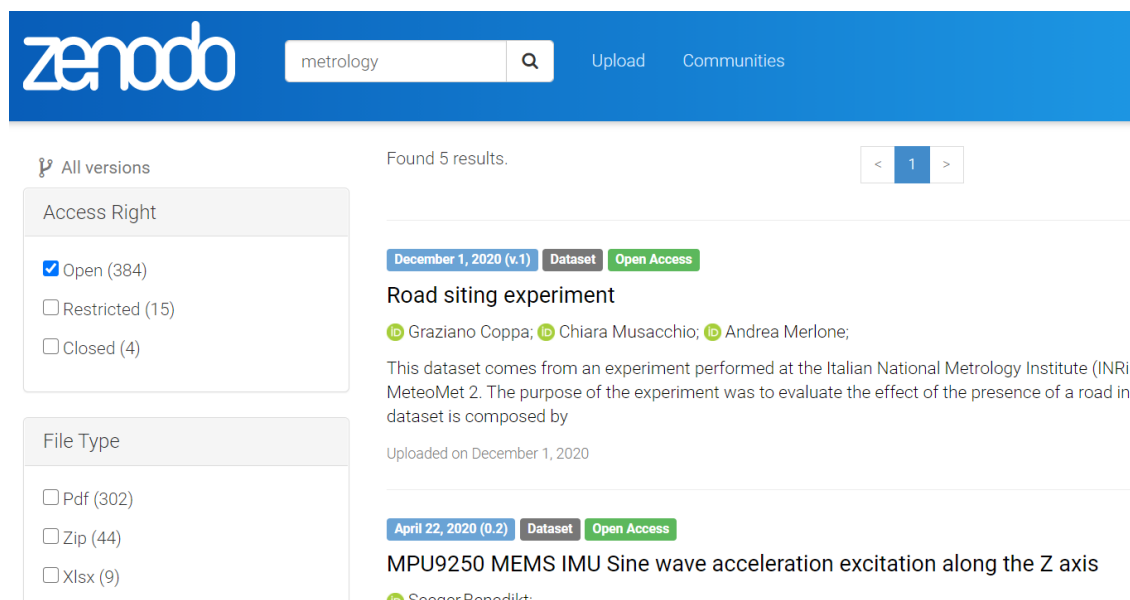
- Generating metadata is not enough to make datasets accessible
 - Datasets+metadata must be uniquely identifiable online
 - Associated metadata must be made searchable
- Restricted VS open repositories
- Cross-domain VS domain-specific

Making datasets accessible

- **Zenodo** an open-access repository hosted by CERN
- 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 displays the Zenodo website interface. At the top, there is a blue header with the Zenodo logo, a search bar containing the text 'metrology', and links for 'Upload' and 'Communities'. Below the header, the page shows search results. On the left, there are two filter panels: 'Access Right' with options 'Open (384)', 'Restricted (15)', and 'Closed (4)'; and 'File Type' with options 'Pdf (302)', 'Zip (44)', and 'Xlsx (9)'. The main content area shows 'Found 5 results.' with a pagination control showing page 1. The first result is titled 'Road siting experiment', dated 'December 1, 2020 (v.1)', and is marked as a 'Dataset' with 'Open Access'. It lists authors: 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 titled 'MPU9250 MEMS IMU Sine wave acceleration excitation along the Z axis', dated 'April 22, 2020 (0.2)', and is also marked as a 'Dataset' with 'Open Access'. It lists the author: 'Seonar Benadik'. The Zenodo logo is visible in the top left of the page content area.

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

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.

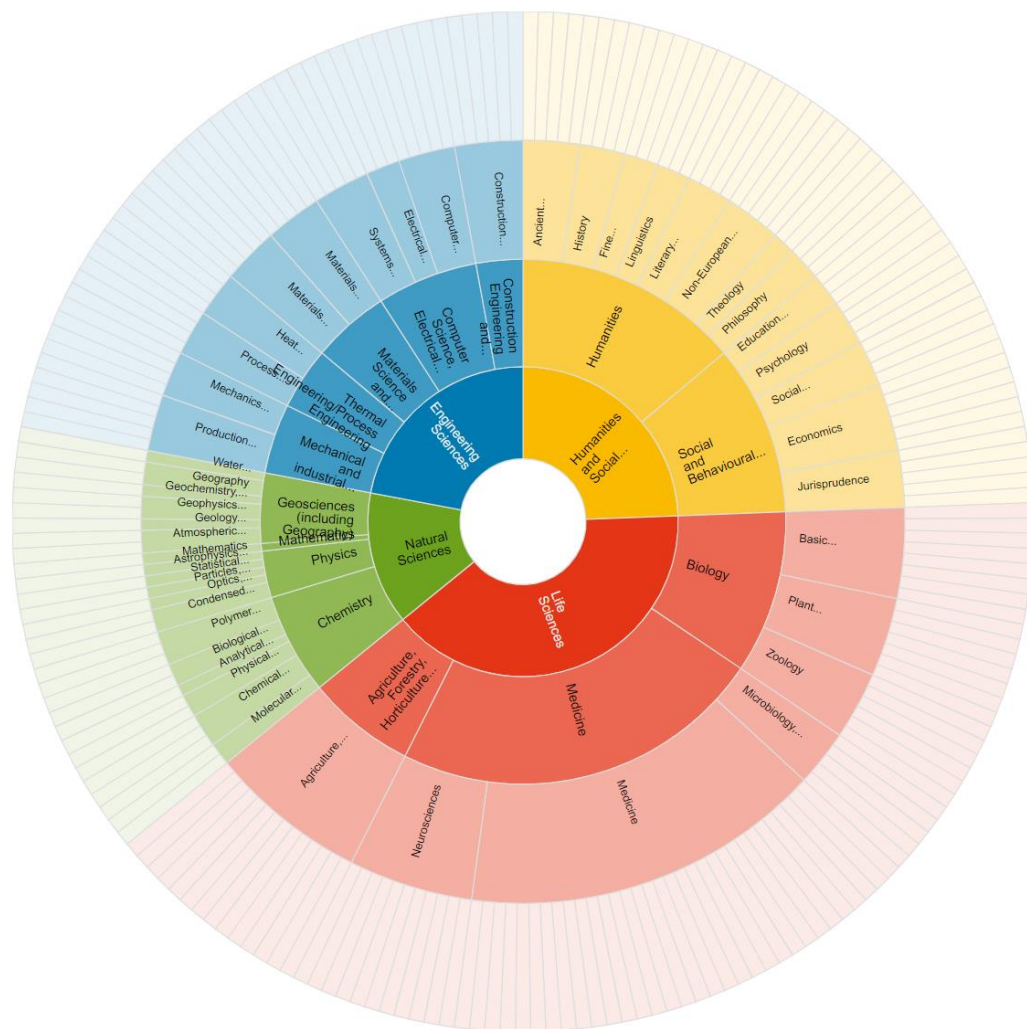
<https://datacite.org/>

Specifying metadata requirements

- Zenodo and Datacite 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

- Datasite 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



Specifying metadata requirements

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

The screenshot displays the OntoBee web application. At the top, there is a navigation bar with links: Home, Intro, Statistics, SPARQL, Ontobee, Annotator, Tutorial, and FAQs. Below this is a search interface with a dropdown menu for selecting an ontology (optional), a text input field containing the keyword 'staining', and buttons for 'Search terms' and 'Batch Search'. The results section, titled 'Terms with 'staining' included in their label:', lists three items:

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):

The first result is expanded to show details for the class 'staining':

- 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

Outline

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



METAS



PTB

NPL



Scientific metadata processing at NPL

- Knowledge Management System

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

The screenshot displays the NPL Knowledge Management System interface. It features a 'Wildcard Search' bar with the search term 'nuclear fission' and a 'Search' button. Below this is a 'Metadata Search' section. The main content area shows 'Search Results' for 'Total 11 result/s found'. A list of records is displayed, including Record ID 252, 249, 269, 266, 257, 263, and 268. Each record entry includes the document type, responsible author, last modified date, and document title. A detailed view of Record ID 252 is shown, including a process status bar (DRAFT DOCUMENT, GL/SAL REVIEW, IP OFFICE REVIEW, PRE-PUBLICATION, POST-PUBLICATION, REPROGRAPHICS REVIEW, APPROVED) and a timeline of events. The 'Document Details' section includes fields for Document Type, Classification, Document Title, Responsible Author, Group Leader, Science Area Leader, Group, Funding Source, and Technical Review Team. An abstract is also provided at the bottom.

NPL Knowledge Management System
National Physical Laboratory

Wildcard Search
Search Terms: nuclear fission Search

Metadata Search
Document Type: Find items
Document Title: DOI: Search

Search Results
Total 11 result/s found

Record ID: 252
Responsible Author: Paddy Regan
Last Modified: 26/02/2021 15:56
Document Type: Article
Document Title: Angular momentum generation in nuclear fission
Process Status: Approved

Record ID: 249
Responsible Author: Paddy Regan
Last Modified: 26/02/2021 15:56

Record ID: 269
Responsible Author: Paddy Regan
Last Modified: 18/02/2021 15:56

Record ID: 266
Responsible Author: Paddy Regan
Last Modified: 17/02/2021 15:56

Record ID: 257
Responsible Author: Paddy Regan
Last Modified: 17/02/2021 15:56

Record ID: 263
Responsible Author: Paddy Regan
Last Modified: 19/02/2021 15:56

Record ID: 268
Responsible Author: Paddy Regan
Last Modified: 19/02/2021 15:56

Document Details

Document Type: Article
Classification: Public
Document Title: Angular momentum generation in nuclear fission

Responsible Author: Paddy Regan
Group Leader: Angelo Bella
Science Area Leader: Peter Ivanov
Group: SED/MMN/NUCLEAR

Funding Source: NMS
Technical Review Team: 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 are systems in which the angular momentum is conserved and the angular momentum is carried away by the fragments.

© National Physical Laboratory 2021

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

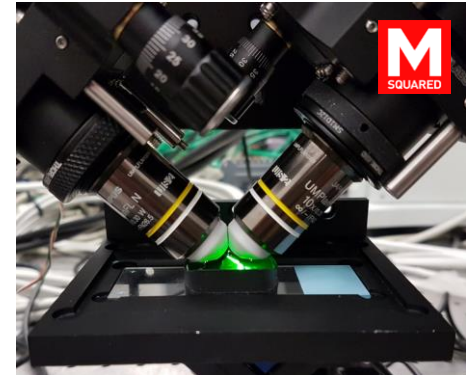
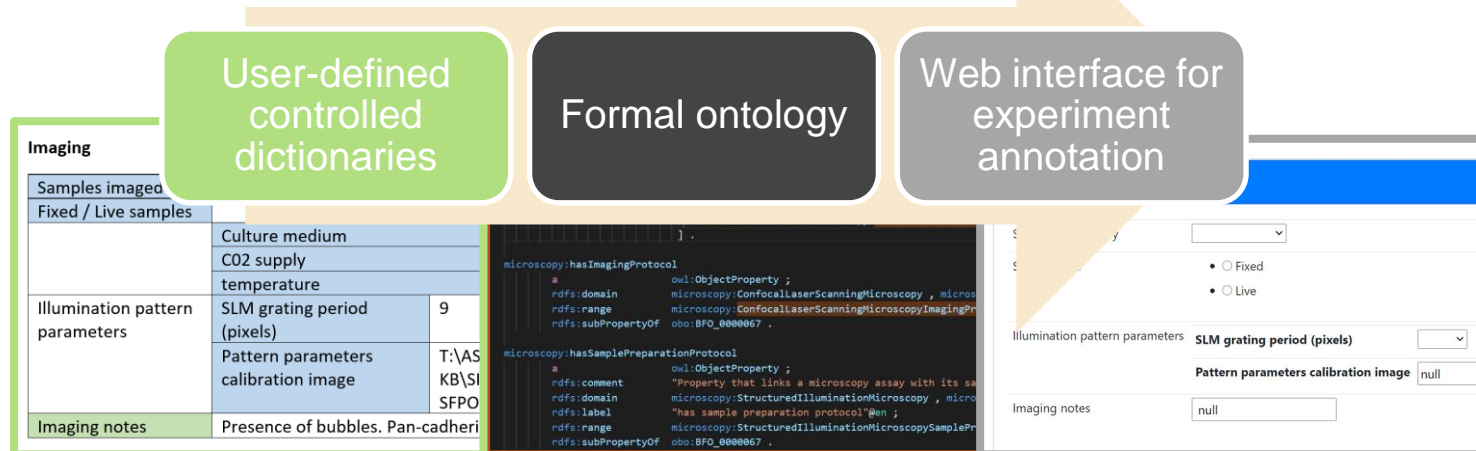
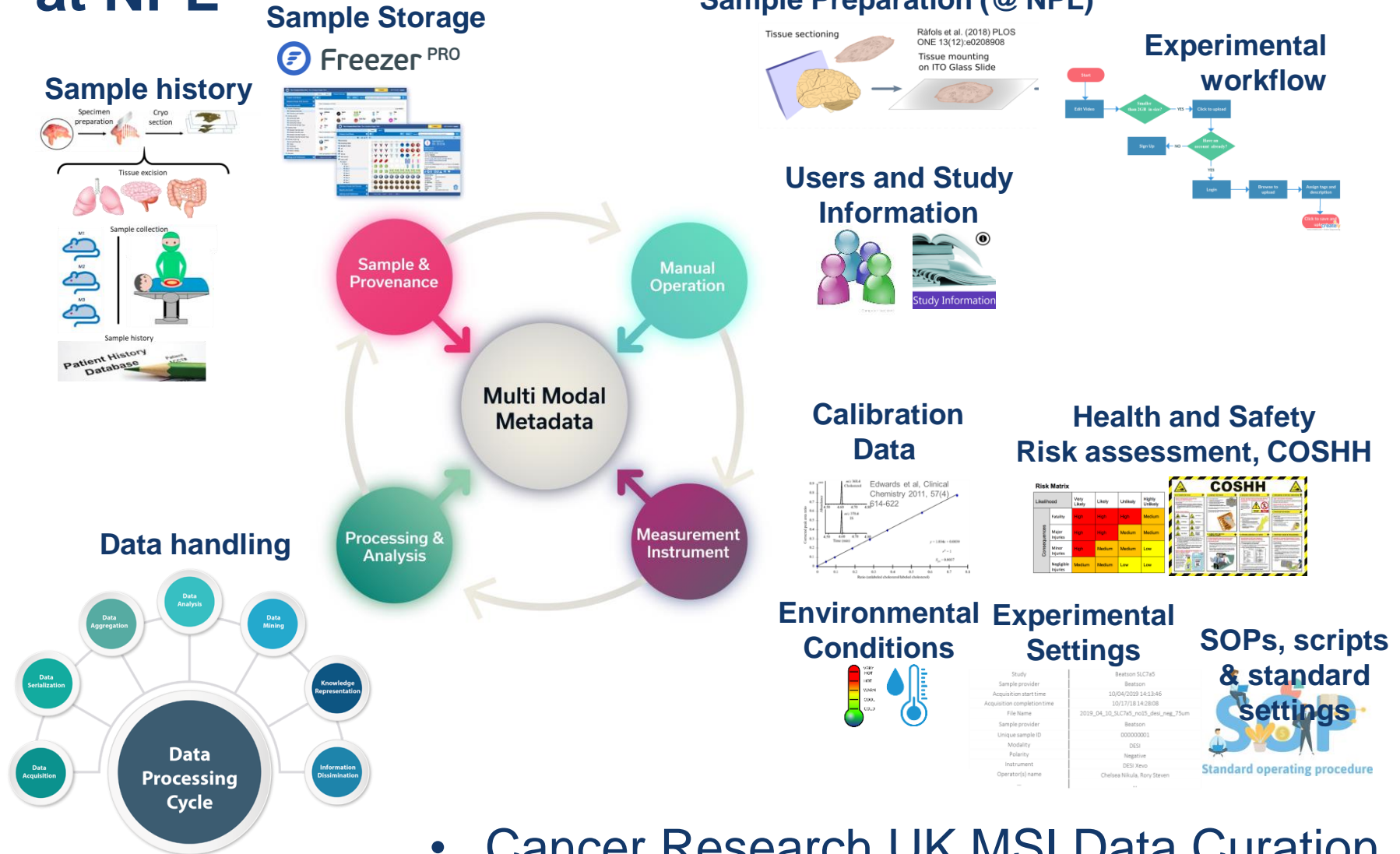


Figure credit: Ebeling, C. G., Nat. Biotech., **31** (2013)



at NPL



- Cancer Research UK MSI Data Curation

**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.



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>