Metadata from an Experimental Perspective

Brian Matthews

Facilities Programme Manager
Scientific Computing Department, STFC

Co-Chair RDA Photon and Neutron Science Interest Group

Task lead, NA on metadata standardisation, NFFA-Europe
Large-Scale Analytic Facilities

Key challenges of the 21st century
- energy, global climate, health and security
- study matter at the scales
  - from single atoms \((10^{-10} \text{ m})\) to living cells \((10^{-6} \text{ m})\) to whole systems \((10^{-3} - 1 \text{ m})\)

High resolution “microscopes” \(\rightarrow\) intense beams of particles \(\rightarrow\) Specialist sources

Requires large scale research infrastructures that are beyond the capability of any single university or research group

**Photons** (X-Rays) “see” electric charge – high atomic number nuclei

**Neutrons** “see” nucleons – including hydrogen atoms
Facility Data Lifecycle

Proposal

Approval

Scheduling

Experiment

Data reduction

Data analysis

Publication

Metadata Catalogue

http://www.icatproject.org
The Core Metadata model forms the information model for ICAT.

Designed to describe facilities based experiments in Structural Science throughout a facility’s scientific workflow.

- Uses a **Facilities** view of an experiment – an “Investigation” (proposal)

For use within the repository for organising data

http://purl.org/net/CSMD
http://icatproject.org/CSM
CSMD : A Successful Metadata Model
Metadata as Middleware

• Implemented in ICAT
  • worked for many years.

• Representations
  • RDF, XML, JSON,
  • As a Database schema

• Tracking provenance: modified model to support:
  • Derived data
  • Software, jobs
  • Linking between these

```xml
<owl:Class rdf:about="csmd:Investigation"> ... </owl:Class>
<owl:Class rdf:about="csmd:Facility"> ... </owl:Class>
<owl:Class rdf:about="csmd:Dataset"> ... </owl:Class>
```
Publishing and Sharing Metadata

- Publish metadata to general purpose harvesters, and search engines which provide search tools across disciplines
  - Being developed by other e-Infrastructure projects

- Published Data Sets
  - A DOI from DataCite [www.datacite.org](http://www.datacite.org)

- Worked with the EUDat project
  - B2Find Data Discovery Service
    - [www.eudat.eu](http://www.eudat.eu)

- Made core metadata available to B2Find
  - OAI-PMH interface
  - Published Data (e.g. with DOIs).

- Mapping of CSMD metadata to Dublin Core and EUDat metadata requirements.

<table>
<thead>
<tr>
<th>Metadata Field</th>
<th>EUDAT Field(s)</th>
<th>ICAT Field(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc:identifier</td>
<td>-</td>
<td>Investigation---&gt;doi</td>
</tr>
<tr>
<td>dc:title</td>
<td>title</td>
<td>Investigation---&gt;title</td>
</tr>
<tr>
<td>dc:description</td>
<td>notes</td>
<td>Investigation---&gt;summary</td>
</tr>
<tr>
<td>dc:relation</td>
<td>tags</td>
<td>Instrument---&gt;fullName</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Investigation---&gt;name</td>
</tr>
<tr>
<td></td>
<td></td>
<td>InvestigationParameter---&gt;name (multiple)</td>
</tr>
<tr>
<td>dcterms:references</td>
<td>URL</td>
<td>“dx.doi.org/” + Investigation---&gt;doi</td>
</tr>
<tr>
<td>dc:creator</td>
<td>author</td>
<td>User---&gt;fullName</td>
</tr>
<tr>
<td></td>
<td>spatial</td>
<td>-</td>
</tr>
<tr>
<td>dc:contributor</td>
<td>maintainer</td>
<td>Science and Technology Facility Council, ISIS</td>
</tr>
<tr>
<td>dc:subject</td>
<td>discipline</td>
<td>“Clean energy and the environment, pharmaceuticals and health care, nanotechnology and materials engineering, catalysis and polymers, fundamental studies of materials”</td>
</tr>
<tr>
<td></td>
<td>PublicationYear</td>
<td>-</td>
</tr>
<tr>
<td>dcterms:issued</td>
<td>PublicationTimestamp</td>
<td>Investigation---&gt;releaseDate</td>
</tr>
<tr>
<td>dcterms:temporal</td>
<td>Temporal Coverage: EndDate</td>
<td>Investigation---&gt;startDate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An open access resource for experimental & theoretical nanoscience

**Information and Data Management Repository Platform for nanoscience**

- An integrated platform
  - covering the full research cycle by the users.
  - automatic acquisition of key metadata
  - a data repository for future data access

*Defining metadata standards for data sharing in nanoscience*

- To represent data from nanoscience experiment and theoretical analysis.
- Use currently available standards e.g. from PaNData project.
- STFC, CNR-IOM, ESRF, KIT, FORTH
- Work within the RDA – with the Materials IG
Metadata for Nanomaterials Data

- Workflow for Nano-structured Science
  - Metadata focussed around the Project
    - A user centred view
  - NFFA Deliverable 11.2: Draft Metadata Standard
    - 29th February 2016
# Core vocabulary for Entities

<table>
<thead>
<tr>
<th>Experiment Concepts</th>
<th>Data Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Research User</td>
<td>• Raw Data</td>
</tr>
<tr>
<td>• Instrument Scientist</td>
<td>• Analyzed Data</td>
</tr>
<tr>
<td>• Project</td>
<td>• Data Asset</td>
</tr>
<tr>
<td>• Proposal</td>
<td>• Data Analysis</td>
</tr>
<tr>
<td>• Facility</td>
<td>• Data Analysis Software</td>
</tr>
<tr>
<td>• Instrument</td>
<td>• Data Archive</td>
</tr>
<tr>
<td>• Experiment</td>
<td>• Data Policy</td>
</tr>
<tr>
<td>• Measurement</td>
<td>• Data Manager</td>
</tr>
<tr>
<td>• Sample</td>
<td>• Data Curation Activity</td>
</tr>
</tbody>
</table>

RDA Metadata IG is working on common concepts
Relations between Entities

- Proposal
- Experiment
- Measurement
- Raw Data
- Data Analysis
- Data Analysis Software
- Project
- Facility
- Instrument
- Sample
- Instrument Scientist
- Research User
- NFFA Portal
- Data Asset
- Analyzed Data
- Data Archive
- Data Archive Policy
- Data Curation Activity
- Data Manager
## Entity Attributes (part)

<table>
<thead>
<tr>
<th>Metadata elements and subgroups</th>
<th>Related Information entity</th>
<th>Value Type</th>
<th>Cardinality</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User ID</strong></td>
<td>Research User</td>
<td>Identifier</td>
<td>Mandatory; Unique</td>
<td>Unique identifier for the user</td>
</tr>
<tr>
<td><strong>User name</strong></td>
<td>Research User</td>
<td>Text</td>
<td>Mandatory</td>
<td>Commonly user name of the user</td>
</tr>
<tr>
<td><strong>User Identifier</strong></td>
<td>Research User</td>
<td>Text</td>
<td>Optional</td>
<td>Unique identifier (PID/URI) assigned to the user by an external organization e.g. ORCID.</td>
</tr>
<tr>
<td><strong>User affiliation</strong></td>
<td>Research User</td>
<td>Text</td>
<td>Optional</td>
<td>Institutional affiliation of the user.</td>
</tr>
<tr>
<td><strong>Instrument Scientist ID</strong></td>
<td>Instrument Scientist</td>
<td>Identifier</td>
<td>Mandatory; Unique</td>
<td>Unique identifier for the Instrument Scientist</td>
</tr>
<tr>
<td><strong>Instrument Scientist name</strong></td>
<td>Instrument Scientist</td>
<td>Text</td>
<td>Mandatory</td>
<td>Commonly user name of the Instrument Scientist</td>
</tr>
<tr>
<td><strong>Instrument Scientist Identifier</strong></td>
<td>Instrument Scientist</td>
<td>Text</td>
<td>Optional</td>
<td>Unique identifier (PID/URI) assigned to the Instrument Scientist by an external organization e.g. ORCID.</td>
</tr>
<tr>
<td><strong>Project ID</strong></td>
<td>Project</td>
<td>Identifier</td>
<td>Mandatory, Unique</td>
<td>Unique identifier for the Project. It may be assigned by a funding body (so known prior to applying for the facility time slot), or the project ID generation feature/service can be offered by a proposal registration system</td>
</tr>
<tr>
<td><strong>Project name</strong></td>
<td>Project</td>
<td>Text</td>
<td>Mandatory</td>
<td>Name for the project</td>
</tr>
<tr>
<td><strong>Project description</strong></td>
<td>Project</td>
<td>Text</td>
<td>Optional</td>
<td>Textual description of the project</td>
</tr>
<tr>
<td><strong>Proposal ID</strong></td>
<td>Proposal</td>
<td>Identifier</td>
<td>Mandatory, Unique</td>
<td>Unique identifier for the Proposal</td>
</tr>
<tr>
<td><strong>Proposal description</strong></td>
<td>Proposal</td>
<td>Text</td>
<td>Optional</td>
<td>Textual description of the project</td>
</tr>
<tr>
<td><strong>Facility ID</strong></td>
<td>Facility</td>
<td>Identifier</td>
<td>Mandatory, Unique</td>
<td>Unique identifier for the Facility</td>
</tr>
<tr>
<td><strong>Facility Identifier</strong></td>
<td>Facility</td>
<td>Text</td>
<td>Optional</td>
<td>Unique identifier (e.g. PID/URI) assigned to the Facility by an external organization.</td>
</tr>
<tr>
<td><strong>Facility name</strong></td>
<td>Facility</td>
<td>Text</td>
<td>Mandatory</td>
<td>Common name for the Facility</td>
</tr>
</tbody>
</table>
Conclusions

- Want to agree
  - Core Entities in the Metadata Vocabulary
    - Agreed definitions
  - Nature of Relationship between entities
  - Attributes for all Entities
- Use this as a basis for a common metadata directory of nano-structure experiments and their associated data.
  - Will need to map onto common metadata standards
- Controlled vocabulary for fields
  - Techniques, instruments, facilities, parameters
- Overlap with the Materials IG
  - Want to join in the new working group
    - RDA Working Group International Materials Resource Registries
  - Especially in defining metadata standards