Core semantic model for generic research activity

Vasily Bunakov
Science and Technology Facilities Council
United Kingdom
STFC

Funds and operates large scale instruments for the UK and visitor researchers in:
- physics, astronomy
- chemistry, materials
- biology, medicine

Scientific Computing develops and operates computing infrastructure:
- High Performance Computing
- Petabyte data store
- CERN LHC Tier 1 hub
also conducts applied research and does software development
Big Facilities for Small Science

- Central Laser Facility
- Diamond Light Source
- ISIS neutron and muon source
PaNdata projects

PaNdata Europe
2010 – 2011
Preparation: common policies and standards
http://pan-data.eu/pandata/?q=PaNdataEurope

PaNdata ODI
2011 – 2014
Implementation: delivering new infrastructure
http://pan-data.eu/pandata/?q=ODIWP
Facilities Research Lifecycle

Scientist submits application for beamtime

Facility committee approves application

Facility registers, trains, and schedules scientist’s visit

Scientists visits, facility run’s experiment

Raw data filtered, and stored

Tools for processing made available

Subsequent publication registered with facility

Data catalogue software: http://code.google.com/p/icatproject/
CSMD: Core Scientific MetaData Model

- Topic
- Publication
- Keyword
- Authorisation
- Investigation
- Investigator
- Dataset
- Sample
- Sample Parameter
- Datafile
- Dataset Parameter
- Datafile Parameter
- Related Datafile
- Parameter

CSMD forms the information model for facilities data catalogues
We joined DataCite

Much cheaper DOIs than directly from DOI Foundation

www.DataCite.org
Is it really about data?
Our DOIs landing pages are in fact for Investigations (Experiments)

Red is for “data” notion, and green is for “investigation”

**ISIS Data**

**Investigation title:** Crystal and magnetic structures of EuWO$_1$+xN$_2$-x.

**Creator:** Kusmartseva, A  
**Creator:** Rodgers, J A  
**Creator:** Atfield, J P

**DOI:** 10.5286/ISIS.E.24071239

**Date of Experiment:** Tue Aug 04 14:38:23 BST 2009

**Publisher:** STFC ISIS Facility

**Data format:** RAF/INexus  
Select the data format above to find out more about it.

**Data Citation**

The recommended format for citing this dataset in a research publication is as:

[author], [date], [title], [publisher], [doi]

For Example:


**Abstract**

Eu$_2$+ d0- transition metal perovskites are of interest as potential multiferroics when undoped, or as CMR materials. EuWO$_1$+xN$_2$-x is a new magnetoresistive material and exists over a broad range of x = -0.2 to 0.5. It has a ferromagnetic ordering transition at TC = 12 K. Neutron diffraction is needed to determine the $I\bar{1}12/m$ monoclinic superstructure evidenced by TEM that arises from O/N ordering and octahedral tilting, and the magnetic order. This may include a coexistence of antiferromagnetic/ferromagnetic orders (as found in a previous GEM study of the analogue EuNbO2N). 2 days on GEM are needed to study 2 samples with different x values (one stoichiometric x =0, the other highly doped x = 0.5) because of high absorption by Eu.
We are not alone in DataCite “abuse”

Enhanced corrosion protection by microstructural control of aluminium brazing sheet

Attachments

Cite or link this publication as: doi:10.4233/uuid:b7113b48-08a4-41a1-871bf524df5db

Author: Norouzi Afshar, F.
Promotor: De Wit, J.H.W. · Terryn, H.
Faculty: Mechanical, Maritime and Materials Engineering
Department: Materials Science & Engineering
Type: Dissertation
Date: 2013-08-30
ISBN: 9789077172933
Keywords: aluminium brazing sheet · accelerated corrosion test · microstructural characterization · electrochemical characterization · heat treatment
We used to think our metadata is for “data” but in fact, quite often it is for “activity”, e.g. Experiment or Study.
Research activity is not restricted to Experiment or Study and can be a part of a longer “value chain”

DDI record for social science Study decomposed

Archives:  
www.data-archive.ac.uk  
www.gesis.org  
and many more

DDI portal:  
www.ddialliance.org

Project:  www.engage-project.eu  
Platform:  www.engagedata.eu
ENGAGE vision: promotion of Open Data to Linked Open Data through collaborative data curation

Project: www.engage-project.eu  Platform: www.engagedata.eu
To make research data linkable, we need to reasonably model research activity

- Keep the model generic enough
- Keep it simple for better adoption and “opportunistic” application
- Aim it not at humans only but at machines / software agents, too
Do we have reasonable research activity models?

DARIAH Scholarly Research Activity

I2S2 Scientific Research Activity Lifecycle

www.dariah.eu

www.ukoln.ac.uk/projects/I2S2/
Concerns about existing research activity models

- Domain-specific
- Elements seem well defined but are open to different interpretations
- Are not “Linked Data ready”
- Overdone to be easily adopted and consistently used
Possible response: offering a (simple) generic research activity model suitable for adoption by different stakeholders
## Research activity cell

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Something that is taken in or operated on by Activity</td>
<td>Previous research, Raw data</td>
</tr>
<tr>
<td>Output</td>
<td>Something that is intentionally produced by Activity</td>
<td>Raw data, Derived (analyzed) data</td>
</tr>
<tr>
<td>Scope</td>
<td>Something that Activity is aimed at or deals with</td>
<td>Sample properties, One or more experiments</td>
</tr>
<tr>
<td>Condition</td>
<td>Something that affects or supports Activity, or gives it a specific context</td>
<td>Scientific instrument, IT environment</td>
</tr>
<tr>
<td>Actor</td>
<td>Something or somebody who participates in Activity</td>
<td>Investigator, Data analyst</td>
</tr>
<tr>
<td>Effect</td>
<td>Something that is a consequence of Activity</td>
<td>Environment pollution, New software module</td>
</tr>
</tbody>
</table>
What we (different stakeholders of the research lifecycle) actually want to monitor and exploit is “research value chains”, to ensure the golden-eggs-laying goose of research is productive = brings enough eggs for everyone involved.

Research activity cells combined in “grid” should result in better research navigation and research contextualization for everyone involved.
RDFS Plus representation (see in paper) and model extensions

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rm: <http://example.org/stuff/ResearchModel#>.
# For Conditions
rm:Regulation rdfs:subClassOf rm:Condition .
rm:DataManagementPolicy rdfs:subClassOf rm:Regulation .
# For Output
rm:Publication rdfs:subClassOf rm:Output .
rm:Dataset rdfs:subClassOf rm:Output .
# For Scope
rm:ExperimentalTechnique rdfs:subClassOf rm:Scope .
rm:SubjectCoverage rdfs:subClassOf rm:Scope .
# For properties
rm:activity_location rdfs:subPropertyOf rm:hasScope .
rm:activity_subject rdfs:subPropertyOf rm:hasScope .
SPARQL queries in support of use cases

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rm: <http://example.org/stuff/ResearchModel#>.

# How much research output, and how much of each type is out there:
SELECT ?output_type (COUNT(?output) as ?total)
WHERE {?output_type rdfs:subClassOf rm:Output .
   ?output a ?output_type .
}
GROUP BY ?output_type

# Discover the chains of interrelated activities:
SELECT ?previous_activity ?current_activity
WHERE {?previous_activity rm:hasOutput ?output .
   ?output am:inputFor ?current_activity .}
Possible application: research provenance

<table>
<thead>
<tr>
<th>RA model class</th>
<th>Funding</th>
<th>Research per se</th>
<th>Results processing</th>
<th>Results dissemination</th>
<th>Results consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Research proposal</td>
<td>Award (grant)</td>
<td>Dataset</td>
<td>DDI record</td>
<td>DDI record or its manifestation</td>
</tr>
<tr>
<td>Output</td>
<td>Award (grant)</td>
<td>Dataset</td>
<td>DDI record</td>
<td>Web service</td>
<td>Feedback</td>
</tr>
<tr>
<td>Actor</td>
<td>Researcher candidate</td>
<td>Contributor (author)</td>
<td>Data archive</td>
<td>Dissemination service</td>
<td>Web service user</td>
</tr>
<tr>
<td>Effect</td>
<td>Researcher's department budget</td>
<td>Whatever is claimed in proposal</td>
<td>Economical effect of processing</td>
<td>Economical effect of distribution</td>
<td>Impact on further research</td>
</tr>
<tr>
<td>Condition</td>
<td>Funding body rules &amp; regs</td>
<td>Microdata regulation</td>
<td>Data processing guidelines</td>
<td>Data access regulation</td>
<td>Research purpose statement</td>
</tr>
<tr>
<td>Scope</td>
<td>Certain branch of science</td>
<td>Certain geolocations</td>
<td>National research</td>
<td>International research</td>
<td>Certain HASSET keywords</td>
</tr>
</tbody>
</table>
Collaborative curation of research data in “cloud of clouds”
The model selling points 😊

• Small
• Extendable
• Allows widely adopted RDFS Plus manifestation
• (Right) balance between simplicity and expressivity
• (Right) balance between modeller’s freedom and results interpretability
Use cases for applying the model

• Research provenance, navigation and contextualization
• Semantic analysis and annotation of domain-specific metadata (DDI, CSMD, …)
• Distributed discovery, curation, and re-use of the research information
• Long-term digital preservation
Thank you!