Core semantic model for generic research activity[€]

© Vasily Bunakov
Scientific Computing Department, Science and Technology Facilities Council,
Harwell OX11 0QX, United Kingdom
vasily.bunakov@stfc.ac.uk

Abstract

A simple research activity model is suggested that is agnostic to research domain and allows independent curation of the research information lifecycle by a variety of its stakeholders with a potential to further link individual activities into meaningful research provenance or research value chains. We consider the drivers for conceiving the model, its main aspects, an RDF manifestation of it, a particular business case for its application, and discuss its potential for future applications.

1 Introduction

Different stages of the research lifecycle in natural sciences as well as in social and economic research produce multiple data artefacts under control of different data management solutions and software platforms. (We use the term "data" here and there in a broad sense: not necessarily numeric data resulting from measurements but research proposals, software components, configuration files, electronic publications, etc.) Data curators working in a particular research domain tend to develop a specific metadata model that aims to cover the entire research lifecycle from the research inception to the research outputs dissemination. Such a metadata model quite often serves as a foundation for the design of the actual information systems and services. The example of a comprehensive metadata model for the research performed at large facilities like synchrotrons, powerful lasers or neutron sources is the Core Scientific MetaData model [5]; the example in social research is DDI-Lifecycle [7].

Proceedings of the 15th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2013, Yaroslavl, Russia, October 14-18 2013.

[€] This work is related to the ENGAGE project www.engage-project.eu and the projects of PaNdata collaboration www.pan-data.eu supported by the EU 7th Framework Programme for Research and Technological Development. The author would like to thank his colleagues in ENGAGE and PaNdata for their input for this paper although the views expressed are the views of the author and not necessarily of the projects.

Substantial effort of renown information experts has been spent in order to extend some established metadata models with new semantic features; the example in social research will be DDI semantic modelling ([8], [9]). The richness and the expressivity of metadata model that has evolved through decades can be considered a limitation that makes it harder to agree on what should constitute the "true" semantic representation, or what format of it should be a "canonical" one. Also the attempts to transform the entire domain-specific metadata model into semantic representation, and then offer it for common adoption and data linkage may contradict the social nature of Linked Data as its curation can be reasonably considered an incremental and opportunistic effort of multiple parties (as brilliantly illustrated by [1]).

This is not to say that semantic modelling of the entire research domain is not sensible or do not have a potential for implementation. Collaborative projects of a multinational scale such as PaNdata-ODI ([2], also see under [16]) consider semantic representation of the popular domain-specific metadata model [5] with the purpose of system integration. The motive for this consideration is that, despite the actual information systems in different research centres may be based on the implementations of the same generic metadata model and even on the same software platform for data catalogue [14], the practices of the catalogue configuration, the interpretation and the use of the model elements, and hence the actual semantics of these elements may vary dramatically. A common semantic layer, probably in the form of ontology, is considered then a viable architecture solution that should allow retaining the existing local practices of data cataloguing and at the same time, should give the IT teams an ability to meaningfully integrate distributed data and services.

That semantic layer, however, will require an inclusion into a certain best practices framework to sustain it through time [4], otherwise divergent business needs and business practices of the collaboration participants can make a thoroughly designed semantic model obsolete the next day after its implementation in a real IT solution. Keeping a comprehensive semantic model actual can be quite an expensive endeavour with substantial overheads on continuous business analysis and communication with multiple parties.

Another concern about the attempts of semantic representation of comprehensive metadata models is a tendency for them to reflect the information needs of

only a few types of the research lifecycle stakeholders: this is commonly Researchers and Data Archivists. The information needs of other stakeholders from Funding, Industry, or Education are often under-represented. To resolve this issue, one can take two approaches:

A) As a responsible information curator, conduct thorough business analysis of the research lifecycle stakeholders' types and their information needs then incorporate the knowledge acquired into a comprehensive model that, in order to be effective, should be validated by the stakeholders themselves (then, ideally, permanently amended).
B) Give different stakeholders a reasonable modeling means to express their role in the research lifecycle so that each of them becomes an information curator who cares about the quality and the actuality of her contribution into the shared pool of information.

The latter approach seems more adequate in the present situation when the advance of Linked Data principles allows various stakeholders to meaningfully model their part of information universe, also re-use the results of similar modeling effort made elsewhere.

We suggest a small but quite universal "core" model in the spirit of Linked Data principles [1] with low barriers for its adoption and use for semantic annotation of the research activity in different local information contexts, with their further inclusion into a global information context. We think that such a model should not focus on data but on common patterns of research activity observed in different research domains (for which we give examples further in this paper); various data then can be considered artefacts or "footprint" of different types of research activity.

2 Research activity model

2.1 Types and common patterns of research activity

Research lifecycles analyzed and structured by digital curators in the respective research domains can be a good source for discovering granular research activities and their interrelations. In this work, we consider two lifecycles: in facilities science[1] and in social research; they are most relevant to the projects which contributed to the development of our model ([11], [16]) and their respective research domains stay quite far apart so may help us with testing our model universality.

Lifecycle in facilities science that underpins CSMD model [5] includes the submission of a research proposal to the facility user office in order to get the facility resource for research (e.g. beam time on synchrotron); the further approval of the proposal by the facility's user office; experiment scheduling; conduct of the actual experiment with data collection; data storage; data analysis; and eventually publishing research results with record keeping for them. Beyond this lifecycle that is supported by facility itself, there is research funding activity, or research policy making, or the researchers' social communication that all can be considered elements of a larger "research value chain".

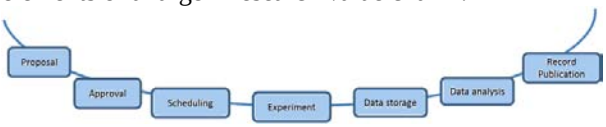

Figure 1. Research lifecycle in facilities science (as captured by CSMD model).

The lifecycle of social research that underpins DDI-Lifecycle model [7] includes the formulation of the study concept, further data collection, its processing, archiving, distribution, discovery, analysis, and repurposing. Funding, or policy making, or social communication, despite there are some placeholders for references to these types of activity – are again beyond the immediate scope of DDI.

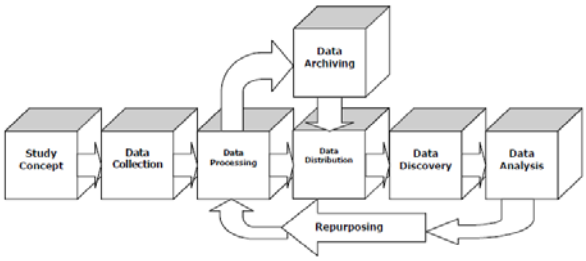

Figure 2. Research lifecycle in social science (as captured by DDI-L model).

Each activity yields certain outputs, e.g. in facilities science, the research proposal preparation results in the investigation (experiment) description, data analysis yields derived data etc. Previous activity may provide an input for other activity or give it a context, e.g. it is quite common for researchers to refer to the previous investigations (experiments) when they apply for a new investigation to be conducted at the same facility.

Despite there are similarities between the two aforementioned lifecycles and between the roles of stakeholders involved in them, there are differences, too. Even more differences come up if we consider context or scope of each research activity, or means for their description that are present in each model. As an example, in facilities science, the scope of experiment can be understood by considering what samples or chemical substances have been under investigation; in social research, it can be meaningful parameters describing the human audience which the study has been aimed upon. Not these details that may be different but the very presence of Context and Scope, as well as the Inputs and Outputs for the research activity, or Actors who perform it, or Effects of the research do represent a common pattern – very generic but universal

[1] For the sake of clarity, we use the term "facilities science" for the research performed on large-scale scientific instruments (synchrotrons, powerful lasers and alike) by visitor teams or individual researchers who obtain, via the application process, access to the common facility resource in order to conduct their experiments or observations, and to collect the resulting data.

across research fields.

These patterns are common not only across different research domains for the similar types of research activity (when we draw parallels e.g. between facility science Experiment and social research Study); this is also the case for different types of research activity within the same lifecycle, e.g. funding or data analysis or record publication have their Inputs and Outputs, their Actors, Effects, Context (Conditions) and Scope.

These basic patterns contribute to a reasonable model that should not be too burdensome for the respective stakeholders (or information specialists working for them) to apply, yet is expressive enough to promote the principles and best practices of Linked Data in various research domains. We consider a potential for such an application below in the section devoted to a particular business case; in the meanwhile, we are going to formally introduce the major aspects of a generic research activity, and suggest a practical RDF-based manifestation for them.

2.2 Generic research activity (research activity "cell")

We deem important the following aspects of a generic research activity:

Aspect	Description	Examples	
		Research per se	Research data analysis
Input	Something that is taken in or operated on by Activity	Previous research	Raw data
Output	Something that is intentionally produced by Activity	Raw data	Derived (analyzed) data
Scope	Something that Activity is aimed at or deals with	Sample properties	One or more experiments
Condition	Something that affects or supports Activity, or gives it a specific context	Scientific instrument	IT environment
Actor	Something or somebody who participates in Activity	Investigator	Data analyst
Effect	Something that is a consequence of Activity	Environment pollution	New software module

Schematically, the granular research activity can be represented by the following diagram:

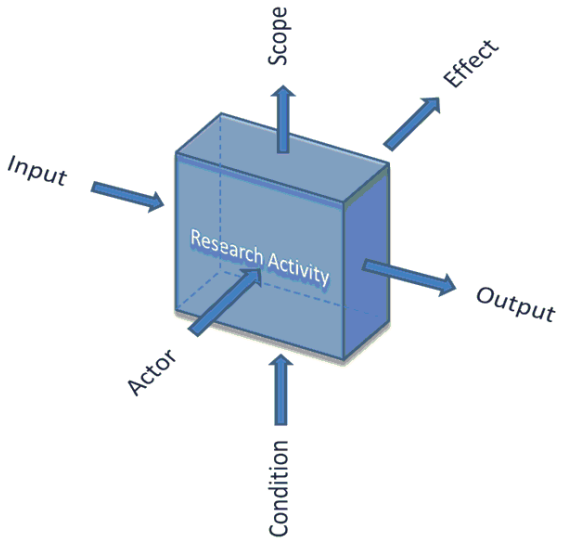

Figure 3. Research activity "cell".

Research activities can be combined as "cells" in chains where Output of one can be an Input to another but in fact, the model allows other sorts of links between activities. As an example, a piece of regulation such as data management policy can be an Output of one activity (policy making), and a Condition that affects another activity (research per se); a new software module that is a side Effect of a certain activity (data analysis) can be a non-human Actor that participates in other activity (e.g. automated indexing of experimental data). This shows that activity aspects in fact do not have "types": a modeler can use and combine them as dictated by the semantics of the respective subject area.

This view is inspired, to some extent, by SADT activity model [17] with its idea of combining activities into the hierarchy or a grid but is quite different by introducing some other activity aspects and not imposing their typization. Also SADT promotes a top-down approach to structured analysis and systems design when we suggest a bottom-up approach that allows combining the granular activities in more complex information structures.

Compared to other project-driven attempts to model research activity ([10], [15]) our model is going to be simpler, more universal, and deliberately aimed at semantic modeling of a granular activity rather than of the entire research lifecycle thus providing a "building block" for a more sophisticated information modeling as and when required.

2.3 RDF manifestation of activity model

The outlined model may imply different manifestations; we feel that one expressed in RDFS Plus (RDF Schema with a few OWL terms) has a good potential for adoption by information curators and implementation in real IT solutions. This paper Appendix suggests the

RDFS Plus manifestation of the activity model that can be extended by domain specific entities and properties. As an example, an information modeler in facilities science might want to extend the model as follows:

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix am: <http://example.org/stuff/ActivityModel#>.
@prefix rm: <http://example.org/stuff/ResearchModel#>.
# For Activities
    rm:Research rdfs:subClassOf am:Activity .
    rm:Experiment rdfs:subClassOf rm:Research .
# For Conditions
    rm:Condition rdfs:subClassOf am:Condition .
    rm:Regulation rdfs:subClassOf rm:Condition .
    rm:DataManagementPolicy rdfs:subClassOf rm:Regulation .
# For Output
    rm:Output rdfs:subClassOf am:Output .
    rm:Publication rdfs:subClassOf rm:Output .
    rm:Dataset rdfs:subClassOf rm:Output .
# For Scope
    rm:Scope rdfs:subClassOf am:Scope .
    rm:ExperimentalTechnique rdfs:subClassOf rm:Scope .
    rm:SubjectCoverage rdfs:subClassOf rm:Scope .
# For properties
    rm:activity_location rdfs:subPropertyOf am:hasScope .
    rm:activity_subject rdfs:subPropertyOf am:hasScope .
```

The user of the information system where the RDF data prepared according to our model is published can then use reasonable SPARQL requests to inquire for different aspects of research activities, e.g. trying to realize first 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 am:Output .
  ?output a ?output_type .
    }
GROUP BY ?output_type
```

or try to discover the chains of interrelated activities:

```
SELECT ?previous_activity ?current_activity
WHERE {?previous_activity am:hasOutput ?output .
     ?output am:inputFor ?current_activity .}
```

User may be familiar with just our activity model knowing very little about a certain research domain at start, then accumulating more and more knowledge through sensible incremental requests. In case the information modeler, in addition to our basic activity model, has followed good practices of data curation so that e.g. instances of Scope or Condition subclasses are not literals but dereferenceable URIs, the User will have even more opportunities of getting familiarized with the semantics of a particular research domain. When we tell of "User" we of course mean the software agents, too, as the prospect of employing them is a strong incentive for any semantic modeling.

2.4 Business case for semantic categorization and annotation of existing metadata

As we mentioned, it may not be easy to give birth to the semantic representation of a comprehensive metadata model because of its richness and complexity, and because of substantial overheads for communication among information curators who apply the model in different contexts. Another observation is that detailed metadata records may in fact represent different activities performed by different stakeholders of the research information lifecycle – while the records that in fact circulate in the information management solutions are focused on particular types of stakeholders only and support their specific roles in the first place. A certain stakeholder, e.g. Data Librarian or Data Archivist may claim that Her information management solution is focused on *data* in pursuit of some common interest when, in fact, the information management solution primarily supports this particular stakeholder specific *role* in the information lifecycle with only some types of other stakeholders well served.

As an example, DDI [7] suggests some means to model information about funding but European funding bodies are likely to use their own information systems, many of them based on CERIF standard [6]. So the richness and expressivity of DDI, as well as the actual information systems based on it are in fact aimed at researchers in social science and data archivists, not at funders who are likely to have their own information systems based on other metadata standards, and not at other types of stakeholders in Business, Education, or researchers in other research domains.

We feel that it will be more productive to admit this natural attitude of the information management solutions and their owners to cater for only one or a few roles; it may be better to provide a reasonable means to model different roles and their activities on a granular level than try to capture an elusive information context in more and more complex versions of a comprehensive semantic model. If we take the existing records in a certain rich metadata format, this approach results in categorization and annotation of the entire metadata records with other metadata based on a smaller but semantically meaningful and universal information model – like our activity model.

Let us see how our core semantic model may serve DDI metadata categorization and annotation.[2] The analysis shows that one DDI record typically represents different types of research activity:

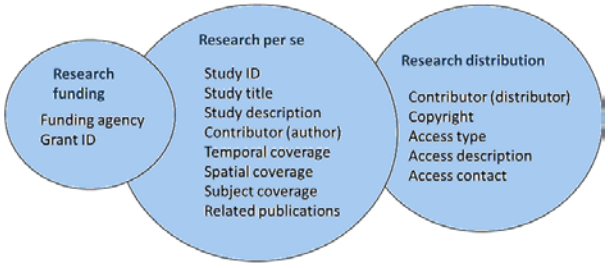

Figure 4. Research activities represented by a DDI record.

As we have identified different types of research activity, we can model them accordingly; we can also

[2] This approach was applied to DDI records harvested from the UK Data Archive and GESIS archive ([18], [13]) in the interests of the ENGAGE project [11] and was communicated in [3] as a prolegomenon to the generic model that we are presenting now.

identify specific Actors (Funding Agency, Author, Distributor), activity Outputs (Publication, Dataset), Scopes (Spatial Coverage, Subject Coverage) and Conditions (Copyright, Access Terms). Different granular activities will be modeled then with different amount of detail but we can enrich them with data from other information systems: for research funding – through funding agency portals, for research – through the project and the individual investigators' Web pages. This information enrichment should ideally be done by the Actors of the respective Activities (Funding, Research per se, Distribution) as they best understand the information context and the semantics of their business.

Our activity model then should allow curating the data and data context (metadata) in a distributed manner, and the combination of granular activities in sensible information context chains. This should eventually give us a more dispersed but a more complete description of the research discourse for a particular Study – more complete if compared to what the Data Archivist deemed valuable to capture and describe in a DDI record for the same. Our core model then serves as a "glue" to support the common information context and facilitate the interoperability of different digital curation frameworks that are operated by different Actors in support of their own Activities.

The existing well curated archives of DDI records can be considered then a valuable "fuel" to support the launch of the research discourse "Web" or "grid". The role-centric nodes of it will be performing their part of digital curation, with sharing its results via simple and commonly understandable semantic model that can be interpreted not only by data archivists or researchers in social science but by various stakeholders from other research domains, or business, or education, or policy making.

2.5 Conclusion

We outlined the motivation for why a simple model would be valuable for the semantic representation of a generic research lifecycle. We introduced the major aspects of the model, suggested an RDF manifestation for them and showed how the domain-agnostic requests might work for information discovery. We then considered a particular business case of applying the model to the existing rich metadata records in social science but there are more promising cases to consider.

One of the immediate candidates is facilities science with its CSMD metadata [5] that we already mentioned. The diverse business practices for using the existing mature data management solutions based on CSMD model [14] may become a barrier to the meaningful sharing of facilities science data as Linked Data. Our model then may be of help for the re-engineering of the existing data archives in spirit of Linked Data and Semantic Web principles, through semantic annotation of the CSMD metadata records (which may involve some decomposition, too, similarly to what we demonstrated for DDI metadata).

Another prospective area where we think our model may prove to be valuable is long-term digital preservation with its two well-known problems of the accountable data provenance and of the meaningful data representation for the future (and changing) community of data consumers. The ability of our model to combine individual data curation activities into the traceable chains of them, as well as its very focus on the Activity (with data being an artefact or footprint of it) may contribute to the satisfactory resolution of the data provenance problem. The model's data discovery capabilities based on standard information requests and profiles of them when it is enough for the User to be familiar with our basic semantic model in order to start the incremental knowledge discovery – may contribute to the meaningful data representation.

Also we find the multi-disciplinary and *distributed* curation, discovery and re-use of the research information to be in high demand; it is already in the agenda of a few actual European projects (see under [11], [12], [16]) and it is reasonable to expect more of them to come. The domain-agnostic nature of our model, as well as its very manageable core size and expandability where required let us hope for its application in some of the existing and future e-infrastructure initiatives.

3 Appendix: RDFS Plus manifestation of the activity model

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix am: <http://example.org/stuff/ActivityModel#> .

############### Core entities of Activity model ###############

# Comments are based on the Oxford dictionary, with some generalization or amendment where appropriate

am:Activity rdf:type rdfs:Class ;
        rdfs:label "Activity" ;
        rdfs:comment "Something that Actor does, or has done,
            or is going to do, or can do" .
am:Input rdf:type rdfs:Class ;
        rdfs:label "Activity Input" ;
        rdfs:comment "Something that is taken in or operated on
            by Activity" .
am:Output rdf:type rdfs:Class ;
        rdfs:label "Activity Output" ;
        rdfs:comment "Something that is intentially produced
            by Activity" .
am:Actor rdf:type rdfs:Class ;
        rdfs:label "Activity Actor" ;
        rdfs:comment "Something or somebody who participates
            in Activity" .
am:Effect rdf:type rdfs:Class ;
        rdfs:label "Activity Effect" ;
        rdfs:comment "Something that is a consequence
            of Activity" .
am:Condition rdf:type rdfs:Class ;
        rdfs:label "Activity Condition" ;
        rdfs:comment "Something that affects or supports
            Activity, or gives it a specific context" .
am:Scope rdf:type rdfs:Class ;
        rdfs:label "Activity Scope" ;
        rdfs:comment "Something that Activity is aimed at
            or deals with" .
```

```
########## Core properties of Activity model ##########

# am:hasInput or am:inputFor
# links Activity to its Input
am:hasInput owl:inverseOf am:inputFor .

# am:hasOutput or am:outputOf
# links Acttivity to its Output
am:hasOutput owl:inverseOf am:outputOf .

# am:hasActor or am:actorFor
# links Activity to its Actor
am:hasActor owl:inverseOf am:actorFor .

# am:hasEffect or am:effectOf
# links Activity to its Effect
am:hasEffect owl:inverseOf am:effectOf .

# am:hasCondition or am:ConditionFor
# links activity to its Condition
am:hasCondition owl:inverseOf am:ConditionFor .

# am:hasScope or am:ScopeOf
# links Activity to its Scope
am:hasScope owl:inverseOf am:scopeOf .
```

References

[1] Tim Berners-Lee. Open, Linked Data for a Global Community. A talk given on Gov 2.0 Expo, Washington, DC, 26 May 2010. http://www.gov2expo.com/gov2expo2010/public/schedule/detail/14247

[2] Juan Bicarregui, Vasily Bunakov, and Michael Wilson. PANdata international information infrastructure for synchrotrons: opportunity for collaboration. Presentation on the 19th Russian Synchrotron Radiation Conference (SR-2012), Novosibirsk, Russia, 25-28 June 2012. http://epubs.stfc.ac.uk/work-details?w=63074

[3] Vasily Bunakov. Semantic categorization of DDI metadata. Presentation on the 4th Annual European DDI User Conference (EDDI12), Bergen, Norway, 03-04 Dec 2012. http://epubs.stfc.ac.uk/work-details?w=64315

[4] Vasily Bunakov and Brian Matthews. Data curation framework for facilities science. In Proceedings of DATA 2013: the 2nd International Conference on Data Management Technologies and Applications, p.211-216, Reykjavík, Iceland, 29-31 July 2013.

[5] Brian Matthews et al., 2012. Model of the data continuum in Photon and Neutron Facilities. PaNdata ODI, Deliverable D6.1. http://pan-data.eu/sites/pan-data.eu/files/PaNdataODI-D6.1.pdf

[6] Common European Research Information Format. See under www.eurocris.org

[7] Data Documentation Initiative – Lifecycle Specification. http://www.ddialliance.org/Specification/DDI-Lifecycle/

[8] Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Web. Schloss Dagstuhl, September 11 – 16, 2011. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=11372

[9] DDI Lifecycle: Moving Forward. Schloss Dagstuhl, October 21 – 26, 2012. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=12432

[10] DARIAH-EU: Digital Research Infrastructure for the Arts and Humanities. http://www.dariah.eu/

[11] ENGAGE: An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens. http://www.engage-project.eu/

[12] EUDAT: European Data Infrastructure. http://www.eudat.eu/

[13] GESIS - Leibniz-Institut für Sozialwissenschaften. http://www.gesis.org/

[14] ICAT project. http://www.icatproject.org/

[15] Infrastructure for Integration in Structural Sciences (I2S2) Project. http://www.ukoln.ac.uk/projects/I2S2/

[16] PaNdata: Photon and Neutron Data Infrastructure. http://pan-data.eu/

[17] Structured Analysis and Design Technique. http://en.wikipedia.org/wiki/Structured_Analysis_and_Design_Technique

[18] UK Data Archive (for social sciences and humanities). http://data-archive.ac.uk/