Investigation as a member of research discourse[€]

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

Abstract

Investigations are specific intellectual entities that circulate in large research facilities with shared access by multiple research teams; investigations have some common features with research papers (publications) and can be included in citation networks. We consider different approaches to modelling the relations between research papers and investigations and discuss opportunities for matching these two members of common research discourse. The analysis undertaken can be of interest for research centres that consider information services based on data and publications contextualization.

1 Introduction

The journal articles, e-prints, reports and other similar artefacts that irrespective of their physical manifestation can be seen as derived from their paper-based "document" ancestors are the well-established means of research communication and a popular aide for tracking the state and the trends of research discourse. The "papers" have clear identity, allow review (of different kinds) and participate in citation networks; this supports performing the aforementioned functions of the quality research communication and measurable research tracking; this also makes "papers" valuable intellectual entities worth capturing in library catalogues, and worth sharing via advanced information services.

We suggest that other type of intellectual entities, *investigations*, have essential features similar to the document-like entities hence are the natural candidates to supplement "papers" as valuable members of research discourse. We consider the types of relation between the document-like and investigation entities, and take a look at the simultaneous circulation of them in our own research domain of experimental science utilizing large research facilities: neutron sources, synchrotrons, and powerful lasers shared by multiple researchers (visiting scientists).

Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13-16, 2014.

[€] This work is related to 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 PaNdata for their input for this paper although the views expressed are the views of the author and not necessarily of the collaboration.

2 Facilities research lifecycle and data modelling

2.1 Facilities science landscape

Research facilities can be thought of as well-equipped hubs where research teams or individual researchers come to perform their experiments on their own samples. The research facility core is typically represented by a unique scientific instrument: a particle accelerator, a neutron source, a powerful laser, a telescope, or a supercomputer that allows detailed simulation of natural phenomena, or by a few such instruments that offer researchers different research techniques. The examples include European Synchrotron Radiation Facility (www.esrf.eu), neutron source in The Institut Laue-Langevin (www.ill.eu), Siberian Synchrotron and Terahertz Radiation Centre (http://ssrc.inp.nsk.su/CKP/eng/) or the future Extreme Light Infrastructure (www.eli-beams.eu).

Research conducted in facilities bears characteristics of "big science" such as a long-term capital investment, permanent support staff, scalable computing infrastructure; and "bench science" with individual scientists and small research teams that may have specific and short-time research goals. The user community of European facilities counts tens of thousands scientists who pursue different applications: crystallography reveals the structures of proteins important for the development of new drugs; neutron scattering identifies stresses within engineering components such as turbine blades, and tomography can image microscopic details of biological tissues ([1]).

A business model for user research on large facilities that emerged a few decades ago has been influenced by the advances in instrumentation and data analysis that are now more automated and more user friendly than in early days of facilities a few decades ago. This has led, among other effects, to a lesser significance of the instrumentation "gurus" ([2]), and to the emergence of specific services for research and industry that allow users sending their samples for remote investigation according to one of the service plans ([3]).

Yet the facilities business model has proved to be effective and is a foundation for a specific research lifecycle, and for specific information modelling and information services in support of it.

2.2 Generic research lifecycle

Despite the variety of facilities instruments and experimental techniques, the following distinct stages are typical across facilities and thus represent a generic facilities lifecycle:

- **Research Proposal:** the facilities are often oversubscribed so the researcher (investigator) should justify the value of her research and the suitability of a particular experimental technique
- **Approval Process:** multilateral assessment by the facility, including risk assessment (as the experiment may involve hazardous materials or techniques)
- **Experiment Scheduling:** allocation of the time slot within a facility operating cycle, and registration of all visitor scientists
- **Series of Experiments (that altogether constitute Investigation with the proclaimed goals):** the user will bring samples, and sometimes an additional equipment to the facility, calibrate the experimental environment and actually take measurements
- **Data Archiving:** facilities offer high-throughput data collection and archiving services; archiving of raw data collected in the facility data storage is often a policy requirement
- **Data Analysis:** it can be done through multi-layer computing environment where some tools are offered by facilities, and others applied by scientist individually
- **Results Publication:** journal articles and alike; facilities often require the visitor scientist to report back on any publications derived from the experiments

This generic lifecycle is illustrated by Figure 1.

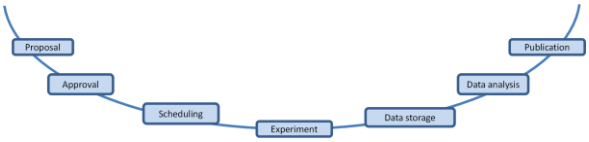

Figure 1. Facilities research lifecycle.

2.3 Data modelling effort so far

Facilities collect raw experimental data in a variety of formats yet there is a movement towards unification best represented by NeXuS standard and community around it (www.nexusformat.org). There are of course data checks and data replication services, as well as some recent attempts to form and curate archival packages according to OAIS reference model ([4]).

The aforementioned generic lifecycle gave birth to the rich CSMD metadata model ([5], [6]) which is implemented, with some modifications, in the popular ICAT software platform ([7]).

Some facilities started assigning persistent identifiers to datasets ([8]) and there is a recent effort of having persistent identifiers for other aspects of facilities research such as instruments or experimental techniques ([9]).

The promotion of the research idea through the facilities lifecycle has inspired the concept of Research Objects for facilities science ([10]) that acquire more and more detail whilst the investigation proceeds from its conceptual stage through the experiment to the research paper and associated artefacts.

An interesting recent development is the intention of some facilities to start publishing the descriptions of the approved research proposals (grants) – that are the "cores" to the future investigation entities – on the national research portals, e.g. ISIS neutron and muon source (www.isis.stfc.ac.uk) intends to publish the descriptions of all approved proposals on the UK common gateway to publicly funded research (http://gtr.rcuk.ac.uk/). The internal representation format for these entities is going to be CERIF (see under www.eurocris.org) that is widely used in the European grant information systems.

3 Research data in research discourse

3.1 The modes and purposes of sharing research data

The earlier mentioned NeXuS format, Research Objects and persistent identifiers for data present three different modes of sharing research data.

NeXuS file includes both data and data context (metadata) and thus offers research result as a "package" that can be interpreted by other researchers – or the same research team in future – with the help of format-compatible software. It is a responsibility of the "package" creator to embed all essential information in there; the boundaries of information context are very well defined (it is literally one data file).

Research Objects suggest the enrichment of information according to a specific model while the intellectual entity moves through the research lifecycle; this implies that there is a "creator" to the model and the "curator" of intellectual entity on each phase of lifecycle; the boundaries of intellectual entity are more flexible (it may be an aggregation of various components) but are still well-defined.

The supply of nothing more but persistent identifiers for data, perhaps associated with some moderate contextual description (metadata), implies the paradigm of "open world" where intellectual entities can be deliberately constructed by various agents, hence there are no clear (predefined) boundaries to the entities, and virtually everyone can be considered a data "curator".

Sharing data or information, however, is not the end in itself and can be considered a means to empower research discourse, to supply some intellectual entities into it. So quite often, when people speak of "research data" they actually mean intellectual entities where data may be just a component, or something associated with

a “quantum” of research discourse.

This can be illustrated by observations over DataCite (www.datacite.org) – a platform that proclaimed goal is supplying data with dereferencaeble persistent identifiers (well-formed DOIs). The data centres who actually use DataCite in fact tend to assign DOIs not to datasets but to “quantums” of research discourse, e.g. to doctoral theses (that may of course contain some data but is not the data per se). In case of facilities science, we observe that DataCite DOIs are in fact dereferenceable to the landing Web pages that contain descriptions of *investigations* which are, as we explained it earlier, the series of experiments performed with a certain research goal on the assigned instrument within a dedicated timeslot.

So when a researcher cites “data” via DataCite DOI, she in fact quite often cites an intellectual entity – which can be a paper or something else, e.g. event (such as an earthquake) in geophysics, or investigation in the case of facilities science. [1] This attitude towards “data” DOIs assignment is only natural as what researchers tend to cite may not be “data” per se but certain identifiable elements of research discourse.

3.2 The place of investigation and the place of data in facilities research discourse

Investigation as an intellectual entity bears some features that are common with traditional research paper. Indeed, an investigation proposal is peer-reviewed; investigation can be cited from papers by the well-formed DOI and from other investigations, too, as when a researcher submits proposal, she refers to the relevant past publications and past investigations.[2]

The Figure 2 illustrates provenance relations between investigations and research papers that are a foundation for appropriate “citations”.

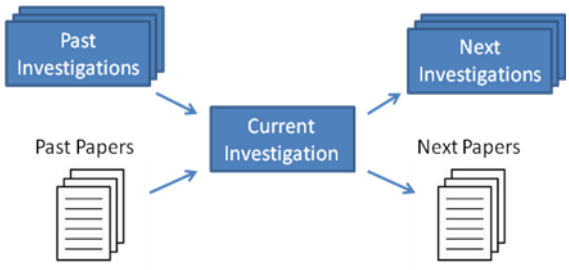

Figure 2. Research provenance chain.

Similarities between investigation and research paper as intellectual entities are summarized in the Table 1.

Feature / aspect	Publication (research paper)	Investigation
Is an intellectual entity	Yes	Yes
Is a subject of peer review	Yes	Yes (via proposal approval)
Can cite all significant intellectual entities of research discourse	Yes	Yes
Citation chains exist (steps of discourse observed	Yes	Yes
Universal identifiers available	Yes	Yes

Table 1. Common features of investigations and research papers.

Looking into what intellectual entities can refer to what other intellectual entities (with the inclusion of datasets and software – which may or may not bear a clear identity) suggests the asymmetry in the direction of references so that e.g. a research paper can cite a dataset but not vice versa:

References (“from” row “to” column)	Paper	Investigation	Data set	Software
Paper	Yes	Yes	Yes	Yes
Investigation	Yes	Yes	Yes	Yes
Dataset	No	Yes	Yes	Yes (e.g. simulation)
Software	Yes (e.g. to paper about algorithm)	No	Yes	Yes

Table 2. Cross-references of intellectual entities.

[1] Examples of dereferenceable “data” DOIs that in fact resolve in investigation or research paper descriptions: http://dx.doi.org/10.5286/ISIS.E.24066298 http://dx.doi.org/10.5167/UZH-27029

[2] Looking into the ICAT database for ISIS facility indicates the existence of investigation “chains” when the next investigation refers to the previous one, with as many as four investigations in a row undertaken in the last 10 years.

In fact, research discourse in facilities science splits into the two distinctive layers that can be called "research per se" and "data management"; this is illustrated by Figure 3.

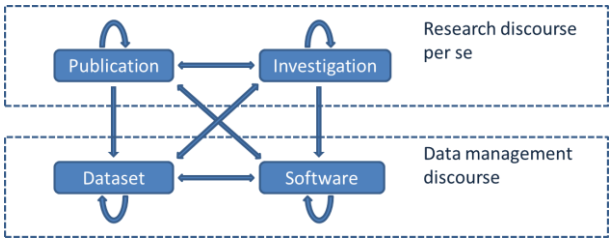

Figure 3. Directions of typical references and two layers of research discourse.

The two layers only loosely interact with each other and the bottom one can be considered a service layer in support of the top one, despite recent attempts to promote a view that information departments can play a role of data curation units, hence expanding their remit from the mere support of information technology to catering for richer tastes of researchers interested in semantic representation of information and in its sensible reuse ([11]).

3.3 Problems, challenges and opportunities

The above analysis contributes to modelling of research discourse in facilities science with the suggestion that data and software should play a modest (supportive) role compared to research papers and well-defined investigations. Different information models that can be applied to the same facilities research discourse. One of them is the model based on Research Objects ([10]) that suggest the "enrichment" of the core Investigation entity while it moves down the facilities research lifecycle illustrated by Figure 1 – turning into a rich aggregation of data, data context (metadata), and software. Another view is seeing research discourse as "grid" composed of provenance chains similar to that in Figure 2; the Research Activity model ([12]) offers a basic semantic means to support this view.

Irrespective of what of the two models we adhere to, they are likely to use the same techniques, e.g. for matching research papers with investigations.

One problem here is that, despite it is a requirement of facilities to submit the "input" to the investigation proposal and then the "output" of it in terms of research papers that led to the idea of the experiment, or have been resulted from it – there is no good curation of these bibliographic records, or a clear requirement for their format. On the other hand, when the institutional library eventually and independently collects the facility output in the form of research papers, they do it in a systematic way with good coverage and according to the best cataloguing practice but there is no record of the investigation that the paper has been resulted from as there is no requirement to capture it in the bibliographic record, also the investigation is often mentioned only implicitly in the paper. So if we want more context for the research papers and for the investigations, there is a task of matching bibliographic records coming from facilities User Office (the unit that looks after investigations lifecycle) and those in the institutional library catalogues.

To estimate the viability of automated techniques, we tried to match the bibliographic records for the papers that were the "input" to the investigations performed on ARGUS muon spectrometer.[3] We managed to visually identify the small number of the well-formed bibliographic records in the institutional repository that for sure match the corresponding poorly-formed ARGUS bibliographic records. We then applied different modifications to the ARGUS records in combination with measuring the Levenstein distance ([13]) between them and those in the library catalogue.

The first experiments suggest that bibliographic records from two systems: ePubs which is the institutional papers repository and ISIS ICAT which is the data catalogue supported by ISIS neutron and muon facility, can be successfully matched if we measure Levenstein distance between modified bibliographic records. A particular pretty simple technique could be the extraction and normalization of the numeric components from the bibliographic record (volume, pages and year), measuring distances between such normalized extracts – in effect, between two strings with only numbers in them – then playing with the threshold (the particular Levenstein distance) that allows to distinguish between matches and non-matches. This technique was tried out via bespoke Java software module and is illustrated by Table 3.

ICAT Reference	ePubs reference	Levenstein distance between full bibliographic references	Levenstein distance between "numeric" parts	Levenstein distance between "numeric" parts with the year normalized and the last page removed
Pratt et al, Phys. Rev. Lett. 96, 247203 (2006)	Phys Rev Lett 96 247203 (2006)	17	0	0
Lancaster et al, Phys. Rev B73, 020410(R) (2005)	Phys Rev B 73 020410 (2006)	24	1	1
Blundell and Pratt, J. Phys.: Condens. Matter 16, R771 (2004)	J Phys Condens Matter 16 R771-R828 (2004)	30	3	0

[3] http://www.isis.stfc.ac.uk/instruments/argus/argus6461.html

M.T.F.Telling and S.H.Kilcoyne, Electron transfer in dextran, J. Phys.: Condens. Matter 19 No 2 (17 January 2007)	J Phys Condens Matter 19 2 026221 (2007)	81	6	6
J Tomkinson and M.T.F Telling, Ammonium ions in alkali metal halide crystals: Tunnelling and spin relaxation, PCCP 2006 8 38 4434	Phys Chem Chem Phys 8 4434-4440 (2006)	113	12	5

Table3. Matching bibliographic records in ICAT data catalogue and ePubs papers repository (ARGUS case).

The technique tuning, including the measurements of precision and recall, should be done with the larger numbers of bibliographic records; there is about a thousand records in ICAT data catalogue that have bibliographic components – candidates for matching them with bibliographic records in ePubs papers repository. Yet it has to be understood that mere matching bibliographic records is just the first step in what we aspire to: a reasonably automated technique for linking investigations to research papers in situations where there are no bibliographic records catalogued for investigations, only investigations textual descriptions and other metadata.

There are more than ten thousand papers in ePubs repository that are marked up by the librarians as having relation to ISIS neutron and muon facility with no indication which investigation (series of experiments) or instrumental work they actually relate to. For the majority of these papers, there are no corresponding bibliographic records in the facility investigations catalogue; hence other techniques are required to match the papers to investigations. We consider decomposition of, on one hand, the bibliographic records from ePubs institutional repository and, on the other hand, the investigation descriptions from the ISIS investigations database into the corresponding elements, then looking into distances between elements with the further aggregation of them into sensible metrics. The analysis of bibliographic records and investigation descriptions suggests the following elements as the candidates for mutual mapping:

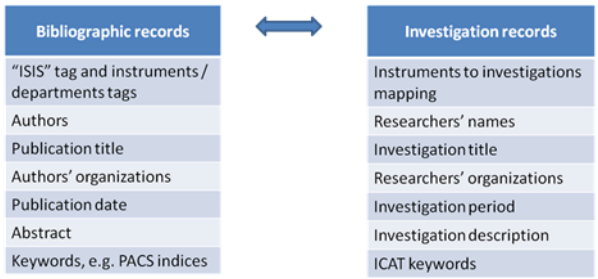

Figure 4. Mapping research papers bibliography to investigations metadata.

The mentioned massive of records in the ISIS ICAT data catalogue (about a thousand of them) – for which the association with ePubs papers catalogue can be established via the earlier outlined bibliographic records matching technique – can be used for the validation of automated matching between investigation metadata records and (more than ten thousand) bibliographic records for all ISIS instruments. Then validation by the researchers themselves will be required, as well as some technical means in support of that validation – such as online polls.

Another opportunity for the validation of the investigations-to-publications matching technique will be looking into descriptions of research proposals (grants) in the research information portals. For ISIS facility, it will be Gateway to Research portal (http://gtr.rcuk.ac.uk/) that is about to start collecting investigation proposals in a systematic manner so that sometime after the investigations are over, they will be supplemented by the submission of research papers resulted from them. It will be possible then to use the newer investigations accompanied by papers resulted from them (as submitted by the researchers themselves) for the calibration of the automated matching technique that can be applied to the large corpus of past investigations and research papers.

Validated via two independent sources of bibliography: ePubs institutional repository and (forthcoming) records in the Gateway to Research portal, the automated matching technique may become a useful tool for research contextualization and for enrichment of the existing records in publications and data catalogues.

Apart from matching research papers with investigations, an interesting theme for further research could be looking into the cases of "indirect citations" when (see Figure 2) one research paper does not directly cite another one but there is an identifiable connection from one to another through the intermediary investigation; or the similar consideration from the investigations network perspective where one investigation does not explicitly refer to another but they are in fact connected through the intermediary research paper(s). Discovering these sorts of "indirect citations" may contribute to the development of alternative metrics for measuring research output, in addition to traditional paper citation metrics.

4 Conclusion

Our analysis indicates that Investigation in facilities science is an intellectual entity that has a clear identity, is involved in structured information exchange and bears some essential features similar to traditional research papers. There are various opportunities for the information modelling and for the formation of links between investigations and other intellectual entities, namely research papers that can be either an input to the investigation, or an outcome of it.

This study can be considered an analysis and a roadmap that precede the scalable experiments on the information contextualization in the domain of facilities science. It is also a call for information practitioners to share their views on the research information contextualization and on the role of various intellectual entities in their research domains, as the popular notion of "data" and its widely accepted importance may sometimes misrepresent the actual content of research discourse where other domain-specific intellectual entities could be more appropriate for sensible information management and for measuring research output.

References

[1] Vasily Bunakov, Brian Matthews and Catherine Jones. Towards the Interoperable Data Environment for Facilities Science. A chapter in "Collaborative Knowledge in Scientific Research Networks" (AKATM book series). In press by IGI Global.

[2] J.Mesot. A need to rethink the business model of user labs? Neutron News, 2012, 23 (4), 2-3.

[3] S.J.Coles and P.A.Gale. Changing and Challenging Times for Service Crystallography. Chemical Science, 2012, 3 (3), 683-689.

[4] Reference Model for an Open Archival Information System. CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

[5] B.Matthews et al. Using a Core Scientific Metadata Model in Large-Scale Facilities. 5th International Digital Curation Conference, London, UK

[6] B.Matthews et al. 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

[7] D. Flannery et al. ICAT: Integrating Data Infrastructure for Facilities Based Science. In e-Science: Fifth IEEE International Conference on e-Science.

[8] Michael Wilson. Meeting a scientific facility provider's duty to maximise the value of data. Talk in DataCite Summer Meeting, Digital Research Data in Practice (DataCite2012), Copenhagen, Denmark. http://purl.org/net/epubs/work/62852

[9] PaNKOS: Proton and Neutron Knowledge Organisation System. www.purl.org/pankos

[10] B.Matthews et al. Investigations as research objects within facilities science. In 1st Workshop on Linking and Contextualizing Publications and Datasets, Malta, September 26th, 2013. http://purl.org/net/epubs/work/11912059

[11] Vasily Bunakov and Brian Matthews. Data Curation Framework for Facilities Science. In 2nd International Conference on Data Technologies and Applications, Reykjavik, Iceland, 29-31 Jul 2013, (2013): 211-216. http://purl.org/net/epubs/work/10938269

[12] Vasily Bunakov. Core semantic model for generic research activity. In 15th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, 14-17 Oct 2013, CEUR Workshop Proceedings (ISSN 1613-0073) 1108 (2013): 79-84. http://ceur-ws.org/Vol-1108/paper10.pdf

[13] Владимир Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академии Наук СССР (in Russian) 163 (4): 845–8, 1965.
Appeared in English as: Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (8): 707–710, 1966.