Persistent Identifiers for Facilities Research: Current Practices and Opportunities

© Vasily Bunakov © Simon Lambert
© Brian Matthews
Science and Technology Facilities Council, Harwell Campus,
United Kingdom
vasily.bunakov@stfc.ac.uk simon.lambert@stfc.ac.uk
brian.matthews@stfc.ac.uk

Abstract. The paper reports on the ongoing effort to define the scope and practical cases for the use of persistent identifiers in research organizations that operate large-scale facilities.

Keywords: persistent identifier, facilities research, community of practice, FAIR principles

1 Introduction

Facilities science or "lab science" is a branch of research performed by visitor scientists on large-scale scientific instruments: synchrotron radiation and neutron sources, powerful lasers, telescopes, or supercomputers. Facilities science business model and research lifecycle are similar across different instruments and geographical locations, and involve extensive management of data and other information artefacts [1], [2].

The progress of information technology makes it possible to mint and manage many types of persistent identifiers for a variety of uses beyond traditional Digital Object Identifiers (DOIs) for research papers. The use of persistent identifiers is often considered in the context of Open Science as a practice in support of FAIR principles [5] that aim to ensure that research results are Findable, Accessible, Interoperable and Reusable.

This paper outlines an ongoing effort to define the scope and practical cases for the use of persistent identifiers in facilities research as a contribution to FREYA project [3] that is taking over from the completed THOR project [4]. We first explain the current popular uses of persistent identifiers in facilities research, then indicate opportunities for the adoption of new types of persistent identifiers, or new use cases for their application.

2 Current uses of persistent identifiers in facilities research

2.1 Persistent identifiers for research papers

The most well established use of persistent identifiers (PIDs) in facilities research context is assigning them to research papers published by visitor scientists and/or personnel who support and operate facilities instruments. Digital Object Identifier (DOI) is possibly the most popular type of a PID for research papers and it can be assigned by different parties: journal publishers (typically issued via CrossRef [26]), or open repositories such as Zenodo [9].

DOI is not the only PID for papers that is actually in use, there are also:

- Persistent URLs (PURLs) [10] used by institutional publication repositories such as ePubs [11];
- arXiv.org IDs [14] that are assigned to items in this widely used pre-print service;
- institution-specific technical report identifiers;
- Handle identifiers [12] that are used in common e-infrastructure services such as EUDAT B2SHARE [13] and in some institutional repositories.

Many types of these PIDs can be used for both peer-reviewed papers and for all kinds of "grey literature". No type of PID, including DOI (e.g. minted by Zenodo [9] through a self-publishing process) is in itself an indicator of research quality or reproducibility, so other factors should be taken into account to judge on the paper compliance to FAIR principles. However, having a PID assigned to publication should give a level of assurance of persistence and integrity, and thus makes it citable and accessible. Therefore, PIDs for research papers and other research outcomes can be considered facilitators for at least the "F" – Findability and "A" - Accessibility, as the first steps towards FAIR principles implementation.

2.2 Persistent identifiers for investigations

Investigation is a generalization of a notion of an experiment performed on a large-scale research facility. Investigation may include one or more experiments, with some of the experiments potentially used for instrument calibration and other experiments used for the purposeful data acquisition.

Investigation time granted by facilities can be considered a non-monetary form of a research grant. Facilities typically assign unique identifiers to investigations; these IDs are not universal but facility-specific. If we extend the analogy of the investigation time award to research grant, an investigation ID is then similar to a grant ID assigned by funding agencies.

Some facilities such as STFC ISIS neutron and muon source [18] assign persistent identifiers to

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

investigations, using DOIs minted through the Application Programming Interface of the DataCite service [16].

Each DOI assigned to an investigation is resolved in a landing Web page supported by the facility on its own Web server. The DOI can be assigned and the landing page for it created right after the time slot is granted to a visitor scientist and before the conduct of the actual experiments. The landing page is then populated with metadata collected from the research proposal managed by a facility-specific proposal system. Once the investigation is actually conducted and experimental data is collected, the landing page is supplied with a link to the data holdings, which may be restricted for an embargo period to the scientists who actually performed the experiments.

Systematic assignment of DOIs to investigations, as well as having other structured metadata for them, make the circulation of investigations in research discourse in many respects similar to the circulation of research papers [7]. If the practice of minting DOIs for investigations becomes universal across facilities, this resemblance of investigations to universally citeable research papers will further grow.

2.3 Persistent identifiers for data

Data collected by Facilities is "raw" experimental data with, typically, no persistent identifiers assigned to it. Having this said, the data can be sometimes indirectly accessed (and indirectly cited) via the respective landing page and the DOI that resolves in that landing page. As an example, ISIS neutron and muon facility [18] publish DOIs for its investigations, with links to data archive from the DOIs landing pages.

Some other facilities publish datasets on the Web or in the FTP archives [15] in which case URLs, if stable and consistently minted, can be used as persistent identifiers for data.

Facilities' practice when there is no directly resolvable persistent identifiers for data, or they are URLs at best, is not unique. Even in cases when a dedicated service for data citation such as DataCite [16] has been created, with certain quality assurance and governance mechanisms for managing PIDs, the actual practices of minting these PIDs vary significantly across participating data centres. Information entities behind "data" PIDs are often not data per se but other information artefacts; a multi-aspect analysis of what is actually being published under the disguise of DataCite "data" DOIs is provided in [8]. One natural reason for this is that data assets often belong to "IT discourse" whilst minting PIDs and circulating them, including for the purpose of citation, makes more sense for information artefacts belonging to "research discourse" [7].

2.4 Persistent identifiers for researchers

ORCID [17] provides PIDs for identifying researchers, with its most prominent role in designating authors of publications, which in part stems from the fact that some quality publishers require ORCIDs when submitting a manuscript. With close to five million IDs issued, ORCID is becoming much more than widespread than other schemes for identifying researchers, such as the proprietary ResearcherID [30].

Facilities do mint their own identifiers for visitor scientists but these identifiers circulate only within a local proposal system. Some facilities, such as Oak Ridge National Laboratory (ORNL) [19] and Argonne National Laboratory (ANL) [20], have started asking visitors to submit their ORCIDs along with their research proposal, in hope that this will eventually allow a better mapping of research outputs such as published articles to the facilities that supported the research.

3 Emerging uses of persistent identifiers in facilities, and opportunities for the advanced uses of existing PID types

3.1 Persistent identifiers for instruments

One centre of expertise about emerging practices and recommendations for minting and using PIDs for facilities instruments is ORCID User Facilities and Publications Working Group [21] with main contributions from information specialists in American large-scale research facilities, such as the aforementioned ORNL and ANL. The interim results of this working group were reported in PIDapalooza workshop in January 2018 [22]. These largely focussed on popularization of ORCID identifiers for visitor scientists, as well as on commonly agreed recommendations for citing facilities instruments.

Another effort of minting PIDs for facilities instruments is Journal of Large-Scale Research Facilities (JLSRF) [23] with the core editorial team from Jülich Research Centre [24]. The structured description of a facility instrument can be published as an article in this journal, with a DOI assigned. An instrument upgrade that qualifies as a new instrument can be published as another article in JLSRF with a new DOI.

The DOIs minted by JLSRF are intended mostly for their citation in research papers, and JLSRF articles (the DOIs landing pages) are intended for reading by humans. The JLSRF editorial team supported by other researchers across the globe have recently formed a Research Data Alliance Working Group specifically devoted to PIDs for instruments [25]. The group are currently collecting case studies from various research centres, and developing a common metadata model for instrument descriptions. The outputs of this group are likely to be skewed towards the use of instrument PIDs and PID-associated metadata by machine agents; hence the RDA work complements the ongoing publishing of "instrument articles" by JLSRF.

Persistent identifiers for facilities instruments (beamlines) will make them citeable in research papers and therefore will allow to better measure the research impact of particular instruments and facilities as a whole, and will contribute to implementation of FAIR principles for research outcomes by giving more context to research

papers and the associated research data. To better cater for impact measurements, the instrument PIDs and the facility PIDs can be included in a common vocabulary that allows a certain level of machine reasoning (semantic inference). This will allow citation-based measuring of facilities impact when researchers cite not a facility as a whole but a particular instrument of it.

3.2 Persistent identifiers for researchers: room for improvement

ORCIDs proliferation in facilities research, which is a multi-disciplinary research by its nature, is uneven across different disciplines. ORCIDs are promoted by libraries and IT departments of the organizations that operate facilities, yet key stakeholders for the wider ORCIDs adoption are facilities' user offices that manage research proposals from visitor investigators.

The aforementioned ORCID User Facilities and Publications Working Group [21] advocates for the entering of facilities beamtime in the ORCID record Funding section or Research Resources section; if this practice is well adopted by visitor scientists, it will allow facilities to better measure the research impact of particular beamlines (facilities instruments). This could be another mechanism for measuring research impact of facilities, in addition to the earlier mentioned possibility of counting instrument and facility PIDs citations by research papers.

Another opportunity for better ORCIDs adoption is doing some work on behalf of researchers, e.g. the autopopulation of institutional repositories, such as ePubs [11], with bibliography from ORCID records (maintained by researchers themselves). This will require integration of institutional software platforms with ORCID, which can be bidirectional, as ORCID, despite being primarily a platform for researchers identification, may be interested in the automated ingest of well-curated bibliographic records from institutional repositories and matching this bibliography to the researchers' records.

3.3 Persistent identifiers for data staging

As mentioned above in Section 2.3, persistent identifiers that presumably designate data can in fact be resolved in different information artefacts, not necessarily data per se. This presents difficulties for data staging to computation and visualization by virtue of the respective PIDs resolution.

For the purposes of data staging, PID resolution can be seen as an API call with one parameter [8]. The implementation of the API though should be inevitably specific to the actual protocol of how data can be accessed from the PID landing page. The complexity involves resolving a path to data assets, identifying data format, as well as a potential need for the authorization in data archive if data access is granted only to those authorized.

The protocol for data staging via PIDs can be implemented using content negotiation mechanisms offered by PID management services such as DataCite [16] or CrossRef [26]. How the protocol can be actually modelled in order to be machine-interpretable is an open question; just having a machine-interpretable metadata associated with the PID is unlikely to be enough. One possible approach suggested in [8] could be semantic annotation service with a mechanism for assembling granular RDF statements and for their transformation into executable workflows that perform data staging. Assigning PIDs at the data file or dataset level can help in formation of such executable workflows, or in some cases (specifically, when such granular PIDs are supplied with machine-interpretable metadata) can be a mechanism of its own for data staging to computation.

3.4 Persistent identifiers for records enrichment

Bibliographic records for research papers as well as other records of science such as landing pages for investigations or records of research awards (grants) have a potential for their enrichment with references to persistent identifiers such as PIDs for researchers, organizations, or well-established systems of identifiers for chemical substances and molecular structures [29].

The records enrichment, in order to be scalable and maintainable, should involve methods of textual analysis and machine learning. This presents excellent opportunities for data scientists and data engineers to showcase their methods and tools that can be used for matching textual names of researchers, organizations or chemical structures with commonly used identifiers, and then for building graphs that interlink these uncovered identities.

3.5 New representations of facilities research discourse

For facilities, the potential for better – well-structured and open – representation of research supported by them is in part related to the aforementioned opportunities for records enrichment (see Section 3.4). This can include better identification of instruments and researchers (see Sections 3.1 and 3.2).

Another opportunity for novel representations of facilities research is the further promotion of investigations (structured descriptions of them) as true components of research discourse across different facilities that are ready to accept common practices of minting PIDs for investigations and encouraging visitor scientists to use these PIDs as references in their research papers. In fact, not only a well-formed investigation record can be cited from a research paper, but investigations can cite previous research, too, such as previous investigations and research papers that contribute to research behind a proposed investigation. This allows mixing up research papers and investigations citation in a common citation graph where anything can refer to anything [7].

PID-rich descriptions of facilities research may benefit from emerging Open Access information services such as Open Citations [27] that has a potential to

challenge the Web of Science [28] decades-long monopoly as an authoritative source for citations data. Open Citations records can be matched with lists of publications resulted from facilities research, thus giving a bigger and more structured picture of facilities impact and contributing to the facilities' Open Science agenda.

New information-rich representations of facilities research discourse may become possible only with support of the proper governance, data curation and social practices; technology alone is just a facilitator of change. This understanding of the importance of practices and attitudes towards PIDs use is getting traction in projects such as FREYA [3] with its notions of "PID forum" – a communication hub for PIDs practitioners, and "PID Commons" – emerging communities of practice around PIDs management and PIDs use in research.

4 Conclusions and further developments

Wider and well-curated use of persistent identifiers can support FAIR principles for research data management and stewardship. Multidisciplinary research supported by large-scale facilities presents good opportunities for the application of information technology for PIDs management, as well as for the development of best practices and communities of practice around PIDs use cases that we briefly describe in this paper.

The practical implementation of these use cases can be supported by research facilities (in particular, their user offices), by information and IT departments of organizations where facilities operate, as well as through collaborative projects and volunteer work in international forums such as Research Data Alliance.

Acknowledgments. This work is supported by funding from the Horizon 2020 FREYA project, Grant Agreement number 777523. The views expressed are the views of the authors and not necessarily of the project or the funding agency.

References

[1] Bunakov, V., Jones, C., Matthews, B., 2015. Towards the Interoperable Data Environment for Facilities Science. In Collaborative Knowledge in Scientific Research Networks, chapter 7, 127-153. IGI Global, 2015. doi:10.4018/978-1-4666-6567-5.ch007

[2] Bunakov, V., Matthews, B. Data Curation Framework for Facilities Science. In 2nd International Conference on Data Technologies and Applications (DATA 2013), Reykjavik, Iceland, 29-31 Jul 2013, (2013): 211-216. doi: 10.5220/0004593302110216

[3] FREYA project. www.project-freya.eu Accessed in May 2018.

[4] THOR project. www.project-thor.eu Accessed in May 2018.

[5] Wilkinson, M.D. et al. The FAIR Guiding Principles for scientific data management and stewardship. In Scientific Data 3, Article number: 160018 (2016). doi: 10.1038/sdata.2016.18

[6] FREYA project deliverable D3.1 "Survey of Current PID Services Landscape". (Expected to be published on the Web by the conference time)

[7] Bunakov, V. Investigation as a member of research discourse. In 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Dubna, Russia, 13-16 Oct 2014. CEUR Workshop Proceedings Vol-1297 (2014): 160-165.

[8] Bunakov, V. Service for data retrieval via persistent identifiers. In DATA 2015: 4th International Conference on Data Management Technologies and Applications. Colmar, Alsace, France, 20-22 July, 2015, pp.177-182. doi:10.5220/0005554401780183

[9] Zenodo repository. https://zenodo.org/ Accessed in May 2018.

[10] PURL Administration. https://archive.org/services/purl/ Accessed in May 2018.

[11] ePubs repository. https://epubs.stfc.ac.uk/ Accessed in May 2018.

[12] Handle.Net registry. http://handle.net/ Accessed in May 2018.

[13] EUDAT B2SHARE service. https://b2share.eudat.eu/ Retrieved in May, 2018.

[14] arXiv.org preprints repository. https://arxiv.org/ Accessed in May 2018.

[15] FTP access to ESRF (European Synchrotron Radiation Facility) data. http://www.esrf.eu/Infrastructure/Computing/ComputingOffsite/FtpWindows Accessed in May 2018.

[16] DataCite service and consortium. www.datacite.org Accessed in May 2018.

[17] ORCID: persistent identifiers for researchers. https://orcid.org/ Accessed in May 2018.

[18] ISIS neutron and muon source. https://www.isis.stfc.ac.uk/ Accessed in May 2018.

[19] Oak Ridge National Laboratory (ORNL). https://www.ornl.gov/ Accessed in May 2018.

[20] Argonne National Laboratory (ANL). http://www.anl.gov/ Accessed in May 2018.

[21] ORCID User Facilities and Publications Working Group. https://orcid.org/content/user-facilities-and-publications-working-group Accessed in May 2018.

[22] PIDapalooza workshop. https://pidapalooza.org/ Accessed in May 2018.

[23] Journal of large-scale research facilities (JLSRF). https://jlsrf.org/ Accessed in May 2018.

[24] Forschungszentrum Jülich. http://www.fz-juelich.de/ Accessed in May 2018.

[25] RDA Persistent Identification of Instruments Working Group. https://www.rd-alliance.org/groups/persistent-identification-instruments Accessed in May 2018.

[26] CrossRef consortium. www.crossref.org Accessed in May 2018.

[27] Open Citations initiative. http://opencitations.net/ Accessed in May 2018.

[28] Web of Science indexing service. https://en.wikipedia.org/wiki/Web_of_Science

[29] Henry Rzepa. Bridging the gaps between current practice and FAIR data (molecular sciences and chemistry case). See abstract at http://sched.co/Cwma and the presentation slides at https://doi.org/ccs4 , both accessed in May 2018.

[30] Researcher ID. Thomson Reuters http://www.researcherid.com Accessed in May 2018.