Metadata for large-scale research instruments

Vasily Bunakov

Published version information

Citation: V Bunakov. "Metadata for large-scale research instruments." In: E Garoufallou, F Sartori, R Siatri, M Zervas (Eds.) **Metadata and Semantic Research**. MTSR 2018. Communications in Computer and Information Science, vol. 846. Springer (2019): 324-329.

DOI: 10.1007/978-3-030-14401-2_30

The final authenticated version is available online at Springer via https://doi.org/10.1007/978-3-030-14401-2_30

Metadata for large-scale research instruments

Vasily Bunakov[1]

[1] Science and Technology Facilities Council, Harwell OX11 0QX, United Kingdom
vasily.bunakov@stfc.ac.uk

Abstract. The work outlines diverse effort of a few initiatives for metadata and attribution mechanisms that can be used for large-scale instruments hosted by shared research facilities. Specifically, the role of persistent identifiers and associated metadata is considered, in relation to cases where the use of references to large-scale instruments can support research impact studies and Open Science agenda. A few routes for the adoption of large-scale instruments metadata are outlined, with indication of their advantages and limitations.

Keywords: large-scale instruments, research facilities, research attribution, persistent identifiers, research information management, impact studies, Open Science

1 Introduction

Large-scale research facilities such as synchrotron radiation sources, neutron sources or powerful lasers offer shared access to a variety of scientific instruments and are a prominent part of the research landscape for the last few decades. The notion of "instrument" in such facilities differs from that in other research contexts, as a facility instrument is often a complex set of equipment that evolves through time, may support multiple experimental techniques and requires specific research and technology expertise for its development and practical use. A facility instrument involves an organizational aspect and may be operated by a dedicated administrative unit; the instrument may have specific sources of funding and specific collaborations that perform the instrument support and upgrades.

Visitor scientists apply for a share of time on large-scale instruments in order to conduct their own research driven by their own research agenda. Depending on the nature of a particular research, the involvement of the host instrument specialists (instrument scientists) may be more of a supporting nature, or can make crucial contribution to research results. This leads to various practices of research attribution across different disciplines and research contexts, with the perceived tendency to the less frequent attribution given to instrument scientists [3].

There is a growing understanding that not only instrument scientists, but the instruments themselves deserve proper attribution in research outputs such as research papers, as this can contribute to impact studies that influence next rounds of investment in the large-scale instruments and in facilities as a whole. The problem of instruments attribution can be addressed using different information management tech-

niques; as an example, larger facilities can afford hiring a dedicated bibliographer who traces research papers down to particular instrument-specific awards that allowed raw data collection in the first place. Another approach is implementing certain policies that require visitor scientists to attribute their research outputs with clear references to instruments. In addition or alternatively, a certain level of information management automation can be introduced, so that when visitor scientists are granted with their timeshare of a large-scale instrument, their personal records in a publically available (harvestable) registry are automatically updated with proper references. Irrespective of the approach to attribution, the large-scale instruments require clear and persistent identity as a part of quality instrument metadata.

This work first introduces a few Open Science cases beyond impact studies that can be supported by quality metadata for large-scale instruments. It then outlines a few approaches to the instruments attribution, and suggests a few routes for the instruments metadata adoption by research facilities. It further suggests reasonable priorities for different adoption routes.

2 Open Science cases for large-scale instruments

Clear research attribution aimed at impact studies can be an immediate driver why research facilities should consider better metadata for their instruments, but this is not the only case where quality instrument metadata is required. FAIR principles [14] that initially promoted research data Findability, Accessibility, Interoperability and Reuse are now advised for their application to related algorithms, tools, workflows, protocols and services [1]. Instruments are now considered an essential part of research workflows that should support Open Science [2].

There are a few aspects of Open Science that facilities may want to explore through better metadata for instruments:

- Research trends and research frontiers studies; this may contribute to evidence-based planning for instruments and facilities upgrades, in order to keep abreast of research interests of applying researchers and their organizations
- Strategic partnership studies, e.g. through discovering and monitoring frequent (or otherwise prominent) funders of visitor scientists, as time slots on large-scale instruments can be considered grants-in-kind, hence recurring co-funding may indicate opportunities for permanent funders cooperation with a facility
- Research provenance chains that can include instruments where raw data was collected; this is important for research reproducibility and for informing potential research applicants about capabilities of particular instruments and facilities
- Giving proper credit to instrument scientists who may be less frequently mentioned nowadays as co-authors of peer-reviewed publications but deserve clear attribution of their work that contributes to quality research

In fact, there is no clear boundary between Open Science and impact studies traditionally supported by all sorts of research information management systems. Information services for Open Science can support impact studies, and potentially in novel

ways, with better granularity and with community review that can raise the quality of impact studies and public trust in them. A few ongoing initiatives on the instruments metadata and attribution can support the Open Science cases and impact studies.

3 Ongoing initiatives on design and implementation of metadata for large-scale instruments

3.1 Journal of large-scale research facilities (JLSRF)

Journal of Large-Scale Research Facilities (JLSRF) [4] is a peer-reviewed online Open Access journal with the editorial team from Jülich Research Centre [6]. The journal publishes articles that describe large-scale equipment intended for use by visitor scientists who are not affiliated with the institution operating the facility.

The articles are peer-reviewed by a reviewer board that is run by the journal; larger institutions that operate several facilities with multiple instruments are encouraged to set up their local reviewing body.

Articles can be attributed to the operating institution or the facility (corporate authors), yet people who compiled the article can be listed as contributors, which gives them a credit for their authoring of the instrument description. In any case, at least one human contact is provided for potential inquiries about the instrument.

An article published in JLSRF allows visitor scientists to cite large-scale instruments in their publications. Operators of large-scale facilities can refer to the respective article in JLSRF, too, e.g. on their websites or in their annual reports.

Every article is assigned with the DOI; an instrument upgrade description can be published as a new article with a new (modified) DOI.

Apart from the DOI, each article is supplied with Dublin Core metadata elements. These include citations (references) to papers and to other citable artefacts that may include other large-scale instruments, or previous versions of the same instrument, or a facility as a whole. This is an opportunity to give a rich information context to the facility instrument descriptions and include them, through citations, in a universal research discourse.

JLSRF is indexed by a few popular indexing platforms including OpenAire [8] and is recorded in Open Access monitoring databases such as SHERPA/RoMEO [9]. The articles metadata is harvestable via the widely known OIA-PMH interface.

Publishing articles about large-scale instruments in JLSRF can be the first reasonable step for facilities to develop best practices for clear instruments identification and for giving visitor scientists a handy mechanism for citing instruments.

3.2 RDA Persistent Identification of Instruments Working Group

The Research Data Alliance have recently endorsed a dedicated Working Group for Persistent Identification of Instruments [5]. The group collects case studies from various research disciplines, and aims to develop a common metadata model for instrument PID descriptions with the main purpose of using them by machine agents (soft-

ware), compared to the case of JLSRF where instrument descriptions are mostly intended for human consumption.

Another difference from the JLSRF is scope: this RDA group is interested in all sorts of instruments, not necessarily large-scale facilities instruments. As an example, a few use cases from geoscience and other disciplines are about the networks of sensors and other serial equipment. This makes the works of this RDA group, on one hand, universal, but on the other hand, the eventual metadata and associated information management practices may happen to be less suitable for a particular case of large-scale facilities instruments. Another limitation is that this RDA group decided to focus on instruments for measurements (data collection) when some large-scale instruments, e.g. photon sources, can be also used for samples modification.

From facilities perspective, this RDA group works should be best viewed as a complementary effort to what the JLSRF have been successfully doing for years.

3.3 ORCID User Facilities and Publications Working Group

ORCID User Facilities and Publications Working Group [7] engages with information management specialists mostly from American large-scale research facilities and aims to promote ORCID persistent identifiers for facilities visitor scientists. The Group have developed a recommendation for facilities user offices to request ORCID IDs and personal ORCID API tokens from researchers who apply for facilities time, as well as to ask the researchers' consent for auto-populating their ORCID accounts with information about facilities time allocation. Once a time slot is allocated, the notice of it can be published by facility in the visitor scientist's ORCID's section devoted to grants or in other section devoted to resources that supported researcher.

There is no functionality within ORCID that allows linking particular papers in the ORCID Publications section with records in the Research Resources or Funding sections. It is the ORCID's view that a publication will get linked to a facility award as a grant-in-kind or to the facility instrument as a research-resource-in-kind when a researcher submits a manuscript for publication. It is the publisher's responsibility then to ask about research awards and resources that supported the paper in question.

Unless the link between a research paper and an instrument is requested by a publisher, it will be only possible to find out, using just an ORCID record, that a particular researcher used a certain facility instrument at a certain time. This will be enough for *some kinds* of pretty coarse impact studies and Open Science use cases but not for fine-grained assessment or for a sensible level of research reproducibility, so this approach that relies on publishers' best practices (that may be diverse across different publishers, too) has its natural limitations. Also facilities' reluctance, owing to privacy concerns and extra effort required, to request ORCID identifiers from visitor scientists and to auto-populate their ORCID records, can be an obstacle for the universal adoption of the ORCID-based mechanism of instruments attribution.

3.4 PANKOS vocabulary

PANKOS vocabulary is an ontology of photon and neutron sources; it was one of the outcomes of PaNdata-ODI project (see under [10]). This kind of a semantic resource will be invaluable in fine-grained impact studies and for Open Science use cases.

The reason for this is that researchers can cite the entire facility in their papers, e.g. a synchrotron light source, or they can cite a particular instrument of it, or a particular experiment (investigation) that corresponds to a facility research award. There should be some means to make aggregations up to the instrument or to the facility level, in order to count all citations towards the impact of a particular instrument or a facility as a whole. More complex and more granular studies of a comparative nature can be considered, too, e.g. comparing impact of only the instruments that use the same or similar experimental technique across a few facilities.

Therefore, semantic links are required within vocabulary that allow to reason over the belonging of the instrument to a facility, as well as over the experimental technique used by the instrument. In PANKOS, this was achieved using Web Ontology Language (OWL) classes, which can be perceived as overcomplicated from modelling point of view and thus prevent the universal vocabulary adoption. The OWL modelling may also present an indirect obstacle for the vocabulary deployment in a variety of IT environments, as triple stores differ in flavours of OWL they can support.

The PANKOS can be a good starting point though for the design of a new universal vocabulary that will have enough expressivity to support the aforementioned modes of reasoning, but will not be overcomplicated or facility-specific. The vocabulary may include the notion of samples modification, not only of their characterization, which will make it applicable when a large-scale instrument is a part of a production line or is otherwise used for the alteration of exposed samples.

4 FREYA project and priorities for instruments metadata

FREYA [11] is a project funded by the European Commission under the Horizon 2020 programme. It aims to extend the infrastructure for persistent identifiers as an essential component of Open Science, in the EU and globally.

It is a view of FREYA that research communities should come up with their own use cases of using PID resolution services such as Crossref [12] or DataCite [13] for the promotion of research FAIR principles. In turn, research communities are expected to contribute their purpose-built research graphs, with all kinds of PIDs as nodes and with sensible relations between the nodes, into a common PID graph, or into the interoperable federation of PID graphs and services built atop of them.

Communities that operate and use large-scale instruments may benefit from this FREYA vision and from services based on PID graphs. Metadata for large-scale facilities instruments will contribute to the graphs and can be adopted through the following routes that complement each other but will be best pursued in the following order:

- Textual descriptions and metadata in JLSRF, including DOIs suitable for citation of instruments in research papers and for their linking to other research artefacts

- Machine-interpretable metadata associated with instrument PIDs if/when the appropriate recommendation will be agreed by the RDA PIDINST WG
- Entries for the large-scale facilities instruments in a semantic vocabulary that allows modest machine reasoning, with the right balance between the vocabulary expressivity and simplicity
- The practice based on facility instruments registration in ORCID records can be further explored, but it involves a variety of stakeholders with different policies, which may hinder the universal adoption of ORCID recommendations

The semantic vocabulary can be a proper tool for eventually incorporating the other flavours of metadata for large-scale instruments. The vocabulary can include DOIs of the JLSRF articles and PIDs for the machine-actionable descriptions according to the RDA PIDINST WG recommendations. These PIDs can be related by a vocabulary entry with an indication of their respective purposes, and with a possibility to cross-walk between different PIDs for the same instrument.

5 Acknowledgements

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

References

1. European Open Science Cloud Declaration. https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf Accessed in June 2018.
2. Open Science Commons. https://documents.egi.eu/public/ShowDocument?docid=2410 Accessed in June 2018.
3. Mesot, J.: A need to rethink the business model of user labs? Neutron News, 23 (4), 2-3 (2012).
4. Journal of large-scale research facilities (JLSRF). https://jlsrf.org/ Accessed in June 2018.
5. RDA Persistent Identification of Instruments Working Group. https://www.rd-alliance.org/groups/persistent-identification-instruments Accessed in June 2018.
6. Forschungszentrum Jülich. http://www.fz-juelich.de/ Accessed in June 2018.
7. ORCID User Facilities and Publications Working Group. https://orcid.org/content/user-facilities-and-publications-working-group Accessed in June 2018.
8. OpenAIRE initiative. https://www.openaire.eu/ Accessed in June 2018.
9. SHERPA/RoMEO. Publisher copyright policies & self-archiving information service. http://www.sherpa.ac.uk/romeo/ Accessed in June 2018.
10. PANdata: Photon and Neutron data infrastructure initiative. http://pan-data.eu/ Accessed in June 2018.
11. FREYA project. https://www.project-freya.eu/ Accessed in June 2018.
12. Crossref consortium. https://www.crossref.org/ Accessed in June 2018.
13. DataCite consortium. https://datacite.org/ Accessed in June 2018.
14. Wilkinson, M. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018 (2016). DOI: 10.1038/sdata.2016.18