University of Oxford

Engineering Science

Shaping a modern approach to open data from a World-leading science facility

Overview of the proposed project for the Project Stakeholder Group

Author

Nian Yang Terence Tan

Wadham College

Supervised by

Prof. Susanna-Assunta Sansone (University of Oxford)

Dr. Philippe Rocca-Serra (University of Oxford)

Dr. Steve Collins (Diamond Light Source)

April 11, 2024

1. Background and Rationale

Diamond Light Source (DLS) is the UK's national synchrotron[1]. Synchrotron facilities produce electromagnetic radiation by accelerating electrons to near the speed of light. The x-rays generated can be used to study objects and structures much too small to be seen by optical microscopes.

The experiments conducted at DLS contribute to many areas of scientific research. The vast quantities of data generated across diverse research fields present an obvious opportunity for the application of the Findable, Accessible, Interoperable, Reusable (FAIR) Principles (Wilkinson et al., 2016) to increase the potential of its data and better address key societal challenges.

The FAIR Principles are a set of principles and best practices for data management to enhance the value of all digital resources and its reuse by humans and machines. Since then, the Principles have been adopted by various organisations such as the European Union (European Commission, 2020). Moreover, the European Commission has estimated that the failure to implement FAIR data can cost up to 10.2 billion Euros per year for the European economy (Research, Innovation., and Services., 2018). The Principles offer DLS the opportunity of transforming its relationship with the vast quantity of research data by ensuring that it is openly available to all humans and machines in a way that enhances the value of the scientific research.

There is an active community of service providers and infrastructure developers who have been working to apply the FAIR Principles to Photon and Neutron data and research. Notably, European science cluster projects such as Photon and Neutron Open Science Cloud (PaNOSC) and European Open Science Cloud Photon and Neutron Data Service (ExPaNDS) have produced tool and services for the purpose of supporting FAIR and Open Science within the Proton and Neutron community in Europe, also known as PaN, which includes facilities like DLS and the European Synchrotron Radiation Facility (ESRF). Organisations such as the League of European Accelerator-based Photon Sources (LEAPS) also aim to promote the impact of research carried out at member facilities to support European science.

Besides the European-focused ones, there are also global initiatives such as the NeXus data format, which is a standardised data format for x-ray, neutron, and muon science. In addition, the Photon and Neutron Science Interest Group (PaNSIG), which is part of the Research Data Alliance (RDA), was established to provide a forum for discussions of shared data-related issues of the global PaN facilities.

[1] https://www.diamond.ac.uk/Home/About.html

The project is aimed at understanding opportunities and barriers in moving towards FAIR and open data in a wide range of science disciplines at DLS, from energy materials and palaeontology to studies of viruses and drugs. Whilst the project is designed to deliver novel conceptual and methodological contributions to enhance the value of DLS research data, it will also leverage on and complement the activities of existing above-mentioned communities and projects. The project will specifically seek to make experiment proposals, submitted to DLS, FAIR-compliant with machine-actionable metadata, using Machine Learning methods and tools. The intention is to make the proposal stage of the DLS science life cycle FAIR-compliant, in order to kickstart the FAIRification process, creating a cascading effect on processes in the other stages. In general, making data FAIR-compliant will also help DLS achieve its five Goals (Diamond Light Source Ltd, 2015) by enabling DLS to deliver highly reusable data to users, maximising its usefulness. More specifically, the machine-actionable metadata extracted from the experiment proposals holds various potential applications. for example, enable more granular classifications of scientific topics, allowing for better analysis of the type of research being conducted.

2. Problem Statement

Although various communities and projects have made great strides in delivering techniques and standards set to make synchrotron data FAIR, the journey is far from over. Many remaining challenges and areas still require work. In particular, reports and outputs from the PaN community and PaNSIG (Ivănoaica et al., 2021; Carboni, 2022; Boscaro-Clarke et al., 2023; Boscaro-Clarke and Roarty, 2019) highlight key implementation challenges, and emphasise the difficulty of introducing a culture change regarding FAIR data, and the need for continuous communication and dissemination of the Principles and their benefits.

Furthermore, FAIR is not "one size fit all", and therefore each organization will also have to analyse their current internal processes, and identify value-added improvements using a cost-benefit assessment (Alharbi et al., 2023). Therefore, it is timely for this project to explore how to realise FAIR at DLS, maintaining the community momentum.

3. Research Questions

This project will answer the following research questions:

1. How would the adoption of FAIR Principles at DLS look like, and affect synchrotron data management?

2. What benefits do the FAIR Principles bring to DLS as an organisation, and to its science life cycle?

In order to answer these two questions, this project has identified a key stage in DLS' science life cycle, which describes the research process at DLS from the experiment proposal stage to eventual publication of results, which machine-actionable metadata will be extracted. In addition, this project will evaluate the impact of the proposed improvements that the metadata may have on the processes of DLS while contributing to the activities of the PaN other relevant communities.

4. Research Methodology

A summary of the overall work plan of the project is outlined below:

1. **Conduct a literature review, and landscape analysis of DLS' processes and science life cycle**
   A reflective account of the research that has been done so far will focus on the work by PaN and other communities, and will be coupled by a picture of the current internal processes, set to identify and confirm the areas that could be prioritised and enhanced by FAIRification.

2. **Create a data pipeline to get machine-actionable metadata from experiment proposals**
   Metadata about experiment proposals will be extracted using Machine Learning and linked with concepts in web-accessible ontologies/databases through semantic annotation to make them machine-actionable. Relevant Machine Learning techniques and ontologies/databases, such as ChemDataExtractor (Swain and Cole, 2016) and the Photon and Neutron Experimental Techniques ontology (Collins et al., 2021) respectively, will be identified and integrated into a data pipeline. A performance metric will also be determined.

3. **Investigate potential applications and pilots**
   It will involve the exploration of how the machine-actionable metadata can be used to enhance selected DLS processes.

4. **Define methodology to evaluate the usefulness of the potential applications**
   It will involve qualitative methods, such as surveys and interviews with DLS staff and researchers.

5. **Derive recommendations to DLS for future work and directions**
   The recommendations will enable the continuation of the work done by this project, which will hopefully help to address the future needs of DLS and other PaN facilities.

To ensure the novelty of the work, its relevance to the PaN community, and to firmly anchor it to the DLS key staff members, a Project Stakeholder Group has been set up, and it will be engaged and consulted with throughout the project. Currently its members are:

- Kirsty Syder (DLS), Research Data Manager
- David Aragao (DLS), Beamline Scientist
- Tim Snow (DLS), Data Analysis Scientist
- Renaud Duyme (PaNOSC, ESRF), Software Engineer
- Brian Matthews (STFC, RDA PaNSIG), Leader of the Open Data Systems Group

5. Risks and Mitigation

1. **Labelling of dataset**

   Supervised Machine Learning models require a labelled dataset for training. Labelling of datasets can be a time-consuming process. Unsupervised models, which do not require labelled training data, should be used wherever possible to mitigate this risk. Additionally, the modular nature of the data pipeline allows us to manage this risk by narrowing our focus down to a select few components of the pipeline.

2. **Not having enough data for training**

   Many Machine Learning models require copious amounts of training data to function, and there may not be enough viable experiment proposals from DLS for this purpose. This risk could be mitigated by using experiment proposals from other facilities.

3. **Resistance to change**

   It is notoriously challenging to encourage a culture change within organisations and communities. It is important that this project highlights the potential ways FAIR data can enhance DLS' processes in a way that benefits the staff and researchers in order to incentivise them to incorporate the Principles into their activities.

References

Alharbi, Ebtisam et al. (Apr. 2023). "A FAIR-Decide framework for pharmaceutical R&D: FAIR data cost–benefit assessment". In: *Drug Discovery Today* 28.4, p. 103510. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2023.103510. URL: http://dx.doi.org/10.1016/j.drudis.2023.103510.

Boscaro-Clarke, Isabelle and Kat Roarty (2019). "ExPaNDS Communication & Dissemination Plan". In: DOI: 10.5281/ZENODO.4714676. URL: https://zenodo.org/record/4714676.

Boscaro-Clarke, Isabelle et al. (Jan. 2023). "ExPaNDS: Laying the Foundations for Achieving Open Science for Everyone". In: *Synchrotron Radiation News* 36.1, pp. 25–28. ISSN: 1931-7344. DOI: 10.1080/08940886.2023.2186664. URL: http://dx.doi.org/10.1080/08940886.2023.2186664.

Carboni, Nicoletta (2022). *PaNOSC D9.4 - Report on dissemination and outreach activities*. en. DOI: 10.5281/ZENODO.7401055. URL: https://zenodo.org/record/7401055.

Collins, Steve P. et al. (2021). "ExPaNDS ontologies v1.0". en. In: DOI: 10.5281/ZENODO.4806026. URL: https://zenodo.org/record/4806026.

Diamond Light Source Ltd (Oct. 2015). *A 10-Year Vision for Diamond Light Source*. URL: https://www.diamond.ac.uk/Home/Company/Vision-and-Strategy.html.

European Commission (Feb. 2020). *A European strategy for data*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066.

Ivănoaica, Teodor et al. (Dec. 2021). *D8.1 – Report on lessons learned and future prospects for adopting best practices data stewardship at the PaNOSC facilities*. URL: https://www.panosc.eu/wp-content/uploads/2022/06/PaNOSC_D8.1_Report-on-lessons-learned_20211201.pdf.

Research, European Commission. Directorate General for, Innovation., and PwC EU Services. (2018). *Cost-benefit analysis for FAIR research data: cost of not having FAIR research data.* Publications Office. DOI: 10.2777/02999. URL: https://data.europa.eu/doi/10.2777/02999.

Swain, Matthew C. and Jacqueline M. Cole (Oct. 2016). "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature". In: *Journal of Chemical Information and Modeling* 56.10, pp. 1894–1904. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.6b00207. URL: http://dx.doi.org/10.1021/acs.jcim.6b00207.

Wilkinson, Mark D. et al. (Mar. 2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: http://dx.doi.org/10.1038/sdata.2016.18.