Data management plans as a tool for making data fair

openaccessgovernment.org/article/data-management-plans-as-a-tool-for-making-data-fair/168861

24 October 2023

Andy Götz (ESRF Data Manager and PaNOSC Coordinator), explores if and how Data Management Plans (DMPs) are essential for making data FAIR

In the previous articles in this series on FAIR data, we explained how the scientific world is undergoing a major change with the widespread adoption of the so-called FAIR principles for research data. FAIR stands for Findable, Accessible, Interoperable and Reusable and was first published in a paper in Nature during 2016. (1)

The FAIR principles were proposed to ensure research data are made available to the scientific community so that they can be found, downloaded, understood and reused. The goal is to make data used in scientific publications available to the community so they can verify the results, reproduce them and eventually derive new results from them.

Applying the FAIR principles systematically to research data will address the reproducibility, also known as the replicability crisis (2) in science. It will make scientific data available for verifying results and use beyond their original purpose.

This article will explore if and how Data Management Plans (DMPs) are essential for making data FAIR.

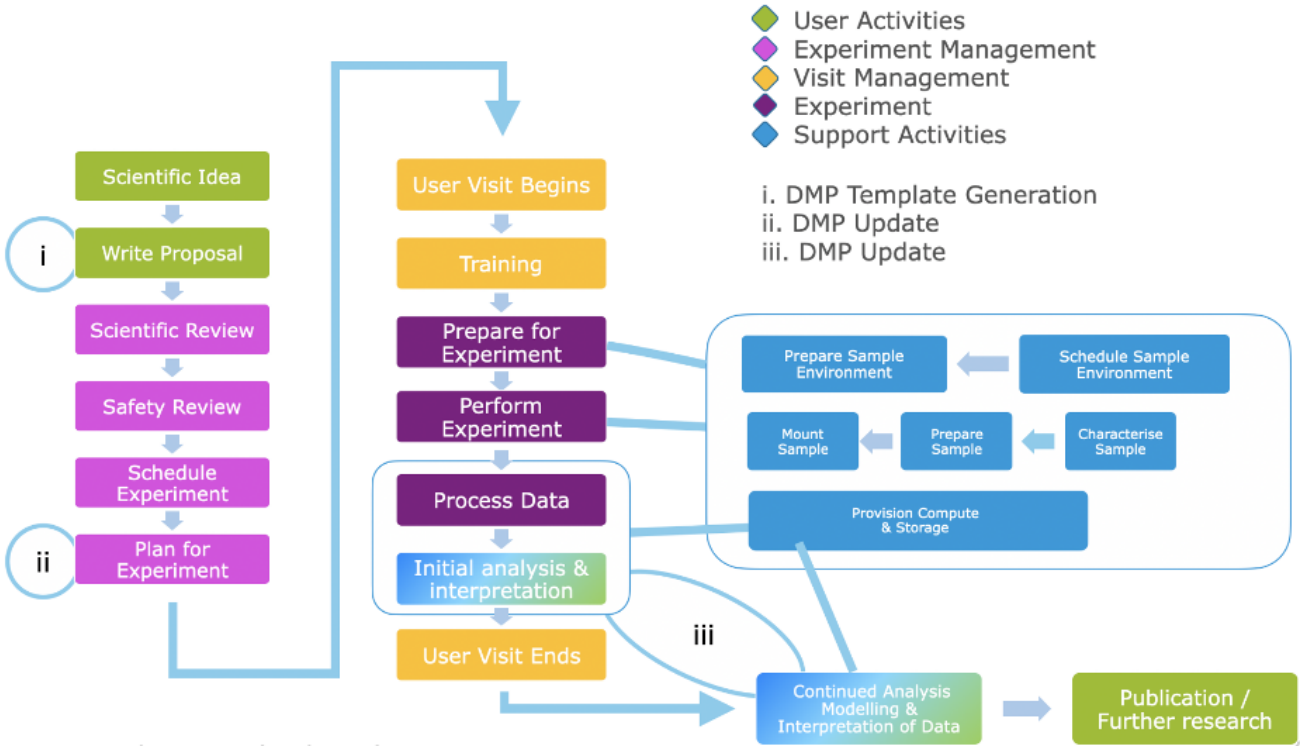

Figure 1 Active DMPs in the PaN facility scientific workflow
(diagram from Bodin, M., Bolmsten, F. et. al., 2023. Data Management Plans for the Photon and Neutron Communities. DOI: https://doi.org/10.5334/dsj-2023-030)

Defining a Data Management Plan

What is a Data Management Plan?

In its simplest form, it is a checklist for scientists of what they need to do to manage the data produced during their study or experiment. The checklist is usually formulated as a set of questions the scientist(s) must answer or keep in mind. The questions should address the following topics [(3)]:

- Type of data: observation, experimental, simulation, derived/compiled.
- Form of data: numeric, text, audio-visual, discipline or instrument-specific.
- File formats:
    1. Research community standards preferred (e.g., NeXus for photons and neutrons).
    2. Preservation formats preferred (e.g., binary in HDF5, docs in PDF, images in JPG).
- Size of data, stable data: plans for where they will be stored them.
- Sensitive data: plans for secure storage and anonymisation.
- Metadata: metadata standards, electronic logbooks, sample description, auxiliary information.

Questions about data, DMPs

The above topics are essential to know what data will be produced and how they will be managed and eventually made FAIR. However, these topics are often lost in questions about legal rights to data, long-term preservation, data volumes, etc., resulting in some data management plans requiring scientists to answer up to 200 questions. This puts an extra burden on the scientists with very little added value for doing the experiment. Scientists often have to fill in DMPs for funders without help and can feel overwhelmed by the technical jargon and number of questions asked. Scientists are motivated by science and not data management.

At the same time, most funding agencies see DMPs nowadays as a tool to ensure data are managed properly by making DMPs mandatory. Consequently, what should be an acceptable process based on best practices for managing scientific data, has become a burden for scientists.

Are data management plans essential for FAIR data? This question is fundamental to accepting and using DMPs as a useful and necessary tool. By reformulating the question, one could ask, “Do scientists need a plan to manage ever-increasing data volumes to share them?”. The obvious answer is YES if we want open science to be a reality. [(4)]

But how to get there without creating unnecessary administrative overhead for scientists? The answer is to help scientists by pre-filling DMPs and providing them with tools to enhance and extract useful information from DMPs. Research institutes must play their role in managing research data and giving high-quality infrastructures for data curation and long-term preservation. Without this, DMPs are doomed to be seen as a waste of time when, in fact, they are an essential part of making data FAIR.

The PaNOSC European H2020 project

In PaNOSC, a different approach has been taken for the case of photon and neutron source, where vast volumes of data need to be managed by the facilities. [(5)] Photon and neutron sources are user facilities for tens of thousands of scientists worldwide. Figure 1 depicts the complex workflow for scientific experiments from a scientific idea to publication.

The PaNOSC European H2020 project has been an occasion to explore how and when to introduce DMPs to the scientific workflow. After analysing when best to introduce DMPs in the workflow, it became evident that DMPs need to be active, i.e., updated during multiple phases in the workflow.

Active DMPs have been implemented at some PaNOSC facilities to demonstrate their added value. To reduce the burden on scientists, the DMPs are pre-filled and updated with all information that can be provided by the facility and experiment at the different phases. The next step is to present the minimum information from the DMP, which scientists need to know to manage their data to make them FAIR so that they have a plan which is easy to understand and use.

The future of DMPs in science

What is the future of DMPs in science now that they have been made mandatory by almost all the large funding bodies for science? Following the example of the photon and neutron sources in Europe (PaNOSC), DMPs must be made simpler and easier to comply with the best practices in each scientific domain to ensure their wide acceptance and usefulness in making data FAIR.

The other area of active development of DMPs is making them machine-readable and actionable so that they can be validated automatically, and valuable information can be extracted to build knowledge graphs of scientific research. A new project (OSTrails) funded by the European Commission will explore creating knowledge graphs from DMPs using data from multiple scientific domains.

Data management plans are necessary for both FAIR data and creating knowledge based on open science but need to made easier and more useful for scientists to adopt them.

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
2. https://en.wikipedia.org/wiki/Replication_crisis
3. Example from https://researchdatamanagement.harvard.edu/data-management-plans
4. “Everyone needs a data-management plan”, Nature 555, 286 (2018) doi: https://doi.org/10.1038/d41586-018-03065-z
5. Bodin, M., Bolmsten, F., Aulin, P., Ivănoaica, T., Olivo, A., Malka, J., Wrona, K. and Götz, A., 2023. Data Management Plans for the Photon and Neutron Communities. Data Science Journal, 22(1), p.30.DOI: https://doi.org/10.5334/dsj-2023-030

This project has received funding from the European Union’s HORIZON 2020 Research and Innovation programme under the Grant Agreement no. 823852.

Please Note: This is a Commercial Profile