

Sharing, curation and metadata as essential components of the data management plan

Jennifer Gibson*

Executive Director, Dryad, Davis, USA

Abstract. This paper stresses the critical importance that data management plans actually be data management and sharing plans, that encourage data-sharing at the early stages of research; that indicate that data must be made as open as possible (and as closed as necessary); and that specify a minimum level of curation and collection of essential metadata. The author stresses that fact that it is no longer sufficient to post material openly and hope that others will be able to make use of it. The use of metadata and careful data curation of the data are essential to facilitate the discovery and successful re-use of data.

Keywords: Data management plan, metadata, data sharing, data curation, Dryad, the Generalist Repository Ecosystem Initiative

It was with foresight that the organizers of the 2023 NISO Plus meeting invited an open data publishing platform (or ‘digital repository’¹) to participate in a discussion on standards for data management plans. As the U.S. National Institutes of Health has demonstrated with a policy taking effect in January 2023, policy makers are shifting their focus from data management to data management and sharing. Ensuring that research data is available, under appropriate terms, to all those who may benefit from or build on prior findings is the next step in the responsible management of research outcomes and maximizing our collective investment in research. And it is the role of data publishing platforms and repositories such as Dryad to do so.

We must think about how the data will be accessed and used. The standard for data management plans is that they must be data management *and sharing* plans, and prescribe for the researcher that data must be made available as openly as possible. For greatest impact, these plans should:

- Encourage data-sharing at the early stages of research.
- Indicate that data must be made as open as possible (and as closed as necessary).
- Specify a minimum level of curation and collection of essential metadata. It’s no longer sufficient to post material openly and hope that others will be able to make use of it.

Here, we delve into the importance of metadata and curation.

First, as context for this presentation, it’s worth pointing out that Dryad is an instrument of open research and intends to enable and perpetuate the reuse of research data. Dryad is a platform for the curation and

*E-mail: jgibson@datadryad.org.

¹Dryad would prefer to be characterised as an open data publishing platform, dedicated to the curation and open publishing of research data that is committed to its reuse.

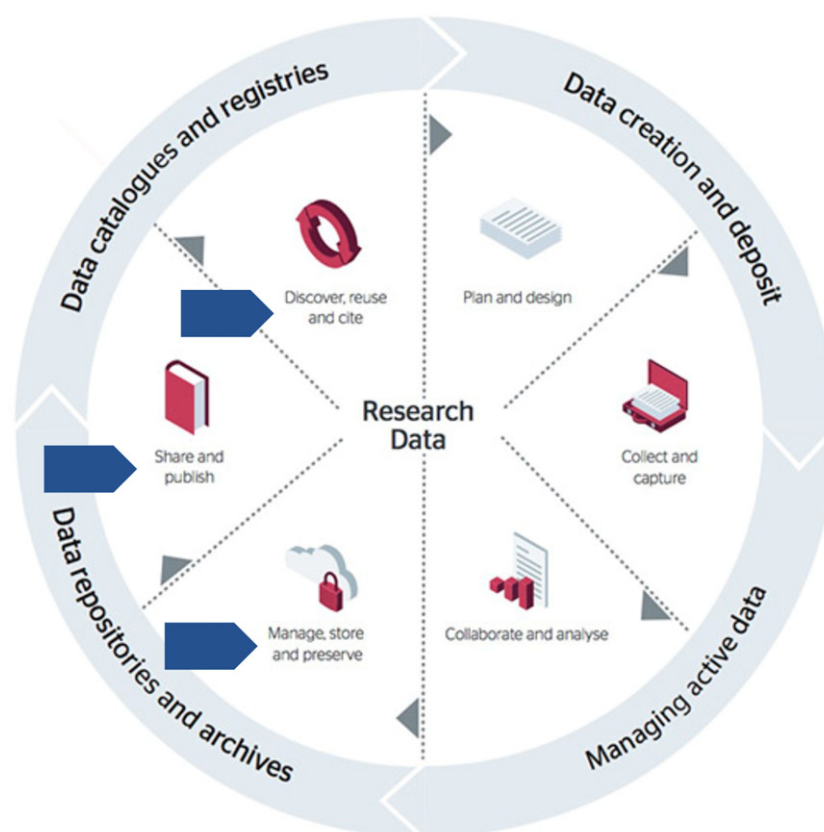


Fig. 1. Source: “[New process: Research data management for awarded projects](#)”. Edinburgh Napier University. © 2021.

open publication of research data from across fields and a multi-stakeholder community of academic and research institutions, societies, publishing organizations, and funders committed to a vision for the open availability and routine reuse of all research data.

Within the broader research data management process (illustrated in Fig. 1), Dryad fits foremost and solidly on the left-hand side - helping researchers to manage, store, preserve, share, publish, discover, reuse and cite data.

With more movement toward sharing data earlier in the process however - among institutions and funders including the U.S. National Institutes of Health (NIH) - Dryad can also be a resource for collecting and capturing data at any stage of research. Data submitted to Dryad may or may not be intended for immediate public sharing, and might be kept private while associated work continues. Ultimately, data shared with Dryad will be published openly, as our platform does not support access restrictions of any kind.

1. The importance of metadata

The essential glue in connecting data management and sharing plans with data resulting from research is metadata. High-level metadata invoking community-supported ontologies is simply critical for unifying



Fig. 2. Objectives of the Generalist Repository Ecosystem Initiative. Source: NIH Office of Data Science Strategy (<https://datascience.nih.gov/data-ecosystem/exploring-a-generalist-repository-for-nih-funded-data>).

the researcher's plan – and associated funding and institutional support – with the published outcomes in data.

At Dryad we invoke the Research Organization Registry (ROR) for institutional affiliation, Funder registry for research funding, and the Organisation for Economic Co-operation and Development (OECD) research classification for subject area. These are required metadata for every data publication submitted to Dryad.

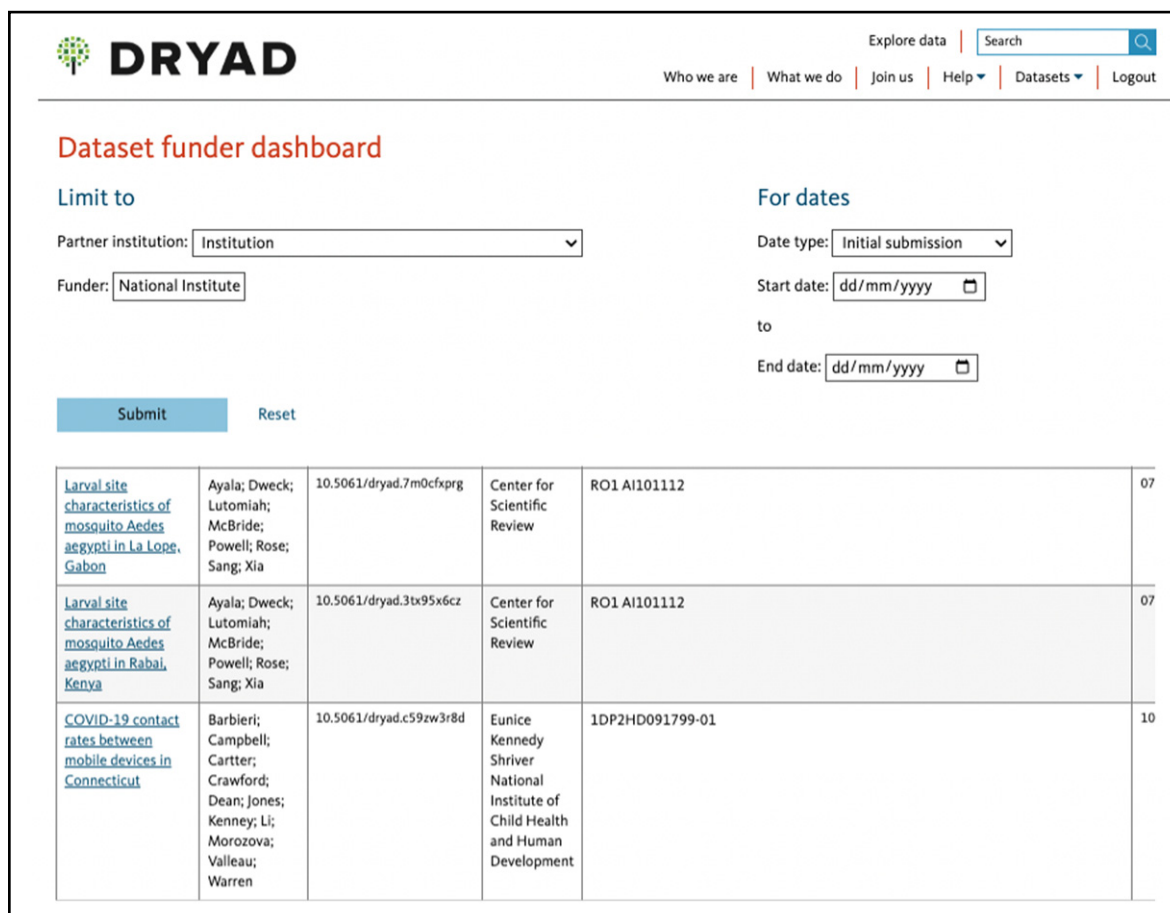
ROR and Funder Registry in particular are powerful tools for connecting investments and plans with outcomes. We can see this highlighted specifically in regard to the NIH policy that came into effect in January of this year.

In support of that policy, the NIH Office of Data Science Strategy has organized the Generalist Repository Ecosystem Initiative (GREI),² a four-year program to foster collaboration and cohesion among the seven so-called generalist repositories in developing common approaches to supporting researchers affected by the NIH data management and sharing policy. The initiative and the collective of organizations participating have identified ten areas of focus, and three that call for strong metadata: a consistent metadata model, discovery of NIH-funded data, and connecting digital objects (see Fig. 2).

The initiative will break ground in many ways, on behalf of the global community of stakeholders interested and invested in making the most of the networked open research environment. The ambition to

²More information about GREI is available on the ODSS website: <https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative>, accessed September 24, 2023.

report on results stemming from support from the twenty-seven NIH institutes and centers, for example, is elevating the need for at least two levels of metadata relating to funding. The outcome - importantly - will be that when the NIH begins to assess the reach and impact of their new policy it will be possible to do so at an appropriately detailed level.



Dataset funder dashboard

Limit to

Partner institution:

Funder:

For dates

Date type:

Start date:

to

End date:

Larval site characteristics of mosquito <i>Aedes aegypti</i> in La Lope, Gabon	Ayala; Dweck; Lutomia; McBride; Powell; Rose; Sang; Xia	10.5061/dryad.7m0cfxprg	Center for Scientific Review	RO1 AI101112	07
Larval site characteristics of mosquito <i>Aedes aegypti</i> in Rabai, Kenya	Ayala; Dweck; Lutomia; McBride; Powell; Rose; Sang; Xia	10.5061/dryad.3tx95x6cz	Center for Scientific Review	RO1 AI101112	07
COVID-19 contact rates between mobile devices in Connecticut	Barbieri; Campbell; Cartter; Crawford; Dean; Jones; Kenney; Li; Morozova; Valteau; Warren	10.5061/dryad.c59zw3r8d	Eunice Kennedy Shriver National Institute of Child Health and Human Development	1DP2HD091799-01	10

2. Admin dashboards and reports are accessible by Dryad members

Strong metadata is not only necessary to connect data with funding and institutions. It is also integral to making data discoverable across systems. Especially community-supported metadata standards, which - though still evolving - are adopted by multiple systems, facilitate travel for data, or information about the data, to travel from one repository to any number of other end points in the global research network, where it may be discovered readily with queries on funding, institution, research classification, grant ID, or related research objects.

This empowers discovery. And, when the data is openly accessible to those who discover it, it empowers open research - helping researchers carry out their work more effectively.

More and more research funders see the value of metadata and are incorporating it into emerging policies for data management and sharing. The U.S. NIH policy for data management and sharing³ aims to “maximize the appropriate sharing of scientific data” through the use of established repositories, timely publication of data, and data quality assurance. The U.S. White House Office of Science and Technology Policy recommendations to make federally funded research openly available⁴ aim to “improve research integrity and reproducibility” through: the immediate, open deposit of data; use of recommended repositories; and robust metadata & persistent identifiers.

3. The importance of curation

Funders are also converging on the importance of curating data before it’s shared. With respect to data quality assurance, the NIH says, “Data should be of sufficient quality to validate and replicate research findings”.⁵ At Dryad, where our team of human curators opens each file to ensure it can be opened and read by another human (not to mention machine) and that the data is adequately described for someone else to use and build on it, this means curation.

Metadata are the key to unifying the data management plan and the data resulting from associated research. But, in the interest of fully leveraging the tools and technologies available to us in 2022 to accelerate the pace of research and, more readily, translate research into benefits for the world, data management plans are evolving into data management and sharing plans, and data management and sharing plans are recognising that sharing must be done with care and attention (read: curation) to be effective. That is what the standard should be.

Acknowledgement

The author would like to thank Maria Praetzellis at the University of California Curation Center for helping to shape this talk.

About the author

Jennifer Gibson joined Dryad as Executive Director in October 2021. Since 2005, she has worked with scientists, funders, publishers, libraries, developers, and others to explore fresh paths toward accelerating discovery through open research communication and open-technology innovation. Prior to Dryad, Jennifer was Head of Open Research Communication for eLife, a non-profit and research funder-backed initiative to transform science publishing. She is Chair of the Board of Directors for OASPA (2020–2022) and a former member of the board for FORCE11 (2018–2020). E-mail: jgibson@datadryad.org.

³https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html#_ftn8, accessed September 24, 2023.

⁴<https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf>, accessed September 24, 2023.

⁵Op. cit.