Making Sense of Data: What You Need to Know About Persistent Identifiers, Best Practices, and Funder Requirements

MODERATOR:

Christine Casey
Editor, Serials
Centers for Disease Control and
    Prevention
Atlanta, Georgia

SPEAKERS:
David Carr
Policy Advisor
Wellcome Trust
London, United Kingdom

Patricia Cruse
Executive Director
DataCite
San Francisco, California

Shelley Stall
Assistant Director, Enterprise
    Data Management
American Geophysical Union
Washington, DC

Kerry Kroffe
Senior Editorial Manager
PLoS One
San Francisco, California

REPORTER:
Darren Early
American Society for Nutrition
Rockville, Maryland

Figure 1. Services to make sense of data (courtesy of Patricia Cruse).

This session focused on data from the perspectives of funders, standards and tools, policy, and operations. David Carr noted that Wellcome Trust (WT) has long been committed to ensuring research outputs can be accessed for societal benefit and has been a persistent, long-term advocate of open access and data sharing. Open sharing of research outputs can accelerate discovery, help to validate and reproduce research findings, and reduce duplication and waste in research efforts. Momentum has been growing internationally for open access and data sharing, and the policies of research funders with respect to data have converged (e.g., requirements for preservation and sharing of data and for data management plans). Expectations have also emerged for specific data types, such as requirements for clinical trial registration. WT's data management and sharing policy (2007, updated 2010) expects all researchers to maximize access to research data, requires a data management and sharing plan, and commits to supporting the costs of such plans as an integral part of WT grants. A survey of WT researchers regarding open research uncovered concerns about misuse of data, loss of publication opportunities, and the resources and time required to share data. An open research team was established at WT in early 2017 to develop policies and

practices and support innovative pilots, such as Wellcome Open Research, which is a new publishing platform for WT researchers that features post-publication peer review. Wellcome Open Research plans to add widgets for Code Ocean to allow readers to access and use the computational code in articles.

Patricia Cruse provided information from JoRD (Journal Research Data Policy Bank) on the data-sharing policies of journals. Of 371 journals surveyed in 2013, only 31 made data sharing a requirement for publication. Reliable and unambiguous access to data is needed for attribution, collaboration and reuse, reproducibility, and faster and more efficient progress. Data sharing is also needed to comply with publisher policies, funder mandates, and institutional requirements. DataCite[1] is a nonprofit, global member organization featuring over 900 data centers; it has issued over 9 million DOIs for research data. It also develops services to promote data sharing and integrate with other community services (Figure 1). According to principle 3 of the 8 principles in the Joint Declaration of Data Citation Principles, data should be cited in scholarly literature when claims rely upon those data. DataCite serves to link data

CONTINUED

in various ways, such as with other data, with researchers and contributors via ORCID, with articles by using FAIR (Findable, Accessible, Interoperable, Reusable) data-sharing principles, and with funders. The last type of linking is problematic, however, because organization identifiers are lacking. To remedy this situation, ORCID launched an Organization Identifier Working Group in January 2017. The Make Data Count project is spearheading efforts to provide a formal recommendation for measuring data usage and to further develop a data-level metrics hub.

Shelley Stall offered the perspective of the American Geophysical Union (AGU), which is the largest society publisher in the earth and space sciences. The AGU is active in several data engagement activities, including *Earth and Space Science*, its journal for data and methods papers; a data management assessment program for repositories; the AGU data blog; and participation in COPDESS (Coalition on Publishing Data in the Earth and Space Sciences). The AGU data policy2 states, "Earth and space sciences data are a world heritage. Properly documented, credited, and preserved, they will help future scientists understand the Earth, planetary, and heliophysics systems." Best practices for research data include depositing data in a leading domain repository or, if one is not available, in a general repository such as Zenodo, Dryad, or Figshare; using supplemental material for small datasets; ensuring data are publicly available at the time of publication; and citing data or code sets as part of the reference list. Publishers and repositories are beginning to work together, as exemplified by the TOP (Transparency and Openness Promotion) guidelines, COPDESS, and the Joint Declaration of Data Citation Principles. TOP includes modular standards, such as those for citations, data transparency, and preregistration of studies, and 4 implementation levels (0–3). Shelley described next steps for publishers as follows: review the Statement of Commitment on copdess.org; review the TOP guidelines; work with organizations on the need for well-documented, citable data; and work on messaging to authors.

In the final presentation, Kerry Kroffe described how an open data policy was implemented at PLoS. The original PLoS data-sharing policy required authors to share data only if requested after publication. In March 2014, the policy was revised to require authors to make all data underlying a manuscript's results fully available without restriction, except in rare cases. Authors are also required to provide a Data Availability Statement at the time of manuscript submission. Implementation of the new policy for *PLoS One* was challenging because the journal has a wide scope, the request was regarded as controversial, and considerable effort, from staff, the editorial board, and reviewers, was required. PLoS provides clear guidance for contributors in the form of consistently updated FAQs and a listing of recommended repositories. PLoS also set up an external data advisory group and has internal checks in place at submission, during review, at acceptance, and after publication. Since the new policy has been in effect, data sharing has increased from 12% to 40%.

Links

1. https://www.datacite.org/
2. https://sciencepolicy.agu.org/files/2013/07/AGU-Data-Position-Statement-Final-2015.pdf