Implementing a System to Enable Credit for Data Sharing
April 9, 2018

WORKSHOP SUMMARY

This workshop, organized by the Association of American Medical Colleges (AAMC), Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard (MRCT Center), and the New England Journal of Medicine (NEJM), brought together a broad, cross-sector community of 70 participants to discuss and facilitate implementation of a system to appropriately credit academic researchers for sharing data. Through a series of panels and table discussions, academic institutions, journals and publishers, and nonprofit and government funding agencies specifically outlined the actions needed from each stakeholder to support and increase data sharing by implementing this process.

Several foundational concepts were raised throughout the day:

- **Technical solutions**: There was broad agreement that despite some remaining challenges, a technically feasible pathway that would result in credit for data sharing is currently possible. The assignment of persistent identifiers (PIDs) for not only a dataset and individual, but also a grant, funder, or organization, would further tie this system together. One of the outputs of this project is a list of stakeholder actions and responsibilities that need to be followed in order for this technical process to work.

- **Data reusability:** Underlying the idea that researchers receive credit for shared data is the need for the dataset to be shared in accordance with FAIR Data Principles, and be Findable, Accessible, Interoperable, and Reusable. While the meeting did not focus on reusability, this is a critical issue that affects all stakeholders in this process. It may not be reasonable to request uniform data sharing in all scenarios, and it would be helpful to identify the types of data that are valuable or make sense to share by discipline or with input from the research community. Researchers will also need to follow guidelines to ensure that their dataset and associated metadata contain enough information to make the data understandable and reusable.

- **Data openness:** It is important to note that this system for awarding credit is compatible with datasets that are not open upon posting, as with clinical datasets involving potentially re-identifiable personal health information. Data should be as open as possible but as restricted as needed, and requesting these data can be based on a gatekeeper model, e.g. Clinical Study Data Request. As long as the identifiers and citations for these datasets are open, the investigators should receive credit for their work via the same mechanism as they would for unrestricted datasets.

- **Value of data-Level metrics:** Assuming a researcher follows the steps needed to track the use of shared datasets and generate data-level metrics, a question remains as to how to academic institutions and funders will interpret this information. As this is a relatively new way to measure the impact of research,

the community does not have a robust method of understanding these metrics to appropriately award credit, beyond determining whether a dataset has been shared. While an individual dataset may not hold much value, together with other data it can be used to advance a field of study. This type of impact might best be described in a narrative form. Institutions could collaborate on best practice to evaluate academic research performance, including data sharing, or share templates to help applicants to build a story for various disciplines.

- **Role of repositories:** In order to treat data as a first class research product in scholarly communication, repositories need to function as data publishers, with all the attendant requirements of publication (curation, permanence) and benefits (income). A dataset can be published without an article associated with it and still fit into a process for credit. Repositories should also consider instituting better data models to represent contributor roles, to include data scientists and curators so they also can track and receive credit for their work. Repositories will benefit from knowing how the data they host are being used, in publications and other research products, as one of the metrics of their success and a mechanism for getting funding. As both funders and journals are looking to repositories to uphold their policies, formal partnerships and financial support of repositories may be a more sustainable model moving forward.

The focus of the workshop was to define specific actions that need to be taken by journals and publishers, funders, and academic institutions in order to facilitate and implement a system of credit for data sharing. The primary recommendations and outcomes from each of these discussions are presented below.

Journals and Publishers

- **Data policies**: Journals should have a requirement that data is deposited in a trusted repository from an approved list and the link is included at the point of submission into a structured field. Likewise, datasets underlying a publication should also be provided to the journal in a similar format. Datasets should be cited in the references at minimum, and also in the metadata if possible, for Crossref accessibility. Journals should determine what is reasonable to request from researchers in addition to the dataset (e.g. metadata, data dictionary, study protocol, analytic code). These policies should support the technical requirements accepted by the data citation community and outlined in the roadmap for scientific data publishers. Journals should also consider requiring data availability statements, as well as an explanation of individual credit roles in generating the dataset.

- **Culture change and resources**: Journals and publishers can provide leadership and venues to establish common data citation standards and data sharing policies, which would facilitate compliance with these requirements in the research community. Existing efforts such as the TOP (Transparency and Openness) Guidelines and Springer Nature Research Data Policies could form a basis for these discussions. Journals might also commit to publishing some articles that are secondary analyses of existing datasets, together with brief editorials to accompany the articles to provide recognition to this type of scholarship. Beyond just credit in a reference, a publication could specifically note that an analysis is based on shared datasets and explain the role of the dataset in generating the research. As part of the review process, publishers and journals can direct reviewers to ensure that datasets underlying a publication meet certain criteria for quality and reusability.

Funders

- **Data Management Plan:** Funders should require a Data Management Plan (DMP) from researchers, which is actionable, evaluated, and held to. To assist researchers with this process, funders should develop and make available a rubric for scoring DMPs in the grant review process, and link to resources such as DMPTool. If a DMP does not meet the necessary criteria, reviewers should provide feedback on how researchers can amend the DMP (funders may also want to remove DMP evaluation from the primary reviewer who is evaluating the merit of the research, and designate a specific individual to go through a checklist for the DMP or statistical plan). Finally, funders should be clear in that limitations on data access or use should not impact the data sharing requirement, and note that datasets should be uploaded to an appropriate repository to meet these needs.

- **Application process:** Funders should be more explicit that data management, curation, and sharing is an allowable cost of research, and can emphasize this by including a standard line in the budget for these costs and encouraging grantees to include percent effort for a librarian or data scientist. For certain data types, funders may mandate a pre-determined percent for data management costs (as is being instituted in the EU). In evaluating applications, funders should request that the biosketch include information on shared datasets and PIDs, ask for a description of the institutional or departmental support for data sharing in the environment section, and require an applicant to describe why no existing dataset can serve their research needs and score this aspect. Reviewer behavior will also have to be modified—funders can charge the review committee to evaluate shared datasets as valuable research products, or if applicable, use an applicant's data sharing record in scoring applications

- **Progress reports and post-award:** Funders should drive adoption of dataset PIDs by requiring that they are submitted along with other research outputs, and that this information is included in funding management systems. There should also be standards for different types of data to ensure complete metadata and quality of data and the use of an appropriate repository. Funders should also integrate data sharing progress as part of the annual review as a way to institutionalize the expectation for data sharing. It is essential that funders enforce data sharing requirements and follow through on the performance of the grantee. Funders may consider withholding funds to the investigator or institution until data are shared. The heterogeneity of data types would need to be taken into account here, as some data are not ready to share during lifetime of a grant. Funders can also make the data sharing plan for a grant public after the grant is awarded to encourage accountability and compliance with policies.

- **General Funding Considerations:** As a general principle, it is unwise for funders to support data collection without a provision to also support data management, preservation, and sharing. Funders should not only have policies for data sharing, but also institute a framework to ensure that all grantees are compliant. They should celebrate data sharing and reuse when possible, and also consider available research data when planning new funding initiatives. Funders placing a high value on data as a research output and making decisions accordingly can help change the culture away from a sole focus on publications. Right now, many funders have independent declarations around open science and data sharing. Standardizing policies across funders based on practices for each discipline or data type would be incredibly helpful to the research community. Finally, to acknowledge the importance of supporting existing infrastructure for data sharing, rather than just funding new repositories or efforts, funders

might consider supporting common infrastructure for data management rather than including it as an allowable cost in each research grant, and allowing user fees for these centralized data centers as a direct research cost.

Academic Institutions

Institutional Policies and Incentives: Institutions should convene appropriate individuals on campus to develop a comprehensive institution-wide data policy, to include research data, and identify and communicate best practices for data sharing by different roles on campus (student, post doc, early stage investigator, senior faculty, department chairs, administrative leadership, and hiring committee). Institutional leadership should recognize the critical role of data in the changing research environment, and acknowledge that there are additional outputs of scholarly research other than journal articles. As much as possible, institutions should support the data sharing and citation ecosystem, and incorporate evaluation of data sharing into hiring and promotion processes e.g. recruiting faculty based on criteria that includes data contribution. Institutions should also publicize institutional efforts that recognize broader contributions, like new career tracks, and highlight champions who have moved their research programs forward through the use of shared data.

- **Infrastructure:** Libraries could be employed to develop needed infrastructure, especially for data storage, curation, and access, as well as training and support. In most instances, libraries are also responsible for the management of institutional data repositories. Libraries can also assist in identifying and leveraging data that already exists within an institution through new tools such as APIs.

- **Researcher Education:** Principles for good data management, citation, and sharing should be incorporated into scientific training and education, both as a part of graduate curriculums and continuing education for researchers. Institutions should provide investigators with resources on where to locate support for data sharing on campus, and also invest in "training the trainers" who will work with researchers on these issues.

Next Steps

Based on the technical process to facilitate credit for data sharing and the recommendations presented here, we hope to further engage with organizations and individuals working in this space to determine how best to move forward and implement these suggestions. We also encourage the use of these documents to inform any relevant discussions, efforts, or initiatives in the community.

For further resources and online copies of the documents linked here, please visit: www.aamc.org/datasharing.