Executive Summary

The **NIH Virtual Workshop on Data Metrics** was held on February 19, 2020, from 9:00 a.m. to 3:30 p.m. EST. This workshop was entirely virtual and used several online services: Webex for the presentations, Slido and IdeaScale for engagement, and Skype for behind-the-scenes coordination. More than 400 people participated in the workshop from organizations around the globe.

The audience included data resource and repository managers, funders, researchers, and others who are interested in understanding the impact and value of data and data resource infrastructure. Individuals representing diverse biomedical data resources and segments of the biomedical research community presented on principles and existing best practices for measuring data usage, utility, and impact. The data metrics workshop recordings and slides are available on the NIH Office of Data Science Strategy (ODSS) website at datascience.nih.gov/data-ecosystem/nih-virtual-workshop-on-data-metrics.

Workshop Content Overview

The workshop was split into two sessions, with the morning focused on measuring data use and utility and the afternoon focused on use cases. Both sessions centered around three questions:

1. What are the long-term positive or negative consequences of having evaluation metrics for research data?
2. Are there existing standards or methodologies for assessing research data value and reach?
3. How might different stakeholders (data resource users, managers, or funders) use data metrics?

The first session was chaired by Daniella Lowenberg, from the California Digital Library, and was focused on evaluating and measuring data use and utility. This session was introduced by keynote speaker and bibliometrician Dr. Stefanie Haustein, from the University of Ottawa, who presented on “Cautions and Lessons Around Development and Implementation of Scholarly Metrics.” The panel session that followed featured presentations by Ms. Lowenberg on the “Current State of Research Data Metrics”; Dr. Valerie Schneider, National Center for Biotechnology Information, NIH, who spoke on “Perspectives from a Data Center with Multiple Repositories”; Dr. Susan Redline, Harvard University, who spoke on “Perspectives from a Domain-Specific Data Resource”; and Dr. Regina Bures, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, NIH, who presented on “Perspectives from a Controlled-Access Data Resource.”

The second session was chaired by Dr. Warren Kibbe, Duke University, and focused on stakeholder use cases for data usage and utility metrics. Dr. Sean Coady, National Heart, Lung, and Blood Institute, NIH, gave a funder's perspective on "How Do You Use Metrics to Evaluate the Return on Investment (ROI)?" Dr. Robert Moritz, the Institute for Systems Biology, gave a manager's perspective titled "How Do Resource Managers Use Metrics to Articulate the Size, Impact, and Scope of Their Resource, and the Stakeholders of the Resource?" Dr. Brian Byrd, University of Michigan, gave a community perspective and examined "How Do Research Communities Help Demonstrate and Maximize the Utility of a Resource and the Data It Holds? How Can Metrics Promote Usage and Utility of a Resource and Justification for Continued Support?"

Highlights

The workshop focused on two types of evaluation criteria: metrics for data themselves, and metrics for data repositories or data resources. Although these evaluation metrics may be different—for instance, citations may be one indicator for datasets while uptime versus downtime may be an indicator for data resources—the speakers agreed on many characteristics of metrics. Following the keynote speaker, the first session had a strong emphasis on the need for bibliometric principles in the construction and understanding of data metrics. Many questions from attendees centered around the kinds of metrics that currently are available, the facets of metrics that are essential (e.g., metadata), and the degree to which data quality plays a role in supplementing quantitative metrics. All speakers on the first panel agreed that data quality should be evaluated through curation and review, that all data require a persistent identifier for access and citation, and that all data need to be linked together (and cited) through persistent identifiers.

Understanding the utility of research data resources, repositories, and datasets themselves, it was clear through the Q&A that the community needs to better understand user behavior (specific to the field represented by the repository), such as how users access different resources, as well as trends in dataset reuse and citation. Referencing the framework of categories and types of acts referring to research objects in Haustein et al.,[1] Dr. Haustein answered many questions about disentangling user behavior driven by infrastructure versus user behavior driven by intrinsic behavior.

The afternoon panel discussion included a discussion of the role of repositories in supporting data reuse, secondary use, and re-analysis. Although there is no perfect metric or way to track the impact of a dataset through secondary use, over time some reuse will translate into citations. For repositories and resources incorporating clinical and other identifiable data, there are additional hurdles to providing reuse and conforming to the intent of signed informed consent documents. In the afternoon session, a clear distinction also was made between usage, value, and impact. Usage (visits, downloads, registered users, etc.) and training and outreach activities are straightforward to quantify. The value of a dataset often is stated in terms of the cost of replication and the uniqueness of the models/specimens used to generate the dataset. Impact is much harder to directly quantify, and there are different axes for considering impact. Does the accessibility of a dataset change the science being done? Does it accelerate innovation? Is it critical for validation? Is it used as a benchmark or a comparator? Typically, it is difficult to quantify the counterfactual for datasets as well as repositories.

Outcomes

With increased investments in research data infrastructure and data-sharing policies, the need for impact and utility metrics is ever-present. Funders, agencies, infrastructure providers, and researchers invested in repositories and data resources need ways to understand their return on investment, as well as how these investments have affected policy (past and future). An original set of goals for the workshop included understanding what core set of metrics may be appropriate for research data and data resources/repositories. Through the two sessions, representing a diverse set of perspectives, it was clear that the community desires evaluation and impact metrics, but it is not yet ready to decide on what these metrics may be.

Two gaps need to be addressed before the community can commit to a set of metrics: (1) there first needs to be an investment in standardization and normalization practices across data repositories and resources for counting data usage[2] and (2) across the disciplines in biomedical sciences, there is a need for bibliometric studies to understand which indicators best represent reuse and impact, instead of defaulting to chosen metrics now and assigning meaning later. Throughout the workshop, Goodhart's Law[3] was mentioned. All speakers agreed that, with metrics' changing research behavior, it is important that impact metrics not be prescribed without accounting for how researcher behavior may be influenced.

In thinking about future directions for workshop participants interested in the development and identification of core data utility metrics, it is important to consider and support initiatives in the works without "reinventing the wheel." Funded community initiatives focused on the development of social and technical infrastructure for research data metrics—such as Make Data Count,[4] the Research Data Alliance Data Usage Metrics Working Group,[5] Research Data Alliance Scholarly Link Exchange (Scholix) Working Group,[6] ongoing evidence-based bibliometrics studies around research behaviors, and open research incentives like the Data Symbiont Awards[7]—should be joined by biomedical community members interested in contributing use cases and feedback to account for diverse perspectives that may not yet be represented.

[1] Haustein, S., T. D. Bowman, and R. Costas. 2016. “Interpreting ‘altmetrics’: Viewing acts on social media through the lens of citation and social theories,” in Cassidy R. Sugimoto (Ed.), *Theories of Informetrics: A Festschrift in Honor of Blaise Cronin.* Available at https://arxiv.org/abs/1502.05701.

[2] COUNTER. 2020. “The COUNTER Code of Practice for Research Data.” www.projectcounter.org/code-practice-research-data. Accessed April 28, 2020.

[3] Goodhart’s law. “When a measure becomes a target, it ceases to be a good measure.”

[4] Make Data Count. 2020. makedatacount.org. Accessed April 28, 2020.

[5] Research Data Alliance. 2020. “Data Metrics Usage WG.” rd-alliance.org/groups/data-usage-metrics-wg. Accessed April 28, 2020.

[6] Scholix. 2020. “Scholix: A Framework for Scholarly Link eXchange.” www.scholix.org. Accessed April 28, 2020.

[7] The Symbionts. 2016. “The Symbiont Awards: Celebrating the sharing of scientific data.” researchsymbionts.org. Accessed April 28, 2020.