Understanding the Impact of Data

By Kristi Holmes, PhD, Associate Dean for Knowledge Management and Strategy

As the scholarly community places more value on data sharing, datasets are increasingly recognized as "first-class research outputs". Tracking data use (and reuse) allows researchers to more comprehensively understand the impact of their work. In some cases, data sets can have as much (or more) impact as a traditional research article. A well-constructed, discoverable, and connected dataset can play a significant role in subsequent work – serving as trusted benchmarks, inspiring new studies, or signaling knowledge translation.

Data metrics provide context to understand how datasets are accessed and used, including counts of views and downloads, and data citations (i.e., structured references to data as part of a scholarly work).[1] Just as with a traditional research article citation, citing data enables researchers to disclose the foundation upon which their work is built, assess the validity of research findings, and reuse data effectively for new studies. By acknowledging data sources, researchers also recognize people and organizations that collect and share data. This openness strengthens the integrity of research, allowing others to critically assess and build on existing work.

> *A data citation, like a traditional publication citation, acknowledges the use of data in research and has specific elements to provide detailed information about the data source.[2] While the specific format of a data citation might differ depending on where it is used, there are key elements:*
>
> - *Author: individuals or organizations responsible for creating the data set.*
> - *Title: name of the data set, which should be descriptive enough to convey its content and scope.*
> - *Year of Publication: date when the data set was published or made available.*
> - *Version: version number, which is important for tracking updates or revisions.*
> - *Publisher: organization that hosts or distributes the data (e.g., data repository or archive).*
> - *Persistent Identifier: unique and permanent identifier, like a DOI (Digital Object Identifier), that provides a direct link to the data.*

Scholarly tools and platforms that support data publishing workflows, automate data citation, integrate with manuscript submission systems, and track data use are critical. Such tools make it easier for researchers to adhere to data citation guidelines and highlight the resulting impact of their data contributions. Make Data Count (MDC) is a critical initiative to make meaningful data metrics possible through advancements in open infrastructure to enable data use evaluation, outreach to drive awareness and adoption, and evidence on the use and impact of data. MDC collaborated with Generalist Repository Ecosystem Initiative (GREI) repositories to develop a set of recommendations for open data metrics.[3,4] A consistent approach to data citations drives meaningful metrics to enable identification and reporting on the reach of NIH-funded data. GREI repositories and several other platforms (including Northwestern's Prism repository) have adopted the recommendations to ensure that data use can be captured and credited.

Good data curation practices throughout the data lifecycle enable better discoverability, linking, tracking, and credit of data as a research output. These practices enable high-quality data repositories, which, in turn, facilitate data sharing and reuse. When researchers are credited for their data, they are more likely to share it in a usable and well-documented format, enabling a virtuous cycle. Data citation is a critical component of modern research to promote transparency, reproducibility, recognizing the contributions of data creators, and facilitating data reuse. By following established guidelines, using persistent identifiers, and showcasing data outputs in CVs and biosketches, researchers can help ensure that data production and reuse are appropriately credited and incentivized, ultimately advancing collaboration and scientific discovery.

Endnotes

1. Puebla I & Lowenberg D. (2024). Building Trust: Data Metrics as a Focal Point for Responsible Data Stewardship. Harvard Data Science Review, (Special Issue 4). https://doi.org/10.1162/99608f92.e1f349c2

2. Data Citation. https://support.datacite.org/docs/data-citation

3. Puebla I, et al. (2024) "GREI Data citation best practices for repositories". Zenodo. doi: 10.5281/zenodo.10562429.

4. GREI recommendations to support consistent practices to collect, expose and aggregate citations to open data https://makedatacount.org/read-our-blog/grei-data-citation-recommendations/