Tracking Citations and Altmetrics for Research Data: Challenges and Opportunities

by Stacy Konkiel

Research Data Access & Preservation

EDITOR'S SUMMARY

Methods for determining research quality have long been debated but with little lasting agreement on standards, leading to the emergence of alternative metrics. Altmetrics are a useful supplement to traditional citation metrics, reflecting a variety of measurement points that give different perspectives on how a dataset is used and by whom. A positive development is the integration of a number of research datasets into the ISI Data Citation Index, making datasets searchable and linking them to published articles. Yet access to data resources and tracking the resulting altmetrics depend on specific qualities of the datasets and the systems where they are archived. Though research on altmetrics use is growing, the lack of standardization across datasets and system architecture undermines its generalizability. Without some standards, stakeholders' adoption of altmetrics will be limited.

KEYWORDS

altmetrics
research data sets
citation analysis
standardization
access to resources

Stacy Konkiel is science data management librarian at Indiana University. She can be reached at skonkiel<at>indiana.edu.

The recently announced San Francisco Declaration on Research Assessment [1], which calls for the abandonment of the journal impact factor as a means to determine the quality of research, highlights how important and contested the measurement of scholarly impact has become. Measuring impact for research data is also complicated. Data citation itself is not yet a standard practice [2, 3], and there is no authoritative agreement on how and when data should be cited [4]. Altmetrics, which track scholarship's usage on the social and scholarly web, comprise a nebulous group of metrics that use an ever-shifting list of web services' APIs as a source of their data [5]. As with data citations, standards do not yet exist to record or report the impact of different types of altmetrics. In light of these challenges, a panel was convened at the ASIS&T Research Data Access & Preservation Summit 2013 (RDAP13) to discuss new developments in exactly how researchers track the impact of data.

Overview of Data Metrics

Though discussions of data citation practices have occurred since the 1980s, it is in recent years that domain specialists, scientometricians and data curators have attempted to define standards for the citation of data and other data-related metrics. The closest the field has come to defining a standard is establishing DataCite [6], an organization that registers permanent identifiers (PIDs) for data and indexes associated metadata for discovery.

Such standards were the subject of the National Academies' Board on Research Data and Information workshop, "For Attribution – Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop" (2012), a full report of which is available at the National Academies Press website [7]. Various stakeholders, including

researchers, librarians and publishers, put forth their positions on what attribution for data should look like (citation versus varied metrics), what functions it should serve (attribution, showing provenance or defining the impact of researchers overall), how its infrastructure should behave (characteristics of host repositories, executable papers or linked data) and which communities are responsible for its development and implementation (libraries, publishers, data centers or researchers). No single position or suite of recommendations emerged from the meeting nor from a similar meeting, "Bridging Data Lifecycles: Tracking Data Use via Data Citations Data Workshop" [8], held earlier that year.

Other researchers are tackling the problem of tracking impact with a bottom-up approach. The Data Usage Index (DUI) has been proposed for the field of biodiversity, based on a variety of metrics culled from the Global Biodiversity Information Facility (GBIF) repository [4]. The authors call for a move beyond data citations, which mimic the citation of traditional publications, primarily because existing metrics do not "recognize all players involved in the life cycle of those data from collection to publication" nor are they yet standardized. Based on usage logs from the GBIF servers, Ingwersen and Chavan conceptualized a set of measures that are either "absolute" or "relative": number of searched records, download frequencies, number of datasets, download densities and number of searches, to name a few of the 14 metrics. These measures are intended to show value to the researcher and to be used to demonstrate impact in a manner analogous to other altmetrics. While the study has implications for further development of related DUIs, the authors acknowledge that their index is specific to the GBIF repository and therefore not generalizable to all research data repositories.

These limitations are the starting point for the study, "The Product and System Specificities of Measuring Impact: Indicators of Use in Research Data Archives," presented at the 2013 International Digital Curation Conference [9]. The overall aim of the group's research is to develop a suite of metrics that can expose the value that data curators add to a dataset, which in itself is an intriguing concept. The researchers' conceptual framework is especially interesting in that they acknowledge that so-called "specificities" of systems and products – that is, the various sociotechnical factors that influence a system's or an organization's design and development – have more to do with the value of metrics that can be extracted than external factors.

Data curation work is related to both system and product specificities. It is reliant on a system's specificities – architecture and arrangement that dictate how the user can interact with an archive – in that such specificities have an influence on metrics like number of search hits and number unique users who can discover the content. Equally important are product specificities, which are "the qualities and properties of the datasets themselves – their file structure, format and size – that affect the way a user can interact with the archive in consuming and discovering data" [9]. Though the researchers do not go into detail about the effects of particular data-curation activities (such as describing data using metadata standards and controlled vocabularies, reorganizing data for understandability and consumption) on data metrics, the area is tantalizingly open for further study.

Another major study in this area that addresses the various metrics, stakeholders and infrastructure considerations from a 10,000-foot view is the report "The Value of Research Data: Metrics for Datasets from a Cultural and Technical Point of View" accessible at the Knowledge Exchange website [10]. The authors give a rich overview of the challenges and opportunities that lie in capturing metrics for data and report on stakeholder views of the viability of the currently available metrics. Chief among the challenges are culture and infrastructure.

The authors posit that researchers have little reason to value data metrics (including citations) as yet, since they are not considered as valuable as citations to traditional publications. They also have little reason to adopt practices that will enable data metrics to be easily tracked, such as standardized citations for data or the assignment of permanent identifiers such as DOIs (digital object identifiers) to datasets, because the technical infrastructure currently does not support such practices for the most part. This presents a chicken-or-egg conundrum for those developing infrastructure for data citation, which is currently suboptimal, as there does not yet seem to be a need for such an infrastructure, given the lack of interest from researchers. Results from stakeholder interviews and environmental scans inform much of their report.

At RDAP13, Kathleen Fear, University of Michigan; Elizabeth Moss, ICPSR; and Heather Piwowar, ImpactStory/Duke University, presented their work and research related to measuring the impact of data. While all three researchers agreed that data citation is a good way to measure scholarly impact, they also shared their ideas on how to capture a fuller picture of the impact of data, including how the data has been reused and by which audiences.

The Impact of Data Reuse: A Pilot Study of Five Measures

Fear, a PhD candidate at University of Michigan, began the panel by sharing her research into the many ways that citations and usage statistics such as downloads can be used to track various degrees of impact for social science datasets [11]. The impacts boil down into five categories: data reuse, quality of publications that reuse data, diversity of publications that reuse data, size of network stemming from a single dataset and number of unique individuals who download a dataset.

The measurement for the number of times the data has been reused is analogous to how many times a dataset has been cited. While most datasets in Fear's sample had never been cited, many were cited two to 10 times over the course of their lifecycle, with some receiving as many as 30 citations in journal articles. Fear measured the quality and diversity of publications that cite (reuse) the data by determining the citation rates for articles that cite the datasets and the breadth of publications. She noted that reuse rates can be affected by the publications in which a dataset was cited and also by disciplinary differences.

By counting the number of unique individuals who download a dataset, repositories can make general estimates of the data's popular impact. However, we cannot be sure if downloads mean that the dataset has been used in any way, just as we cannot be sure that downloads of journal articles guarantee a paper has been read [9].

The final metric, the size of publication network that stems from a single dataset, is still being researched. The other measures are, interestingly, for the most part all interrelated. Fear found that data reuse counts had little to do with unique downloaders or the data's secondary impact.

The results of Fear's study are interesting, but are they generalizable to all data and data repositories? In our current environment, the answer is, "No." In working with social science datasets culled from the Inter-university Consortium for Political and Social Research (ICPSR), Fear was able to track reuse using the repository's *Bibliography of Data-Related Literature* (which is described in more detail below). The bibliography is, by necessity, a manually curated list; data citation standards have not yet been fully developed or implemented in a way that can automate the tracking for all data held in the ICPSR.

However, in a future where data is cited as strictly as prior publications are cited, one could imagine that Fear's measures of impact take on great importance. Data potentially could have a much broader impact than publications, because they are open to interpretation and analysis: different communities often repurpose data in many different ways with many different results. Determining the scope and quality of that impact could speak volumes about the quality and utility of the data itself.

Viable Data Citation: Expanding the Impact of Social Science Research

ICPSR has done much in the years since its launch to track the citations for data stored in its repository via its *Bibliography of Data-Related Literature* – a manually curated list of more than 60,000 articles that are based in whole or part on findings culled from ICPSR data. In her presentation, the bibliography's chief architect, Elizabeth Moss, stressed the importance of cultivating a culture of data citation: "Impact can be better measured if data use is readily discernible." [12]

Impact is broken down by ICPSR to help understand who uses the data and to what effect. There are certain measures that ICPSR's own website tracks easily: download statistics, unique sessions and users and the names of ICPSR member institutions where downloads of datasets occur. These metrics track who uses the data, while the *Bibliography of Data-Related Literature* more broadly tracks the data's impact in the literature.

ICPSR has engineered some aspects of its repository to encourage citation of both data and related publications, as well as to support different uses by its various audiences. Within the bibliography, literature is searchable

and exportable to reference manager programs. Item records for publications link back to related datasets. This tool can be used in teaching students how to conduct and document their own research, helping researchers perform literature reviews, allowing researchers and funders to track how data is used and enabling reporters and policymakers to see both statistics and the related reports [12]. Digital object identifiers (DOIs) are also issued for data, both at the collection and the study level, with links resolving to the web page with the richest metadata that can help users understand the dataset. These system specificities likely have an effect on how the data is cited and on the other metrics that are collected, as described in the previous paragraph.

Despite ICPSR's efforts to encourage good citation practices, Moss finds that data is rarely explicitly referred to in the literature or discoverable within academic databases. Often, ICPSR staff must comb through articles' methods descriptions and figures to uncover the original dataset a project might be based upon. Most academic databases do not index data – it is simply out of their scope – and current full-text search capabilities are not sophisticated enough for the nuanced search techniques that are currently required to uncover references to datasets. Moss's current strategy to overcome these challenges is to combine text-mining scripts with Google Alerts, which can alert Moss whenever a dataset's creator is mentioned or its DOI is referenced.

ICPSR's recent partnership with the Institute for Scientific Information's (ISI) new *Data Citation Index* (DCI) initiative aims to address some of these issues by integrating its datasets and the *Bibliography of Data-Related Literature*, as well as many other repositories' data and related citations, into the DCI database. Within the DCI, datasets are fully searchable and are treated as research objects that are on par with journal articles, conference proceedings and other traditional outputs. The database search functionality for the DCI as well as related databases like the *Web of Knowledg*e is being converted to meet the needs of those searching for data. As a result, articles can be more easily linked to data, leading to increased data discovery, which is itself a reward for data citation and also rewards those who make their data easily citable – all these benefits from a search interface that many researchers are already using to find emerging research.

Moss concluded by explaining how ICPSR helps "build a culture of viable data citation to improve measures of impact" by providing principal investigators and users with citations, metrics and DOIs for data. Moss encouraged the audience to join groups and attend conferences to advocate for viable data citation practices, including DataCite, iASSIST and the Research Data Alliance. She also advocated that journal editors, domain repositories and funders work together to support repositories and change publishing practices, by requiring authors to better steward and clearly cite the data that underpins their studies.

No More Waiting! Tools that Work Today to Reveal Dataset Use"

Heather Piwowar, co-founder of the altmetrics service ImpactStory, discussed the responsibility of librarians, metrics providers and data scientists to go beyond citations when considering the impact of dataset reuse [13]. Altmetrics can track many types of engagement (views, saves, discussions, formal references and recommendations) that many different types of user groups (researchers, teachers, students, policy makers and practitioners) can have with a single dataset. Those are characterized as "impact flavors," and tools such as ImpactStory, Altmetric.com and Plum Analytics are well suited to help aggregate and display them.

Piwowar laid out three ways in which the community can help encourage more diverse research metrics for dataset reuse: by exposing more metrics, supporting more types of engagement with datasets and lobbying and negotiating for Open Access to research.

Taking the ICPSR and its metrics as an example, Piwowar argued that content providers not only should provide information on dataset usage (downloads and pageviews for descriptive information), but also other rich metrics such as institutions from which a dataset was downloaded and classifications of unique users (into categories such as graduate students, undergraduates, university staff or faculty). However, many repositories do not expose any metrics, especially at the dataset level [10]. It is the responsibility of data curators and repository administrators to expose such metrics.

Secondly, datasets are complicated research products. The scholarly community has not yet figured out an efficient or standardized way to support

peer-reviewed data publications. It follows that scholarly social media sites like Faculty of 1000 or Mendeley would have difficulty addressing datasets and their usage. Piwowar called upon service providers – and altmetrics service providers – to report metrics for all types of engagement with data.

Finally, Piwowar advocated for advocacy itself, as it relates to data metrics. As data curators, librarians, researchers and university administrators, Piwowar argued that it is our duty to lobby and negotiate for open access to research, including open-text mining of articles, open data from repositories and open metrics from aggregators.

Piwowar's last point led to a general discussion of whether repositories like ICPSR should allow commercial, toll-access services such as the DCI to index their metadata, much of which is the result of manual curation. Moss proposed the idea that any exposure to data, whether via the open web or a service like the DCI, is beneficial to the data creator and end user alike. Piwowar, as the founder of a service that relies on open APIs to report metrics, acknowledged that toll-access services and closed APIs inhibit both the ability of end-users to find datasets and platforms such as hers to track their impact.

Summary

Data citations are just one metric that can be tracked to determine the impact of datasets made available through repositories. Altmetrics and usage statistics can determine the impact of data and publications beyond the academy and are useful supplements to citations. The technical infrastructure of repositories and the characteristics of the datasets stored in them can sometimes dictate which metrics can be applied to fully evaluate the impact of data. No metrics can be fully implemented until certain standards, such as DOI usage or commonly agreed-upon best practices for data citation, are widely adopted. Even then, manual intervention to link data to publications and other research outputs may be necessary, making the role of repository staff and librarians ever more essential. ■

Resources Mentioned in the Article

[1] *The San Francisco Declaration on Research Assessment* (DORA): http://am.ascb.org/dora/

[2] Borgman, C. L. (2012). Why are the attribution and citation of scientific data important? In P. F. Uhlir (Raporteur) & Board on Research Data and Information Policy and Global Affairs, National Research Council. *For attribution – Developing data attribution and citation practices and standards: Summary of an international workshop* (pp. 1–10). Washington, D.C.: National Academies Press. Retrieved June 19, 2013, from www.nap.edu/catalog.php?record_id=13564.

[3] Mooney, H., & Newton, M. (2012). The anatomy of a data citation: Discovery, reuse and credit. *Journal of Librarianship and Scholarly Communication, 1*(1), eP1035. doi:10.7710/2162-3309.1035. Retrieved June 19, 2013, from http://jlsc-pub.org/cgi/viewcontent.cgi?article=1035&context=jlsc.

[4] Ingwersen, P., & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): An incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics, 12* (Suppl 15), S3. doi:10.1186/1471-2105-12-S15-S3. Retrieved June 19, 2013 from www.biomedcentral.com/1471-2105/12/S15/S3.

[5] Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Alt-metrics: A manifesto.* Retrieved October 26, 2010, from http://altmetrics.org/manifesto/.

[6] *DataCite:* www.datacite.org/

[7] Uhlir, P. F. (Raporteur), & The National Research Council. (2012). *For attribution – Developing data attribution and citation practices and standards: Summary of an international workshop.* Washington, DC: The National Academies Press. Retrieved June 19, 2013, from www.nap.edu/catalog.php?record_id=13564.

[8] University Corporation for Atmospheric Research (UCAR). (2012). *Bridging Data Lifecycles: Tracking Data Use via Data Citations Data Workshop.* Retrieved June 19, 2013, from http://library.ucar.edu/data_workshop/.

[9] Weber, N. M., Thomer, A. K., Mayernik, M. S., Dattore, B., Ji, Z., & Worley, S. (2013). Indicators of use in research data archives. *8th International Digital Curation Conference (IDCC).* Amsterdam, The Netherlands.

Resources continued on next page

Resources Mentioned in the Article, cont.

[10] Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. (2013). *The value of research data: Metrics for datasets from a cultural and technical point of view.* Copenhagen, Denmark. Knowledge Exchange. Retrieved June 19, 2013, from www.knowledge-exchange.info/datametrics.

[11] Fear, K. (2013). The impact of data reuse: A pilot study of five measures [Powerpoint slides]. *Research Data Access & Preservation Summit.* Baltimore, MD. Retrieved June 19, 2013, from www.slideshare.net/asist_org/kfear-rdap.

[12] Moss, E. (2013). Viable Data Citation: Expanding the impact of social science research [Powerpoint slides]. *Research Data Access & Preservation Summit.* Baltimore, MD. Retrieved June 19, 2013, from www.slideshare.net/asist_org/rdap13-moss.

[13] Piwowar, H. A. (2013). No more waiting! Tools that work today to reveal dataset use [Powerpoint slides]. *Research Data Access & Preservation Summit.* Baltimore, MD. Retrieved June 19, 2013, from www.slideshare.net/asist_org/rdap13-piwowar-tools-that-work-today-to-reveal-dataset-use.