NSF Project Making Data Count: Unit 1 Report

1. Introduction

Data is an indispensable part of research, but it isn't recognized as an important component of a researcher's scholarly output. PLOS, in partnership with UC3 and DataONE, has undertaken a project called Make Data Count (http://articlemetrics.github.io/MDC/) to develop data-level metrics (DLM). This 12-month NSF-funded project is aimed at piloting a suite of metrics that track and measure data use so that it can be shared to funders, tenure & promotion committees, and other stakeholders.

The first phase of this project is to gather information about the needs of researchers– how do they want to get credit for the data they produce? What do they want to know about how their data is used? What do they want to know about others' data to evaluate quality? We connected with the community to determine requirements and understand use cases. This involved several activities where we assessed what values and benefits stakeholders would like to receive from data metrics and any major features or characteristics that they considered critical. We took an iterative approach to requirements gathering to provide richer input into the desired data metrics feature set. Activities undertaken included (1) exploring existing literature and works on this subject, (2) conducting focus groups, (3) distributing an online survey, (4) attending altmetrics events, and (5) developing use cases.

Via these activities, we obtained input from two main stakeholder groups: scientists and data professionals. Scientists polled represented several domains, a spectrum of different institution types, and were at different professional levels. Data professionals were primarily individuals employed by data centers and archives who work directly with data or researchers that produce data.

We adopted a two-pronged approach to build off of previous work and deepen our understanding of the community's needs:

- Quantitative assessment comprised of brief surveys that ask the scientist to identify the most desired characteristic of data metrics.

- Qualitative assessment, focus group interviews used to observe scientists' use of data metrics, ask them about their current perceptions and use of metrics, and identify major features of interest for data metrics.

During these sessions, we assessed what values and benefits stakeholders would like to receive from data metrics and any major features or characteristics that they consider critical. We created use cases for the different types of scientists to better address the challenges associated with different types of data, researchers.

2. Review of previous work

Data sharing & publication

To design meaningful metrics of data dissemination and use, it is essential to understand where researchers go to share data. Fortunately, data sharing behavior has been extensively surveyed. Unfortunately, the major channel of sharing is in response to personal requests by email or other individualized file transfer (Akers & Doty, 2013; Wallis et al., 2013; Kratz & Strasser, 2015). Among the drawbacks to this approach– including the effort involved, the potential for access to be denied capriciously, and the absence of any component of preservation– is that it is invisible to measurement. However, more tractable channels are also in use to considerable extent. After a series of interviews with researchers, Kim & Stanton (2012) distilled the channels of data sharing in use to four main categories: in response to personal requests; via personal, project, or lab website; via upload to an external repository; and as supplemental material published with a journal article.

After email/direct contact, personal or lab websites were the second most frequently mentioned method of sharing in a series of interviews with Center for Embedded Network Sensing (CENS) researchers conducted by Wallis et al. (2013). The publisher Wiley conducted a large cross-disciplinary survey of data sharing practices in 2014 and found that 37% of respondents had shared data via a personal, institutional, or project webpage. Another large, cross-disciplinary survey conducted by DataONE found that 54% of respondents had shared data via an organizational website, 43% via a lab website, and 27% via a personal website.

Upload to a database or repository was the second most used channel in Kratz & Strasser (2015; by 54%), and the third most used channel in Akers & Doty (2013). A cross-disciplinary survey conducted by Wiley found that 26% of respondents had shared data via an institutional repository, 19% via a discipline-specific repository, and 6% via a general purpose repository (e.g., Dryad, Figshare). Aydinoglu et al. (2014) found that 14% of astrobiologists had uploaded data to an open access repository. One-third of the CENS interviewees who mentioned a method for sharing data said that they had done so via a repository (Wallis et al, 2013).

Sharing data as supplemental material attached to a journal article was the most common method in the Wiley survey (which did not include sharing on personal request)– used by 67% of respondents. It was the second most common response to Akers & Doty (2013) and the third most common to Kratz & Strasser (2015).

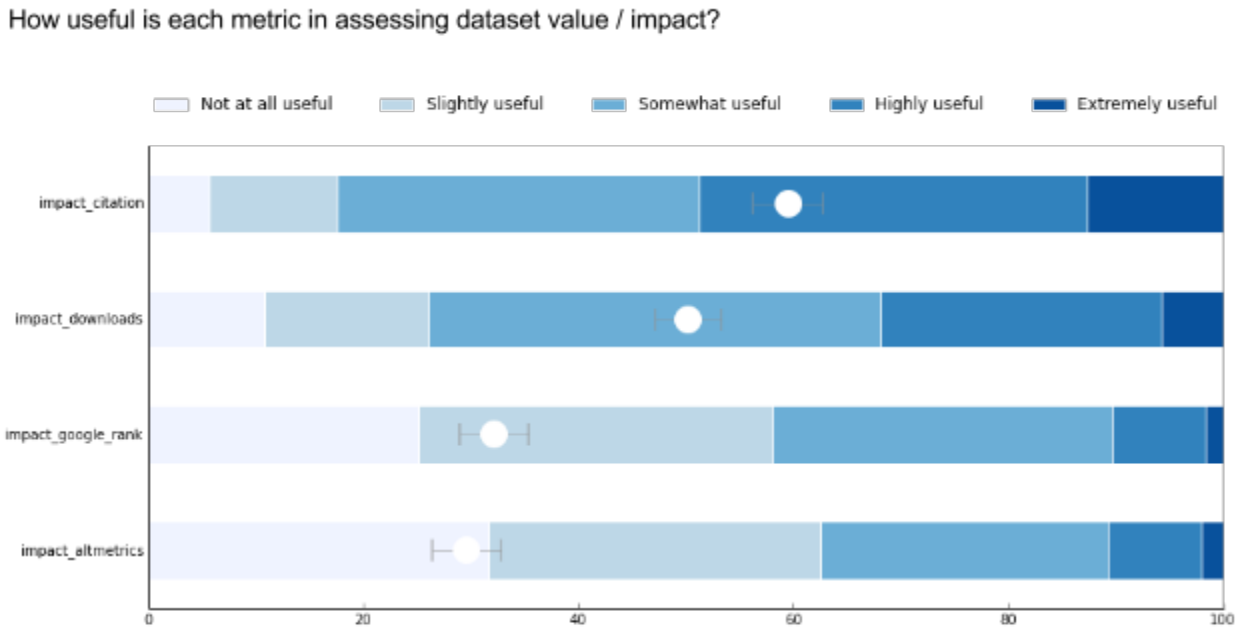


Figure 1. White dots show the mean response for each item; error bars depict bootstrapped 95% confidence intervals. From Kratz & Strasser (2015)

Data metrics

Citations

Citation counts are an established means of evaluating journal articles that can be sensibly applied to datasets. The scholarly communication community recognized the consensus that data used in an article should be cited formally in the reference list with the Joint Declaration of Data Citation Principles (http://www.force11.org/node/4769). Researchers agree: 95% of respondents to the DataONE survey agreed that formal citation was a fair condition for data sharing, as did 87% of astrobiologists in a follow-up survey (Tenopir et al., 2011; Aydinoglu et al., 2014). Citation “in the references like normal publications” is the preferred method of receiving credit for data sharing by 71% of biodiversity researchers (Enke et al., 2012) and by 75% of all respondents to Kratz & Strasser (2015).

Citation counts were seen as the most useful measure of dataset impact in Kratz & Strasser (2015), where 49% of respondents found citations to be highly or extremely useful. To some extent, researchers already track usage of their data in the literature: “details of publications

generated using the data" was the most popular measure used (by 71%) researchers to track use of their own data in a survey conducted by UK Expert Advisory Group on Data Access (EAGDA; Bobrow et al., 2014).

A significant difficulty in using citation counts for data is that despite the consensus, data is cited far less consistently than journal articles. A 1995 survey of 198 social science papers found that only 19% cited the dataset with at least the title in the reference list and, discouragingly, a 2011 follow-up found no improvement in the situation (Sieber & Trumbo, 1995; Mooney, 2011). The most common approach actually seen is an informal description of the source of the data in the text of a paper; 30.8% of papers in 1995 and 69.2% in 2012 included the dataset title somewhere in the text.

Downloads

After citations, download counts were seen as the second most useful measure of dataset impact in Kratz & Strasser (2015): 32% of respondents found citations to be highly or extremely useful. Many (43%) of respondents to the EAGDA survey tracked downloads of their data (Bobrow et al., 2014).

Altmetrics

Altmetrics were far less valued. A 63% majority of respondents to Kratz & Strasser (2015) viewed them as slightly or not at all useful. However, 51% of respondents to the EAGDA survey did pay attention to landing page views (Bobrow et al., 2014).

Researcher attitudes towards altmetrics for journal articles are not well understood (and are likely in flux), and it is not surprising that there is even less information about altmetrics for data.

Relevant conclusions

- Citations are highly valued by researchers as a mechanism for awarding credit and citation counts are seen as a meaningful measure of impact. Complicating matters, current data citation practices are inconsistent at best. But there is broad agreement that this situation should change.
- Download counts are also considered to be valuable, and researchers who currently track use of their data do use them.
- Landing page views are used by researchers to track data use, but altmetrics in general are considered to be of low value.

3. Focus groups

DataONE Focus Group

We convened the first focus group in October 2014 during the DataONE All Hands Meeting in Albuquerque, New Mexico. The DataONE project has a substantial number of individuals involved who work as data managers, are affiliated with data archives, or are otherwise experts in data preservation and sharing. We asked those that were interested in contributing to our project to join us for a one-hour facilitated discussion about data metrics. Attendees included Matt Jones, Bill Michener, David Vieglais, Carly Strasser, Patricia Cruse, Amber Budden, Rebecca Koskela, Bill Michener, Ben Leinfelder, Stephanie Wright, Amy Hodge, David Bloom, Bruce Grant, Bruce Wilson, Gail Steinhart, and others. A total of approximately 25 people participated.

Generally, participants indicated that **researchers** were interested in understanding who is using their data, exactly how they are using it, how it was obtained, and whether any publications resulted from its use. Researchers were interested in similar information regarding others' data they might be interested in using, in addition to information about data quality and restrictions on use. The types of metrics mentioned by participants that might help researchers judge impact or quality of data included citations of associated publications (how many, quality of publication, self-citation instances), data use statistics (number of downloads), and the extent of peer review (formal or informal).

Participants in the group then focused on data metrics issues from the perspective of **archives**. They indicated that repositories were interested in the following types of information about their holdings: impact, usage, citations, click streams, and relationships (with other data, with authors, and with publications). Currently archives were looking at metrics that included related publications, citation counts, and download statistics. Participants indicated that these types of metrics would be used to inform decisions such as deaccession decisions, cost model development, collections development, prioritization of goals, and definition of service tiers.

Society for Neuroscience Sessions

In November, Lin and Strasser attended the Society for Neuroscience 2014 Meeting in Washington DC. They conducted two informal sessions for researchers at the PLOS booth in the conference exhibit hall. The sessions focused on current practices and barriers associated with making data available. Researchers who attended were asked about common repositories for their communities and how they data quality. Although these two small sessions were informative, they primarily provided anecdotal information that will inform the development of data level metrics.

AGU Workshop

In December, we held a workshop at the PLOS offices in San Francisco with researchers in town for the 2014 American Geophysical Union Fall Meeting. Six eminent researchers participated:

- Ben Bond-Lamberty (Pacific Northwest National Laboratory)
- Jim Hansen (Columbia University)
- Andrew Gettelman (University Corporation for Atmospheric Research)
- David Schneider (UCAR)
- Kevin Trenberth (UCAR)
- Maosheng Yao (Peking University, China)

Much of the conversation concerned how to motivate researchers to share data. Sources of external pressure that came up included publishers, funders, and peers. Publishers can require that, at a minimum, the data underlying every figure be available. Funders might refuse to 'count' publications based on unavailable data, and refuse to renew funding for projects that don't release data promptly. Finally, other researchers– in some communities, at least– are already disinclined to work with colleagues who won't share data.

This project is particularly concerned with the inverse– not punishing researchers who don't share, but rewarding those who do. For a researcher, metrics demonstrating data use serve not only to prove to others that their data is valuable, but also to affirm for themselves that taking the time to share their data is worthwhile. The researchers present regarded altmetrics with suspicion and overwhelmingly affirmed that citations are the preferred currency of scholarly prestige.

Many of the technical difficulties with data citation (e.g., citing dynamic data or a particular subset) came up in the course of the conversation. One interesting point was raised by many: when citing a data subset, the needs of reproducibility and credit diverge. For reproducibility, you need to know *exactly* what data has been used– at a maximum level of granularity. But, credit is about resolving to a single product that the researcher gets credit for, regardless of how much of the dataset or what version of it was used– so less granular is better.

4. Online survey

In November and December of 2014, we conducted a pair of online surveys of researchers and data managers (i.e., database or repository staff), asking questions about data sharing, discovery, and metrics.

In recruiting survey respondents, we reached out to researchers from across subject areas, career levels, and affiliations. We used several tools including social media (Twitter feed, blogs), listservs, face-to-face interactions at conferences, meetings, and other venues.

A total of 247 respondents completed the researcher survey (Table 1). Most (78%) are employed by academic institutions. The United States (57%) and United Kingdom (14%) are the best represented nationalities. Researchers from across the academic career spectrum responded; principal investigators, postdocs, and grad students are all well represented.

Biology is the most popular discipline (53%), but environmental (17%) and social (10%) science are also significantly represented. Representatives of 73 repositories completed the data manager survey.

		count	percent
Institution	Academic	173	78%
	Government	20	9%
	Nonprofit	13	6%
	Commerical	10	5%
	Other	5	2%
Country	United States	124	57%
	United Kingdom	31	14%
	Canada	8	4%
	Other	84	25%
Role	PI / Lab Head	93	42%
	Postdoc	47	21%
	Graduate student	41	19%
	Technician / Staff	14	6%
	Other	26	12%
Discipline	Biology	114	53%
	Environmental science	37	17%
	Social science	21	10%
	Computer science	14	7%
	Other	29	12%

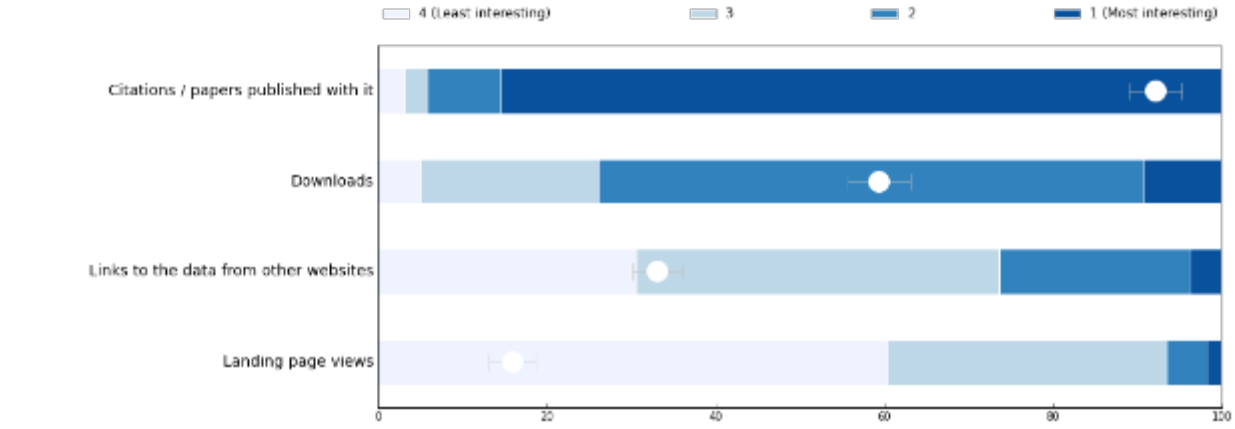
Table 1. Researcher demographics.

Sharing & Discovery

To situate the collection and display of data metrics, we wanted to know where on the internet researchers go to share their data or to discover other's data. Consistent with previous surveys (see section 2), more respondents (90%) shared at least some data by email than by any other method (Figure 2). However, when sharing most or all of their data, more respondents used a database or repository (24%) than email (16%).

A 63% majority of respondents said that they would "definitely" use more than one method to search for data to use. Three of the five methods presented would definitely be used by a majority of respondents: searching via the literature, a discipline-specific database, or a general purpose search engine. When asked to write-in particular sources, more respondents mentioned Dryad (*n*=16) than any other source; Google and "journal articles" were tied for second (*n*=14). Respondents were unlikely to query colleagues via social media or a discussion forum (42% and 40% "no chance" respectively). Email or in-person inquiries to particular colleagues is probably more common, as colleagues were a relatively highly cited write-in source (*n*=12).

A. *Researchers*: How interested would you be to know each of the following about how your data is used?



B. *Data managers*: How interested would you be to know each of the following about how the data you hold is used?

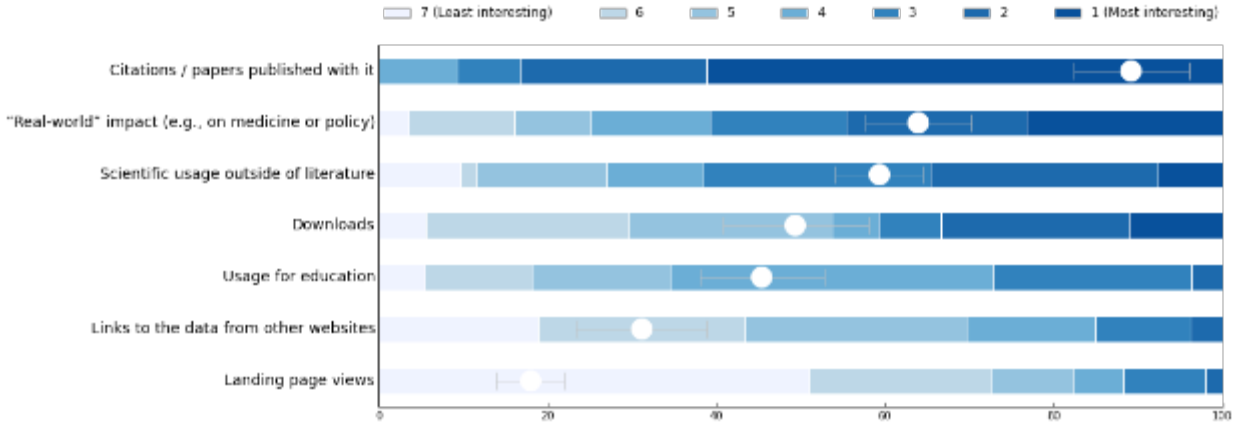


Figure 2. Citations are the most interesting metric of data use to (A.) researchers and (B.) data managers. Respondents were asked to rank their interest in each item. Landing page views were the the least interesting metric to both groups. White dots show the mean response for each item; error bars depict bootstrapped 95% confidence intervals

Measures of impact

Citations are overwhelmingly valued as the coin of scholarly prestige. When asked to rank potential metrics of impact, 85% of researchers and 61% of data managers chose citation count as the most interesting. The preference of researchers is consistent with previous surveys, but it is more surprising that data managers prefer a purely scholarly measure so clearly over “real world” impact (ranked first by only 23%). Download count was a consistent second choice of researchers (by 64%). Landing page views were ranked last by >50% of both researchers and data managers.

Metric collection by repositories

As a practical matter, it is important to understand what metrics are already being tracked and exposed by repositories. We found that most repositories track downloads (85%) and landing page views (66%). However, only 35% of the repositories that track downloads expose them via API or display on the landing page; this ratio is roughly similar for all of the metrics we asked about. Despite the extreme interest in citation counts, relatively few repositories (23%) track them.

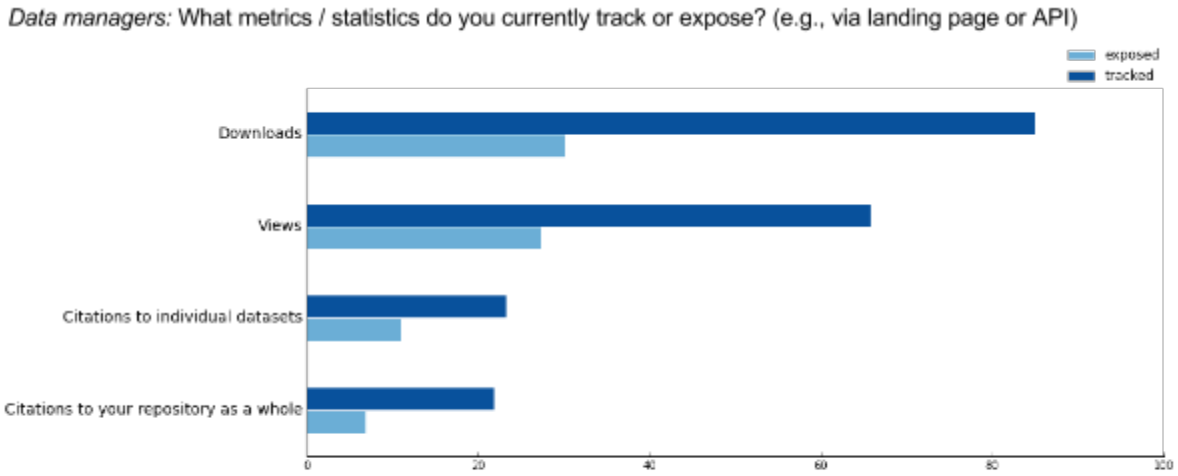


Figure 3. Downloads are the most commonly tracked metric of data use. Data managers reported the metrics they track (dark blue) and expose via landing page or API (light blue). No metric was exposed by more than half of the repositories that track it.

Relevant Conclusions

- Citation count is the metric preferred by both researchers and data managers. Few repositories currently track citations to datasets.
- Download count is a secure second-choice metric for researchers. Unlike citations, downloads are tracked by a majority of repositories.
- Landing page views are of relatively little interest to researchers or repository managers, but are widely tracked.
- More that half of repositories do not expose the metrics they collect via landing page or API.

5. Meetings and workshops on data metrics

While there have been no meetings and workshops on data metrics during the Unit 1 work period, two meetings occurring prior are worthy of note. At the Research Data Alliance Fourth Plenary (https://www.rd-alliance.org/plenary-meetings/rda-fourth-plenary-meeting.html), the Publishing Data Bibliometrics working group led by Sarah Callaghan and Kerstin Lehnert gave a report of the group's current progress investigating requirements for and

recommending necessary steps to be taken for developing data bibliometrics across the following areas: organizational/ecosystem, technical, methodological, and financial.

Secondly, the 1:AM Altmetrics Conference (http://www.altmetricsconference.com/) was held in London on September 25-26, 2014 and brought together a wide cross-section of funders, publishers, librarians, researchers, and technology providers to discuss how altmetrics are integrated into scholarly communication processes — and the challenges that they face along the way. While the conference mostly addressed the online social activity surrounding research articles, one panel session was focused on altmetrics for data and software. Sarah Callaghan represented the RDA working group here once more. Daniel Katz, NSF Program Director, Division of Cyberinfrastructure, presented on NSF's efforts to value software and other research outputs in their funding and evaluation activities.

6. Use cases

- As a data generator, I want to provide evidence of the impact of my research output.
- As a researcher, I want to know what data is being used in scholars in my field and in new fields, which are of interest to my research.
- As a researcher, I am interested in understanding who is using my data, exactly how they are using it, how it was obtained, and whether any publications resulted from its use.
- As a researcher, I am interested in similar information regarding others' data they might be interested in using, in addition to information about data quality and restrictions on use.
- As a data manager, I am interested in the impact and usage of the data in my repository.
- As a data manager, I am interested in the click streams and relationships (with other data, with authors, and with publications) of the data in my repository.
- As a data manager, I want data to inform decisions such as deaccession decisions, cost model development, collections development, prioritization of goals, and definition of service tiers.

7. References

Aydinoglu, Arsev Umur, Todd Suomela, and Jim Malone. "Data Management in Astrobiology: Challenges and Opportunities for an Interdisciplinary Community." *Astrobiology* 14, no. 6 (May 19, 2014): 451–61. doi:10.1089/ast.2013.1127.

Bobrow, Martin, James Banks, Paul Burton, George Davey Smith, Rosalind Eeles, Paul Flicek, Mark Guyer, et al. *Establishing Incentives and Changing Cultures to Support Data Access*. Wellcome Trust, May 2014.

http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_peda/documents/web_document/wtp056495.pdf

Enke, Neela, Anne Thessen, Kerstin Bach, Jörg Bendix, Bernhard Seeger, and Birgit Gemeinholzer. “The User’s View on Biodiversity Data Sharing — Investigating Facts of Acceptance and Requirements to Realize a Sustainable Use of Research Data —.” *Ecological Informatics*, Data platforms in integrative biodiversity research, 11 (September 2012): 25–33. doi:10.1016/j.ecoinf.2012.03.004.

FORCE11 Data Citation Synthesis Group. “Joint Declaration of Data Citation Principles.” The Future of Research Communications and Scholarship, 2014. http://www.force11.org/node/4769.

Ingwersen, Peter, and Vishwas Chavan. “Indicators for the Data Usage Index (DUI): An Incentive for Publishing Primary Biodiversity Data through Global Information Infrastructure.” *BMC Bioinformatics* 12, no. Suppl 15 (December 15, 2011): S3. doi:10.1186/1471-2105-12-S15-S3.

Kim, Youngseek, and Jeffrey M. Stanton. “Institutional and Individual Influences on Scientists’ Data Sharing Practices.” *Journal of Computational Science Education* 3, no. 1 (June 2012): 47–56.

Kratz, John Ernest, and Carly Strasser “Researcher perspectives on publication and peer review of data.” *PLOS ONE*, in press.

Mooney, Hailey. “Citing Data Sources in the Social Sciences: Do Authors Do It?” *Learned Publishing* 24, no. 2 (2011): 99–108. doi:10.1087/20110204.

Sieber, Professor Joan E., and Bruce E. Trumbo. “(Not) Giving Credit Where Credit Is Due: Citation of Data Sets.” *Science and Engineering Ethics* 1, no. 1 (March 1, 1995): 11–20. doi:10.1007/BF02628694.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. “Data Sharing by Scientists: Practices and Perceptions.” *PLoS ONE* 6, no. 6 (June 29, 2011): e21101. doi:10.1371/journal.pone.0021101.

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. “If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology.” *PLoS ONE* 8, no. 7 (July 23, 2013): e67332. doi:10.1371/journal.pone.0067332.

Wiley. *Researcher Data Sharing Insights*, 2014. http://exchanges.wiley.com/blog/wp-content/uploads/2014/11/Researcher-Data-Insights-Infographic-FINAL-REVISED-2.jpg.