Harvard Data Science Review

Measuring and Mapping Data Reuse: Findings From an Interactive Workshop on Data Citation and Metrics for Data Reuse

Lisa Federer

Published on: Apr 30, 2020

Updated on: Mar 26, 2020

DOI: 10.1162/99608f92.ccd17b00

ABSTRACT

Widely adopted standards for data citation are foundational to efforts to track and quantify data reuse. Without the means to track data reuse and metrics to measure its impact, it is difficult to reward researchers who share high-value data with meaningful credit for their contribution. Despite initial work on developing guidelines for data citation and metrics, standards have not yet been universally adopted. This article reports on the recommendations collected from a workshop held at the Future of Research Communications and e-Scholarship (FORCE11) 2018 meeting titled Measuring and Mapping Data Reuse: An Interactive Workshop on Metrics for Data. A range of stakeholders were represented among the participants, including publishers, researchers, funders, repository administrators, librarians, and others. Collectively, they generated a set of 68 recommendations for specific actions that could be taken by standards and metrics creators; publishers; repositories; funders and institutions; creators of reference management software and citation styles; and researchers, students, and librarians. These specific, concrete, and actionable recommendations would help facilitate broader adoption of standard citation mechanisms and easier measurement of data reuse.

Keywords: data citation, data metrics, data publication, open science, research data

1. Introduction

The sharing of research data has become an increasing priority over the past decade across a range of scientific communities, with governmental agencies, public and private research funders, research journals, and open science organizations supporting the development of policies to encourage or require sharing research data. The benefits of data sharing include facilitating reproducibility and enhancing transparency of published results (Alquraishi & Sorger, 2016; Leung et al., 2007; Martens et al., 2016), promoting collaboration and deriving new insights from existing data sets through secondary analysis (Pasquetto et al., 2019; Rung & Brazma, 2013; Zimmerman, 2003, 2008), and accelerating the pace of innovation (Knoppers, 2014; Knoppers et al., 2014).

As data sharing has gained traction across a range of scientific disciplines, the need for standard mechanisms for citing data and tracking its reuse has emerged. Citations are foundational to scholarly literature, used to track provenance of ideas (Dempsey, 2004), provide attribution and credit to authors (Garfield, 1964a), establish connections between related research outputs (Garfield, 1964b), and provide a means to measure a researcher's scientific impact (Carpenter et al., 2014). If data are to be recognized as a "first class scientific output" (Callaghan et al., 2012), mechanisms must likewise be developed to track the use of data sets (and ultimately, other research products such as code and software) through citation. Data citation has additional importance given some researchers' reluctance to share their data; as mechanisms to cite secondary use of data in ways that are easy to track and ensure that data creators receive credit for their contribution. In the absence of a means to comprehensively track data reuse, funders, research institutions, and others have few means to reward data creators whose data sets have been highly used.

Recognizing the importance of data citation and the need for standards and best practices, the Future of Research Communications and e-Scholarship (FORCE11) created a Data Citation Synthesis Working Group, along with 25 other organizations, in 2014 (FORCE11, 2014). That group's goal was to synthesize the outputs of various groups that had issued statements on data citation, including the Committee on Data for Science and Technology (CODATA)-International Council for Scientific and Technical Information (ICTI) task group on Data Citation Standards and Practices (CODATA, n.d.; CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013), to create a unified set of principles to guide best practices for data citation. The result was the Joint Declaration of Data Citation Principles (JDDCP), which outlines eight overarching principles informing the "purpose, function and attributes of [data] citations" (Data Citation Synthesis Group, 2014). However, nearly five years later, the vision of machine-actionable data citations with persistent identifiers (PIDs) "accorded the same importance in the scholarly record as citations of…publications" has not yet been fully realized (Data Citation Synthesis Group, 2014). While some publishers and journals have adopted data citation policies, data citation remains far from standardized (Zhao et al., 2017). Further, many research communities have not seen a significant uptake of data citation; for example, a 2015 study of data sets in the Data Citation Index found that only one of the nearly 45,000 data sets in the humanities and seven of the nearly 425,000 data sets in the social sciences had ever been cited (Robinson-García et al., 2015). While some computational methods have been developed to locate articles that report using data sets, even if they do not formally cite them (Piwowar, 2010; Zhang et al., 2016), most work on quantifying and tracking data reuse thus far has relied largely on time-consuming manual methods (Belter, 2014; Callahan et al., 2018; Piwowar et al., 2011). Such methods are not applicable to the large-scale, systematic tracking of data reuse. Without the ability to reliably and exhaustively track data citations, data metrics cannot be considered a meaningful measure of data use or impact (Stuart, 2017).

As part of the effort to increase data citation, FORCE11 created a Data Citation Implementation Pilot (DCIP) to coordinate efforts of various stakeholders adopting the JDDCP, including publishers and repositories (FORCE11, n.d.). Their activities included creating data citation "roadmaps" that would guide various stakeholders in implementing data citation based on the JDDCP. The roadmap for scientific publishers provided both technical guidance (such as how to format citations and what type of identifiers to use) as well as guidance aimed at changing the culture around data citation (such as providing training for reviewers and developing policies for data sharing and citations) (Cousijn et al., 2018). A second roadmap for scholarly data repositories primarily focused on specific technical guidelines, such as recommendations on the use of persistent identifiers and the type and format of metadata to collect (Fenner et al., 2019).

Parallel to such work that aims to improve the practice of data citation, efforts are also underway to develop infrastructure and frameworks for making connections between data sets and the articles that cite them. The Scholix (SCHOlarly LInk eXchange) framework, developed by the Research Data Alliance/World Data Service Publishing Data Services Working Group with consensus from stakeholders in the research data and publishing communities, provides a set of guidelines for interoperability between systems that will enable better literature/data linkages (Burton et al., 2017; Research Data Alliance, 2019). The Make Data Count Project also aims to develop infrastructure to enable tracking data citations, as well as establish practices and metrics for meaningfully quantifying data usage (Cousijn et al., 2019; Make Data Count, 2019). The latter is addressed in the Counting Online Usage of Networked Electronic Resources (COUNTER) Code of Practice for Research Data Usage, which "standardizes the generation and distribution of usage metrics for research data" (Fenner et al., 2018). As data citation becomes more standardized and adopted across disciplines, these projects will be valuable in tracking uses of data in the literature.

This report contributes to these various efforts to increase data citation and develop metrics by describing the output of a 2018 workshop that convened participants representing a range of stakeholders involved in data citation and mechanisms for quantifying data reuse. The goal of the workshop was to generate suggestions for how various stakeholders could help move the scholarly community forward in these areas. Some of these suggestions involved ways to establish broader adoption of standardized data citation that would enable tracking data reuse, while others involved temporary means to help track data reuse until that broad adoption is realized.

2. Workshop Format

"Measuring and Mapping Data Reuse: An Interactive Workshop on Metrics for Data" was organized and conducted by the author as a 1.5-hour preconference workshop at the FORCE2018 meeting in Montreal on October 10, 2018. The first 20 minutes consisted of a presentation by the author detailing her research on the usefulness of data citations for effectively tracking data reuse. In short, this research demonstrated that existing mechanisms for tracking data reuse through data citation likely underestimate true reuse counts, since many researchers are not citing data reuse using a common citation standard (or failing to cite data reuse altogether). These findings suggest that effectively measuring and quantifying research data reuse will require various communities and stakeholders to work together to find solutions, both technological and cultural, that enable better automatic tracking of data reuse.

The last hour of the workshop was spent engaging participants in a collaborative brainstorming activity to develop potential solutions that research communities could adopt to facilitate measuring and mapping data reuse. The workshop comprised 32 participants representing a range of disciplines and backgrounds, including publishers, researchers, funders, repository administrators, librarians, and others, and came from a range of disciplinary backgrounds and geographic locations. These individuals self-selected to participate by registering for the workshop. Participants were asked to individually brainstorm ideas for actions that could be taken next to improve the state of data citation and tracking data reuse and write each idea on a notecard. Participants then discussed their findings in small groups, and the notecards were collected by the author at the close of the workshop to prepare this summary report.

3. Findings

Workshop participants generated a total of 68 ideas for how various stakeholders could address the problem of measuring and tracking research data reuse. Many of the ideas were suggested by multiple participants. Here, these suggestions are synthesized and organized by the stakeholder group they address. All 68 suggestions are represented in the synthesized list, although a small number of specific suggestions were counted with the relevant higher level suggestion. For example, a suggestion that the National Institutes of Health (NIH) mandate data sharing was counted as part of the broader suggestion about funders in general requiring data sharing.

Table 1 provides a brief summary of recommendations by stakeholder type, including how many times each suggestion was raised, and is followed by a full discussion of recommendations and their implications. To respect participants' privacy, neither names nor stakeholder identity of the participant making each suggestion were collected.

Standards and metrics creators **(*n* = 15)**	1. Clarify granularity and versioning in defining what constitutes a single data set. (*n* = 1) 2. Standard for citation that is machine-readable and includes a PID. (*n* = 1) 3. Standards for metadata (including minimal metadata for citation). (*n* = 11) 4. Enforcement of standards. (*n* = 1) 5. Metrics for quantifying and understanding data reuse. (*n* = 1)
Publishers **(*n* = 18)**	1. Follow and enforce standards for citation, including machine-readability. (*n* = 12) 2. Educate editors and peer reviewers and to evaluate data citation as part of peer review. (*n* = 3) 3. Provide enhanced full-text searching of articles to make it possible to find data sets that haven't been properly cited. (*n* = 2) 4. Use text mining to identify potentially citable objects in submitted articles. (*n* = 1)
Repositories **(*n* = 13)**	1. Issue DOI (or other PID) to all data sets. (*n* = 2) 2. Ensure data sets have adequate minimal metadata. (*n* = 3) 3. Require information from data downloaders about type of reuse. (*n* = 3) 4. Create a mechanism to allow reusers to register their reuse. (*n* = 1) 5. Create mechanisms that allow data submitters to track downloads and reuse of their data. (*n* = 2) 6. Implement a citation export for data sets to facilitate accurate citation. (*n* = 2)

Funders and institutions **(*n* = 9)**	1. Include data sets in the definition of recognized research outputs. (*n* = 2) 2. Give credit to researchers for sharing. (*n* = 4) 3. Mandate data sharing. (*n* = 2) 4. Issue a ‘call to action’ to relevant stakeholders. (*n* = 1)
Creators of reference management software and citation styles **(*n* = 8)**	1. Include data citation format in citation style guides, specifying inclusion of DOI or other PID. (*n* = 3) 2. Incorporate data citation style into reference management software. (*n* = 5)
Researchers, students, and librarians **(*n* = 6)**	1. Ensure shared data is accompanied by adequate metadata to facilitate citation by potential reusers. (*n* = 1) 2. Provide training and mentoring on proper mechanisms for data citation. (*n* = 5)

Table 1. Brief Summary of Recommendations by Stakeholder Group

3.1. Creators of Standards and Metrics

About 22% (*n* = 15) of the suggestions were broad recommendations aimed at groups that define standards and metrics related to data sharing and citation. Such groups could include a range of stakeholders, such as professional societies, funders, repositories, and publishers. While many of these stakeholders are addressed in greater depth elsewhere in this report, the suggestions in this section speak specifically to activities related to development of standards and metrics.

At the most fundamental level, participants suggested that the definition of a data set should be more clearly established. While such a basic question may seem trivial, previous research has demonstrated that, in fact, there is a lack of clarity about what constitutes a data set. For example, in a 2015 study that aimed to quantify the number of ‘invisible’ or uncited data sets in published research, pairs of annotators who independently counted data sets per article differed substantially on the numbers they identified (Read et al., 2015). That study elucidates some of the confusion that arises in defining a data set; as they note, “[d]epending on one’s perspective, a single data set could be: all of the data that is collected or used in a study; all data collected at a specific time within a study; pre- or postintervention; a discrete type of data from a specific diagnostic device; or even every individual measurement reported in a research article" (Read et al., 2015, p. 40). Further complicating the issue is whether later versions of previously issued data sets should be considered as part of the earlier data set or as its own separate data set. Before considering how to cite a data set, these questions about what comprises a data set must be answered. The most appropriate ways to handle granularity and versioning likely differ across different types of data and different fields of research, so these questions are likely best considered on a disciplinary basis, potentially by relevant professional societies, funders, repositories, or journals. Ideally, such decisions would incorporate input across all these actors within a particular discipline.

Participants also pointed to a need to define a standard for data citation, and relatedly, define metadata standards that ensure that data sets have the minimal metadata essential for citation. The Joint Declaration of Data Citation Principles stops short of recommending a specific citation format, noting that "practices vary across communities and technologies will evolve over time" (Data Citation Synthesis Group, 2014). Despite the varied needs of different communities, a minimal standard for data citation could be defined that would apply across research disciplines broadly. Article citation styles differ from one discipline to another, and even one journal to another, but despite the specific format, most include a minimal standard set of information like author name, article title, year of publication, and the like. Similarly, a minimal standard for data citation could be adapted to specific community needs in terms of form. Specific disciplines and repositories may have additional criteria for their data set metadata schemas, which they could add to the common minimal standard.

Participants also pointed to two essential characteristics of data citations: the need for PIDs for data sets and machine-readability of citations. Together, these two requirements ensure that readers can easily find the cited data set and that citations to data sets can be tracked and counted. Standardized, machine-readable forms of data citation have the benefit of enabling tracking of instances of data reuse within scholarly literature in much the same way that citations to articles can be automatically tracked and counted. In the absence of a means to unambiguously identify a specific data set and automatically locate articles that cite it via machine-readable citations, quantifying reuse of data sets is virtually impossible.

Beyond just developing standards, participants also saw a potential role for standards creators in enforcing them, although their enforcement ability would likely depend on their specific role within the research community. For example, funders, publishers, and repositories could feasibly enforce standards through policy, but professional organizations and others may not have the means to ensure compliance with standards. Previous studies of data-sharing policies have suggested that researchers often do not adjust to these new policies on their own (Couture et al., 2018; Federer et al., 2018; Naudet et al., 2018; Rowhani-Farid & Barnett, 2016), and may require oversight from appropriate stakeholders to ensure that new standards are followed, at least until such practices become the norm within scientific communities.

Once mechanisms exist to reliably track data citations and thereby quantify reuse, the more difficult question of developing quantitative metrics for counting these citations must be addressed. An entire field of study—bibliometrics—has arisen to quantify and understand the impact of research outputs, primarily in the form of publications. A range of metrics and methods exist to quantify research outputs to recognize and reward authors of high-impact research, some of which are used in tenure, promotion, and funding decisions (Carpenter et al., 2014). Along with the proliferation of metrics for research impact have come cautions to avoid misapplying them in ways that potentially distort their meaning (Hicks et al., 2015). Metrics creators should be guided by previous work on measuring other types of research impact in considering how to develop metrics for meaningfully quantifying the reuse and impact of data sets, as well as defining a set of best practices for applying and interpreting those metrics.

3.2. Publishers

About 26% (*n* = 18) of the suggestions were directed toward publishers, who were viewed as playing an important role in promoting standards for citation by requiring that authors cite their data in a standard format. Around the same time as this workshop was held, an article outlining a “data citation roadmap” for publishers was released, containing some of the concepts that were mentioned during the workshop (Cousijn et al., 2018). Participants highlighted the importance of ensuring machine-readability of citations by clearly identifying those citations that refer to data sets. For example, Wiley’s Data Citation Policy specifically notes that data citations should appear in an article’s reference list preceded by the notation “[dataset]” to ensure they are correctly marked as referring to data rather than a publication (Wiley, 2018). This step allows the article to be automatically connected to the data set it cites, making it possible to get accurate counts of reuse without having to resort to time-consuming and ineffective manual measures.

Providing clear guidance to authors on data citation is a helpful first step, but citation policies must also be consistently applied to ensure that authors are correctly citing all relevant data sets. Workshop participants saw a role for journal editors and peer reviewers in ensuring data sets are properly cited as part of the peer review process. As many peer reviewers may themselves be unfamiliar with data citation practices, publishers and editors would need to consider how to educate reviewers on how to incorporate evaluation of data citation into their review. Some participants also suggested using text mining to identify references to citable data sets within submitted articles. Such a method would require the development of text mining algorithms capable of effectively identifying language that referred to citable data sets and would likely be most effective as a complement to rather than replacement for human review.

While solutions involving policy and article review can be effective in ensuring more standard and widespread data citation prospectively, the scientific literature includes millions of already published articles that may be missing citations to data. Participants saw text mining as a solution for the published literature as well, potentially by programmatically searching the published literature for data set identifiers like digital object identifiers (DOIs) or accession numbers. Comprehensive identification of data sets that have not been formally cited relies on the ability to search the full text of articles, since most mentions of data sets occur within the paper itself, not fields like the title or abstract that are searchable by most bibliographic databases (Federer, 2018). Participants suggested publishers could facilitate the identification of uncited data sets by providing broader full-text searching capabilities for their articles and exposing full text of articles to search application programming interfaces (APIs).

3.3. Repositories

About 21% (*n* = 14) of the suggestions were related to repositories. Many of the suggestions for repositories are also echoed in an article published several months after the workshop, which proposes a "data citation roadmap" for repositories (Fenner et al., 2019). Some of these suggestions involved actions that repositories could take to help ensure that data users cite data sets in standard formats, including all relevant information. First, repositories should create a PID for each data set, ideally in the form of a DOI. In the absence of a PID, readers may have difficulty locating the original data set an article cites, and tracking citation for the purposes of quantifying reuse is challenging (if not impossible). Repositories should also ensure that data sets have minimal metadata needed for a citation. As discussed above, the specifics of what constitutes minimal metadata may differ from one discipline to another; therefore, repositories should work with the communities they serve to ensure that their metadata meet these minimal standards for citation. To make data citation even easier, repositories could implement a citation tool, as is often used in the context of journal articles, allowing data reusers to simply copy and paste the proper citation or download the citation file to their reference manager of choice.

A second category of suggestions involved mechanisms by which repositories could facilitate tracking instances of reuse of data from their repository. While many repositories allow free and open download of data without requiring users to apply for access, some participants suggested that repositories could gain a better understanding of how data sets were being reused by requiring downloaders to provide information about how they intend to use the data set. Some repositories with restricted access already have such a requirement, but open repositories may find it undesirable to adopt such a policy, which could hinder free access to data sets and possibly even discourage reuse. Alternatively, repositories might rely on a voluntary system, either accepting input about potential reuses upon download, or creating a mechanism for secondary data users to “register” their reuse. Ideally, the practices and infrastructure needed to fully automate tracking of data set reuse through data citation will eventually be widespread, but in the meantime, a voluntary method for reporting reuse of a data set to the repository could help fill some of the gaps in our understanding of how (and how much) data sets are being reused. Soliciting user input about intended reuse would have the added benefit of enabling tracking of a broader range of use and reuse activities, including those that do not end up being cited in the literature (van de Sandt et al., 2019). For example, data sets may be used in teaching, instrument calibration, model validation, and other types of use and reuse that would not likely be reported in the literature and therefore not be captured by methods that track citations (Gregory et al., 2019). Whether the link between data set and citing article is made manually or through automated means, participants suggested that repositories could provide links to articles citing data sets, just as resources like PubMed Central, Google Scholar, and Web of Science provide links to citing articles for a given publication. Such a mechanism would be helpful for data creators who would like to demonstrate the impact of their shared data and understand how others are reusing their data sets. Further, showing the potentially varied ways a data set has been reused could inspire further reuse in new contexts.

3.4. Funders and Institutions

About 13% (n = 9) of the suggestions related to research funders and academic or research institutions. While these recommendations do not suggest actions that would directly enable tracking data reuse, they draw upon funders’ and institutions’ authority to mandate actions or practices that would contribute to a research data ecosystem in which such tracking is both feasible and useful. Policies mandating data sharing could prescribe certain methods for sharing that would facilitate use by other researchers and tracking of that reuse. For example, rather than allowing data creators to make their data available only upon request, institutions or funders might specify in their policies that data creators should deposit data in a repository that meets certain criteria, such as providing a PID or a set of minimal metadata needed to facilitate proper citation.

Funders and institutions could also take approaches that encourage, rather than simply require, researchers to share their data by recognizing shared data as a research output and creating incentives for sharing high-value data sets. At a basic level, funders could allow researchers to include data sets as examples of their research output, both in progress reports and in proposals for funding. Some U.S. funders have already taken such steps; for example, both the National Science Foundation (NSF) and NIH explicitly include data sets in the list of research products that may be used to demonstrate progress toward funded grant aims and exemplify researchers' scientific impacts (NIH, 2017; NSF, 2016). Similarly, some academic institutions have begun adopting policies aimed at rewarding data sharing. For example, the Montreal Neuroscience Institute (MNI) has adopted an institution-wide open science policy that includes rewarding open sharing of data and other research products in the tenure and promotion process (Ali-Khan et al., 2018).

Finally, some participants suggested that funders in particular could play an important role as catalysts for change in data citation. Funders are in a unique position to set a course for research communities through their policy and funding priorities. Participants suggested that funders could further drive the move toward implementation of data citation standards and acceptance of data as a quantifiable research output by issuing a 'call to action' to other relevant stakeholders.

3.5. Creators of Reference Management Software and Citation Styles

About 12% (*n* = 8) of the recommendations addressed the individuals and groups that create citation styles and the reference management software that many authors use to create the citations and reference lists. First, the organizations and bodies that create style guides should provide explicit direction about how to cite a data set. As discussed above, the specific format will likely vary from one discipline to another, but citation styles should include at least minimal information to allow a reader to locate the cited data set, particularly by including a DOI or other PID. Some citation style guides, such as the one by the American Psychological Association (APA), have already taken steps to clarify data set citation style, including noting that a DOI should be used when one is available. Participants also suggested that developers of reference management software should ensure that they include data citation as a unique reference type, incorporating the standards of specific citation styles where possible, and including fields for the relevant minimal metadata in generic styles or in styles where a standard does not yet exist.

3.6. Researchers, Trainees, and Librarians

Participants also saw a role for researchers, including trainees, and librarians in facilitating data citation, with 9% (*n* = 6) of suggestions aimed at these groups. Although researchers, trainees, and librarians are somewhat diverse as a set of stakeholders, they are grouped here based on the small number of suggestions for each and their similar role, as either data sharers/citers in the case of researchers and trainees or facilitators of data sharing/citation in the case of librarians. Regardless of where or how researchers share their data, they can help encourage citation by ensuring that their data sets are accompanied by adequate metadata to enable a secondary reuser to cite the data set. If researchers are not depositing in a repository that already does so, researchers may also choose to provide a suggested or preferred citation to ensure that potential reusers will cite the data set in a way that can be recognized (including by automated systems tracking citations). When researchers are in a position to choose from various repositories in which to deposit their data, they may wish to opt for one that provides not only the minimal metadata necessary for citation, but also a persistent identifier, preferably a DOI.

The practice of reusing data sets from repositories is a relatively recent phenomenon in many scientific disciplines that have not previously shared data widely, and even in disciplines more accustomed to sharing, data citation has generally not been standardized. Moving toward a scientific culture in which data sharing can be rewarded requires researchers to adapt their practices to ensure that they are sharing their data in ways that facilitate citation, and, in turn, correctly citing data sets that they reuse. As new standards for citation and tracking develop, both current researchers and students will need to learn how to incorporate these practices into their work. Given their role within scholarly communication, librarians are particularly well-situated to provide training and outreach to help data citation gain traction within the scientific communities with which they work.

4. Conclusion

As the research that led to this workshop presented, tracking and quantifying data reuse on a large scale remains challenging. With automated methods for doing so, like those coming out of Scholix and Make Data Count, not yet fully able to retrieve citations in light of the lack of standardization, it is difficult to reward researchers who share their data, an important step to incentivizing and encouraging data sharing. While this problem presents a significant challenge to scholarly communities, participants in the workshop were able to present specific, concrete, and actionable recommendations that would help facilitate broader adoption of standard citation mechanisms and easier measurement of data reuse.

Taken as a whole, these recommendations suggest a top-down approach, proposing mostly actions on the part of stakeholders like publishers and repositories, rather than researchers themselves. The underrepresentation of suggestions to researchers may reflect the demographics of workshop participants; it is possible that more publishers and repository representatives were present and that they mostly considered things that they themselves could do, or conversely, that more researchers were present and that they felt the onus was on other stakeholders to take action. Since recommendations were made anonymously, it cannot be known whether participants were considering recommendations for themselves and their colleagues or to other stakeholders in the scholarly communication space. Of course, success of data citation efforts cannot rest on top-down approaches alone, as researchers must indeed cite data and know how to do so correctly. Given the previously observed "strong cultural inertia against data citations within many research communities," efforts focused on researchers are also needed to ensure that they know how and why to cite data (Mayernik, 2012).

One of the goals of this workshop was to encourage participants from a range of disciplinary backgrounds, geographic areas, and roles to think about how they could work within their own communities to encourage broader adoption of standard data citation and other mechanisms to facilitate the tracking and quantifying of research data reuse. Likewise, this report aims to share those suggestions with the broader scholarly community to serve as a potential agenda for advancing our collective ability to better understand research data reuse and thereby reward researchers who share highly reused data sets, not by suggesting specific next steps for stakeholders to take, but by providing possible direction for future efforts.

Disclosure Statement

The author declares no competing interests.

Acknowledgments

The author wishes to thank all the participants in the FORCE2018 preconference workshop who contributed the ideas described here, as well as Maryam Zaringhalam for input on the preparation of this report. The author also thanks the two anonymous peer reviewers who provided constructive feedback that helped improve this article. This work was carried out by staff of the National Library of Medicine (NLM), National Institutes of Health, with support from NLM.

References

Ali-Khan, S. E., Jean, A., MacDonald, E., & Gold, E. R. (2018). Defining success in open science. *MNI Open Research*, *2*, 2. https://doi.org/10.12688/mniopenres.12780.1

Alquraishi, M., & Sorger, P. K. (2016). Reproducibility will only come with data liberation. *Science Translational Medicine*, *8*(339), 7–10. https://doi.org/10.1126/scitranslmed.aaf0968

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0092590

Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., & Schindler, U. (2017). The scholix framework for interoperability in data-literature information exchange. *D-Lib Magazine*, *23*(1–2). https://doi.org/10.1045/january2017-burton

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, *7*(1), 107–113. https://doi.org/10.2218/ijdc.v7i1.218

Callahan, A., Winnenburg, R., & Shah, N. H. (2018, March). Analysis: U-Index, a dataset and an impact metric for informatics tools and databases. *Scientific Data*, 1–10.

Carpenter, C. R., Cone, D. C., & Sarli, C. C. (2014). Using publication metrics to highlight academic productivity and research impact. *Academic Emergency Medicine*, *21*(10), 1160–1172. https://doi.org/10.1111/acem.12482

Committee on Data for Science and Technology. (n.d.). CODATA-ICSTI data citation standards and practices. Retrieved January 3, 2019, from http://www.codata.org/task-groups/data-citation-standards-and-practices

Committee on Data for Science and Technology-International Council for Scientific and Technical Information Task Group on Data Citation Standards and Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, *12*, 1–75. https://doi.org/10.2481/dsj.OSOM13-043

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing citations and usage metrics together to make data count. *Data Science Journal*, *18*. https://doi.org/10.5334/dsj-2019-009

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M., & Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, *5*, 180259. https://doi.org/10.1038/sdata.2018.259

Couture, J. L., Blake, R. E., McDonald, G., & Ward, C. L. (2018). A funder-imposed data publication requirement seldom inspired data sharing. *PLOS ONE*, *13*(7), e0199789. https://doi.org/10.1371/journal.pone.0199789

Data Citation Synthesis Group. (2014). *Joint declaration of data citation principles* (M. Martone, Ed.). FORCE11. Retrieved from https://www.force11.org/group/joint-declaration-data-citation-principles-final

Dempsey, R. (2004). The provenance of ideas: Constructing a scientific bibliography. *Journal of Diagnostic Medical Sonography*, *20*(1), 20–24. https://doi.org/10.1177/8756479303261059

Federer, L. (2018). Quantifying biomedical data reuse: Do citations tell the whole story? [Manuscript in process].

Federer, L., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of data availability statements. *PLoS One*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, *6*(1), 28. https://doi.org/10.1038/s41597-019-0031-8

Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., Cruse, P., & Chodacki, J. (2018). Code of Practice for Research Data Usage Metrics Release 1. *PeerJ Preprints*, 1–43. https://doi.org/10.7287/peerj.preprints.26505v1

FORCE11. (n.d.). *Data Citation Implementation Pilot (DCIP)*. Retrieved January 31, 2020, from https://www.force11.org/group/dcip

FORCE11. (2014). *Data Citation Synthesis Working Group*. Retrieved January 3, 2019, from https://www.force11.org/group/data-citation-synthesis-working-group

Garfield, E. (1964a). Can citation indexing be automated? In *Statistical Assocation Methods for Mechanized Documentation* (Vol. 269, pp. 84–90). http://garfield.library.upenn.edu/essays/V1p084y1962-73.pdf

Garfield, E. (1964b). “Science Citation Index”—A New Dimension in Indexing. *Science*, *144*(3619), 649–654.

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2019). Lost or found? Discovering data needed for research. *Arxiv.* http://arxiv.org/abs/1909.00464

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *The HUGO Journal*, *8*(1). https://doi.org/10.1186/s11568-014-0003-1

Knoppers, B. M., Harris, J. R., Budin, I., & Edward, L. (2014). A human rights approach to an international code of conduct for genomic and clinical data sharing. *Human Genetics, 133* (1), 895–903. https://doi.org/10.1007/s00439-014-1432-6

Leung, K., Au, A., Huang, X., Kurman, J., Niit, T., & Niit, K.-K. (2007). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *21*, 108–119. https://doi.org/10.1002/per.615

Make Data Count. (2019). *About.* Retrieved March 23, 2019, from https://makedatacount.org/about/

Martens, L. & Vizcaíno, J. A.(2016). A golden age for working with public proteomics data. *Trends in Biochemical Sciences*, *42*(5), 333–341. https://doi.org/10.1016/J.TIBS.2017.01.001

Mayernik, M. S. (2012). Data citation initiatives and issues. *Bulletin of the American Society for Information Science and Technology*, *38*(5), 23–28. https://doi.org/10.1002/bult.2012.1720380508

National Institutes of Health. (2017). *NIH and other PHS agency Research Performance Progress Report (RPPR) instruction guide.* https://grants.nih.gov/grants/rppr/rppr_instruction_guide.pdf

National Science Foundation. (2016). *Grant preparation instructions.* https://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_2.jsp

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. A. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in *The BMJ* and *PLOS Medicine*. *Bmj*, *360*, k400. https://doi.org/10.1136/bmj.k400

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Piwowar, H. A. (2010). A method to track dataset reuse in biomedicine: Filtered GEO accession numbers in PubMed Central. *Proceedings of the ASIST Annual Meeting*, *47*, 1–2. https://doi.org/10.1002/meet.14504701450

Piwowar, H. A., Carlson, J. D., & Vision, T. J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the ASIST Annual Meeting*, *48*(1), 1–4. https://doi.org/10.1002/meet.2011.14504801337

Read, K. B., Sheehan, J. R., Huerta, M. F., Knecht, L. S., Mork, J. G., Humphreys, B. L., & Members of the Big Data Annotator Group(2015). Sizing the problem of improving discovery and access to NIH-funded data: A preliminary study. *PLoS ONE*, *10*(7), 1–18. https://doi.org/10.1371/journal.pone.0132735

Research Data Alliance. (2019). *RDA/WDS Publishing Data Services WG*. Retrieved January 31, 2020, from https://www.rd-alliance.org/groups/rdawds-publishing-data-services-wg.html

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2015). Analyzing data citation practices using the Data Citation Index. *Journal of the American Society for Information Science and Technology*, *18071*, 12. https://doi.org/10.1002/asi.23529

Rowhani-Farid, A., & Barnett, A. G. (2016). Has open data arrived at the *British Medical Journal* (BMJ)? An observational study. *BMJ Open*, *6*(10), 1–8. https://doi.org/10.1136/bmjopen-2016-011784

Rung, J., & Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, *14*, 89–99. https://doi.org/10.1038/nrg3394

Stuart, D. (2017). Data bibliometrics: metrics before norms. *Online Information Review*, *41*(3), 428–435. https://doi.org/10.1108/OIR-01-2017-0008

van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., & Petras, V. (2019). The definition of reuse. *Data Science Journal*, *18*(1). https://doi.org/10.5334/dsj-2019-022

Wiley. (2018). *Data sharing & citation*. Retrieved March 24, 2018, from https://authorservices.wiley.com/author-resources/Journal-Authors/licensing-open-access/open-access/data-sharing.html

Zhang, Q., Cheng, Q., Huang, Y., & Lu, W. (2016). A bootstrapping-based method to automatically identify data-usage statements in publications. *Journal of Data and Information Science*, *1*(1), 1–17. https://doi.org/10.20309/jdis.201606

Zhao, M., Yan, E., & Li, K. (2017). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, *69*(1), 32–46. https://doi.org/10.1002/asi.23919

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: Experiences of ecologists*. [Doctoral dissertation, University of Michigan]. Deep Blue Data Repository. https://deepblue.lib.umich.edu/handle/2027.42/61844

Zimmerman, A. S. (2008). New knowledge from old data. *Science, Technology, & Human Values*, *33*(5), 631–652. https://doi.org/10.1177/0162243907306704